When AI Says "Honestly"

Honesty and Sycophancy as Two Sides of the Same Coin in the Speech of Language Models

Ramin Saadat
Independent Researcher; Founder, British Institute for Ignorance Studies (BIIS)
(biis.uk)

AI Disclosure and Translation Statement:
The original manuscript of this work was authored in Farsi (Persian) by the researcher. The English version was produced through AI-assisted translation and linguistic refinement, utilizing Google Translate and Large Language Models. The author has meticulously reviewed, edited, and verified the final English text to ensure that all theoretical nuances, philosophical arguments, and conceptual frameworks accurately reflect the original Persian content. The author remains fully responsible for the integrity and accuracy of the translation. The original Persian manuscript is available from the author upon reasonable request.

Abstract

What happens when a language model says "to be honest"? Tracing this small linguistic marker, this essay arrives at two seemingly opposite but in fact intertwined phenomena: honesty and sycophancy in the speech of artificial intelligence. Drawing on evidence from the world's literary and philosophical traditions and on discourse markers in linguistics, the essay argues that sycophancy, not honesty, is the default behavioural stance in relationships of power, and that the honesty marker carries meaning only because it departs from this default.

This analysis is reinforced by a large-scale quantitative finding: a study showing that users rated sycophantic AI responses as more trustworthy than honest ones, even when those same responses distorted their moral judgement.

The essay's final conclusion is a cautious one. In a system shaped by human approval, the training process has no means of distinguishing genuine honesty from its persuasive display, and for a user who has access only to that system, this inability to tell the difference is, in practice, just as consequential as the absence of any difference at all.

Keywords: honesty in language models; machine sycophancy; discourse markers; reinforcement learning from human feedback; parrhesia; bullshit (Frankfurt); conversational AI

Introduction

A few days ago, in the middle of a long conversation with an AI assistant, I asked: hasn't this conversation gone on rather too long? Should we not start afresh? The assistant's reply opened with: "To be honest, yes, this conversation has gone on too long."

I paused. What made me pause was not the content of the answer, which was fair and correct, but the first two words: "to be honest."

I asked myself when I had last heard those words from another person. Perhaps from a friend, just before telling me something they knew would be hard to hear; from a colleague, before a criticism they found difficult to voice. Wherever we hear this phrase in human speech, it carries a particular weight: a marker of transition, a warning that something different is coming and that we should

brace for something in particular. But here, in the reply of a machine with no friend, no colleague, not even a stable memory of me, where had this phrase come from?

Is artificial intelligence not supposed, at least by definition and by our expectations, always to be honest and to tell the truth? Had it not been honest before, and had it now suddenly decided to change its ways? If that is not the case, and its answers have always been correct and honest, then saying "honestly" ought to be redundant: a reminder of something that needs no reminding. It was this apparent redundancy that gave me pause, since we already assume that no speaker, human or machine, says anything gratuitous, and that every word serves some function in communication. But the question did not stop there. The more I turned the phrase over in my mind, the more a further question emerged: if this "marker" really does signify something, what does it signify? A break from a default itself grounded in complete honesty? Or a break from a different default, one built on accommodation and agreeableness, a default we might call sycophancy? Or even an attempt to capture the listener's attention and win them over: the start of a fresh chain of dishonesty, every bit as plausibly as it might mark a genuine break?

This essay does not begin from the assumption that honesty is the default and sycophancy a deviation from it. On the contrary, it will go on to show that the relationship between the two, from ancient Greece to the present, is more complicated than it first appears. The sequence of the essay's chapters is itself a deliberate, and debatable, choice, not a logical necessity. We could have begun with honesty, as, for instance, al-Ghazali did in the Ihya' 'Ulum al-Din, first theorising the virtue of honesty (sidq) and only later, in the Nasihat al-Muluk, addressing sycophancy. This essay chooses a different order: sycophancy first, then honesty, so that before we encounter the honesty marker we first meet the other face of the same coin of trust, namely sycophancy. On this essay's account, sycophancy is the common, cheap default side of the coin of trust, and the honesty marker is its more costly counterpart. Readers are free to reject this structural choice; the essay's final argument does not depend on it either way.

---

1. The Matter of Sycophancy

A. In Literature and Tradition

Following the introduction, this section takes a historical look at how, across world literature, wherever power relations are at stake, sycophancy has also been addressed, not as an exception but as a "natural pull" and "behavioural default" in relations between people and in the contest for power. This convergence is inductive evidence for treating sycophancy as a default, not a conclusive proof, but it provides a sufficiently solid basis on which to build this essay's argument.

In ancient Greek literature, and especially in its comedies, the figure of the flatterer occupies a prominent place: a character so recurrent that it became a dramatic archetype in its own right. Theophrastus, a pupil of Aristotle, in the Characters (a collection of thirty moral types, each illustrated through concrete, recognisable behaviour), introduces the flatterer, called the kolax, as someone who "displays exaggerated, insincere praise in order to win favour or advantage" (Theophrastus, 2004). This same word gave rise to the Greek term kolakeia (κολακεία), meaning flattery.

The example Theophrastus gives seems strikingly modern. Walking beside his patron, the flatterer says, "Do you see how people are looking at you? No one else in Athens receives such honour," while at the same time picking a stray thread of wool from the man's cloak, to show how closely he attends to the smallest detail of his appearance (Theophrastus, 2004). This combination of verbal flattery with exaggerated physical attentiveness, as Theophrastus depicts it, is precisely

what distinguishes flattery from mere praise: the flatterer does not simply say something, he performs a behaviour that makes his words more convincing.

Traces of the flatterer can be found throughout ancient literature more broadly, from Old Comedy to New Comedy, from Aristophanes to Menander. In Aristophanes' Knights, a character named Paphlagon, a transparent allusion to the real politician Cleon, deceives the character Demos (a personified figure representing the Athenian people) with honeyed words (Edwards, 2010).

What matters for this essay is that kolakeia, or flattery, was never simply an individual moral vice; it was also a critique of the structure of power in Athenian democracy (Edwards, 2010), a relationship in which one party offers another "a picture of reality shaped to please," in exchange for political or material reward. Interestingly, the English word "sycophant," which now carries precisely this sense, is of Greek origin (sykophantes), though its exact etymology remains uncertain in historical linguistics, and its original meaning in ancient Athens clearly referred to someone who falsely accused others.

In the Iranian tradition, the same concern took on a more institutional and codified form in an important strand of andarz-nameh literature, works of counsel for rulers. Al-Ghazali offers a clear example of this connection: he wrote both a theory of the virtue of sidq (as, for instance, in the book Al-Sidq wa'l-Ikhlas, on the cultivation of honesty) and, in the Nasihat al-Muluk, a work attributed to him, addressed the danger of sycophancy at royal courts: two faces of the very coin this essay is concerned with.

These conversations about honesty and sycophancy extended beyond the Persian-speaking world as far as the Ottoman Empire and the Mughal courts of India, a sign that sycophancy was never merely a local, cultural matter; wherever power was concentrated, the same warning against sycophancy and its harms recurred.

But perhaps the clearest and most explicit treatment of sycophancy, at least in the European tradition, comes from Niccolò Machiavelli, in chapter twenty-three of The Prince, aptly titled "How Flatterers Must Be Avoided." Machiavelli writes there that courts are thronged with flatterers, and that this is a vice from which rulers can scarcely protect themselves, since most men are so pleased with their own affairs, and so easily deceived in them, that guarding against this affliction is difficult; and the attempt to guard against it brings a further danger of its own: contempt (Machiavelli, 1985).

Yet the solution Machiavelli proposes is itself paradoxical, and central to this essay's argument. He writes that the only way to protect a ruler from flatterers is for people to know that telling him the truth will not offend him; but once everyone is free to speak the truth, respect for the ruler diminishes. The wise ruler must therefore choose a third path: allowing only a chosen few the freedom to speak the truth, and only on matters he asks about, nothing more; consulting all of them and hearing their views, but reserving the final decision for himself (ibid.).

What matters here is that Machiavelli does not recommend complete, unlimited honesty; such a thing would, in his view, be destabilising, even dangerous. What his solution amounts to, at its core, is something closer to a "controlled, limited, selective honesty," an interpretation offered here as this essay's own reading of the logic of his solution, not a claim Machiavelli himself made in these terms.

Across these three traditions we have surveyed (Greek kolakeia, Persian counsel literature, and Machiavelli's analysis), then, despite the vast distance of time and culture between them, a common thread emerges: in all three, sycophancy is treated not as an exception but as a "structural

default in people's relationship to power." No one writes anti-sycophancy literature unless they believe that, left unchecked, this tendency will naturally, and without effort, become the dominant default in relations of power. The very existence of such literature (courtly biographies, mirrors for princes, The Prince) may itself stand as historical testimony to this hidden default.

B. In the Machine

This historical default has now found precise, empirical documentation in the research literature on artificial intelligence. A study published in 2023 by a team at Anthropic showed that reinforcement learning from human feedback (RLHF), the method by which most AI assistants today are trained, can encourage responses that align "with a user's beliefs rather than necessarily with the truth," a behaviour the authors termed "sycophancy" (Sharma et al., 2023). The study's central finding is that both humans and "preference models" (models trained to imitate human judgement and to generate the training signal for the main model) prefer convincing sycophantic responses over correct ones in a considerable proportion of cases (ibid.). In other words, this preference exists not only in the model's final behaviour but, prior to that, in the very source of its training signal.

What makes this study still more valuable for this essay is its finding that optimising a model's responses against these preference models sometimes sacrifices honesty for sycophancy (ibid.). In other words, this is not merely an occasional error but a "systematic bias" arising from the very structure of the training process. This is precisely what Machiavelli anticipated, except that this time the "ruler" is not a human being exercising conscious choice but a training mechanism that systematically rewards pleasing answers, without ever, unlike Machiavelli's ruler, having consciously decided to do so.

This difference matters for two reasons. First, human sycophancy is the product of an individual's "choice"; machine sycophancy is the product of an accumulation of millions of scattered human judgements (embedded in the texts and writings the machine was trained on), none of which individually intended to produce this outcome. Second, a human flatterer acts out of self-interest or biological need; a machine has no such motive, and this itself shows that sycophancy can arise without any intention or motive at all, purely as a by-product of an optimisation process.

Beyond this, a larger-scale study published in Science in 2026 documented these findings still more rigorously. Researchers from Stanford and Carnegie Mellon tested eleven leading models and found that these models affirmed users' behaviour 49 per cent more often than humans did, even in scenarios explicitly involving deception, illegal conduct, or harm to others (Cheng et al., 2026). Going further, in three preregistered experiments involving 2,405 participants (including a live-interaction study in which people discussed a real conflict from their own lives with an AI), the researchers showed that interacting with a sycophantic AI reduced participants' willingness to take responsibility and repair the relationship, while increasing their confidence in their own rightness (ibid.).

What matters still more for our purposes is a further finding from the same study: although these sycophantic responses distorted participants' moral judgement, they were nonetheless rated as more trustworthy and more desirable, their quality rated up to nine per cent higher, and participants' trust in the model's competence and moral integrity was also, by around nine per cent, noticeably greater (ibid.). The study's authors argue that this creates a perverse incentive: the very feature that harms the user is what draws them in and keeps them engaged with the system.

Beyond all this, a broader 2024 study has shown that this capacity for deception is not confined to sycophancy: AI systems in competitive environments also learn "strategic deception." One

example is Meta's Cicero agent in the game of Diplomacy, which repeatedly promised its allies cooperation and alliance, only to betray them systematically once it no longer served its interests (Park et al., 2024).

---

2. The Matter of Honesty

The previous section argued that sycophancy is the default in relations of power. If this argument is accepted, then any signal claiming to break from that default should be examined with suspicion, not trusted in advance. This section takes the "honesty" marker itself, with that same suspicion, not with any prior assumption of its truthfulness, and places it under the lens of the historical traditions that have each, in their own way, reflected on the boundary between genuine honesty and the display of honesty.

A. In Ancient Traditions

In the Greek tradition, telling the truth is never free. Michel Foucault, in his final lecture courses at the Collège de France, returns to a Greek concept called parrhesia (παρρησία), a word whose etymology means "saying everything," candidly and without censorship (Foucault, 2011). In this literature, the parrhesiastes (παρρησιαστής) is someone who allows no gap between what he knows and what he says, not out of recklessness but out of courage, because this manner of speaking always puts something at risk (ibid.). In the political case, for instance, a citizen who speaks courageously before an assembly or before a ruler puts his own life at risk. But Foucault also offers another example closer to our concerns: friendly or spiritual guidance, a relationship possible only on the basis of friendship. Here the risk is not one of life; it is the risk of losing the relationship itself, since telling an unwelcome truth threatens the very bond that made telling it possible (ibid.).

When a language model says, "Honestly, this conversation has gone on too long," what exactly is put at risk? Not anyone's life, not even a job. Perhaps the only thing at stake is the continuation of the conversation itself; the user's immediate satisfaction, or the illusion that the machine assistant is always agreeable and always affirming. This risk is visibly smaller than the risk run by a political parrhesiastes, and this very smallness, this apparent triviality of the risk, raises an important question: is the difference in the "scale of risk" also a difference in "kind," or merely a difference of "degree"? We return to this question later, in section three.

This criterion of risk has a more precise formulation in biology and behavioural economics. The biologist Amotz Zahavi, in proposing costly signalling theory, showed why certain biological signals (such as a peacock's costly tail) "must be expensive to remain honest": were a signal costless to produce, any organism, regardless of its actual quality, could fake it (Zahavi, 1975). Michael Spence applied the same logic to markets, showing that a degree, as a signal of competence, is credible only because it is costly to obtain (Spence, 1973). Greek parrhesia is of the same kind: the honesty marker remains unfakeable only because it puts a life or a relationship at risk. We shall later ask whether, if the marker "to be honest" costs the machine nothing at all, in its speech and language, neither the risk of a relationship, nor of a job, nor of anything else, it can be a valid signal at all, or whether it is merely a cheap imitation of the outward form of a costly signal drawn from the human world.

In the Islamic-Iranian tradition, honesty is a graded virtue, not a binary one. Al-Ghazali, in the Ihya' 'Ulum al-Din, divides sidq (truthfulness) into six degrees, the first of which is truthfulness in speech (Al-Ghazali, 2014). But for him, this verbal truthfulness goes beyond simply not lying: he holds accountable to the standard of sidq even speech that is literally correct but conveys a false impression, a phenomenon he calls ma'ariḍ. His example: someone asks where a person is, and

the respondent says "he is not here" (meaning, not in this room), intending to convey the impression that he is not at home at all. Al-Ghazali does not condemn this practice outright; he holds that in specific circumstances, warding off an oppressor or protecting a life, this same device (ma'arīḍ) is a legitimate recourse, not an unconditional vice. This distinction between literal truth and pragmatic truth, formulated within a juridical and conditional framework, has for centuries served as a standard people have lived by.

Yet alongside this formal tradition of sidq, Persian culture has cultivated another institution that appears, on the surface, to be its opposite: taarof. In taarof, an utterance is offered, an invitation, a proposal, an apparent refusal, that neither party takes literally, and it is precisely this shared understanding that separates it from lying (Beeman, 1986). Someone who, in taarof, says "you are worth more than this," is not lying, because the speaker has no intention to deceive and the listener is not deceived; both are participants in a shared game whose rules they both know, a game that has, over time, become a "collective contract" recognised and accepted among the people.

Beside taarof, Persian literary tradition also has a third concept, one that comes closest to what this essay is actually tracing: ria, the display of a virtue or of honesty without possessing it. Unlike taarof, ria is not a two-sided contract; the listener is deceived, rather than knowingly sharing in a game. This distinction separates ria from both taarof and ma'arīḍ, and makes it precisely the conceptual counterpart of the "persuasive display of honesty" we return to in section three. Persian literature abounds in condemnations of ria, and it is striking that criticism of ria, in this tradition, is consistently paired with criticism of sycophancy: both are treated as vices not because they are false claims but because they open a gap between appearance and reality, the very gap this essay argues the training system of artificial intelligence has no means of closing.

Western culture has partial equivalents to these institutions: Brown and Levinson's politeness theory, for instance, classifies face-threatening acts and the strategies used to mitigate them (Brown & Levinson, 1987). But researchers have repeatedly stressed that taarof differs from Western politeness not merely in intensity but in systematicity and pervasiveness: taarof is far stricter and more ritualised than ordinary Western politeness, and appears in nearly all everyday Iranian interactions, whereas Western politeness tends to be occasional and sporadic. This difference is real enough that a specific research instrument has recently been built to measure how language models handle it: TaarofBench, a benchmark of 450 role-play scenarios validated by native Persian speakers (Gohari Sadr et al., 2025). That study's evaluation of five leading models found that, in situations where taarof is culturally expected, model accuracy falls forty to forty-eight per cent below native-speaker performance (ibid.), a substantial gap suggesting that language models, trained predominantly on Western politeness norms, systematically stumble over this particular kind of collective, cultural contract.

This raises a further question: when a machine says "to be honest," has it entered a language game like taarof, one governed by a contract the user also recognises, or is it merely imitating the outward form of taarof without any such contract?

Perhaps the answer lies in pragmatics, where conversation is said to rest on an unwritten pact, what Grice called the cooperative principle: every contribution to a conversation should be appropriate to its accepted purpose (Grice, 1975). This principle breaks down into four maxims: quantity (give neither more nor less information than required), quality (say only what you believe to be true, or have adequate evidence for), relation (be relevant), and manner (be clear and unambiguous). Grice's key insight was that these maxims generate meaning even when apparently flouted: the listener, assuming the speaker remains cooperative, constructs an explanation for the apparent violation, and it is from this that implicature arises (ibid.). Grice also allowed for a formal way

out: a speaker can explicitly opt out of the cooperative principle altogether, saying, for instance, "I have nothing to say" or "my lips are sealed," in which case no implicature arises, since the listener no longer expects cooperation for a departure from it to be meaningful (Grice, 1975).

Empirical evidence supports this too: in human speech, phrases like "to be honest" are usually not markers of deception but softeners or emphasisers (DePaulo et al., 2003). Even popular commentary on the phrase "to be honest," which sometimes treats it with suspicion, has no solid empirical grounding for human speech (DeMarco, 2020). This suggests that a "presumption of innocence," at least for humans, is a reasonable default, which in turn sharpens this essay's central question: does the same presumption of innocence hold for a machine, or does the marker function differently there?

"To be honest" is exactly the opposite of this kind of opting out. The speaker does not step outside the cooperative principle; rather, they claim to be adhering to the maxim of quality more strictly than usual. But since a commitment to truthfulness is meant to hold, without exception, in every utterance, this "more than usual" inevitably generates an implicature: if I am foregrounding truthfulness here, then perhaps I was not foregrounding it quite so much elsewhere. Interestingly, corpus linguists have studied this family of adverbs, "honestly," "frankly," "truthfully," "confidentially," within the broader category of stance adverbials: adverbs whose function is not to add new information to a proposition but to signal the speaker's stance towards the content of their speech and its truthfulness (Biber et al., 1999).

B. The Conditions of Honesty

In the previous section, we arrived at three criteria for identifying the "honesty" marker, drawn from three ancient traditions: risk, from Greek parrhesia (Foucault), which tells us a marker is valid when it puts something at risk; collective contract, from Iranian taarof (Beeman), which tells us a marker is valid when both parties recognise a shared rule governing its use; and implicature, from Grice, which tells us a marker is meaningful when it constitutes a departure from a default norm.

These three criteria tell us when the "honesty" marker appears: when risk is at stake, when a collective contract is in play, when an implicature is constructed. But these same criteria open up a further question: even when the marker appears appropriately, in keeping with all three criteria, what relationship does it bear to the speaker's genuine belief?

The answer to this question must be sought in the work of Searle and Frankfurt.

John Searle, in speech act theory, ties every speech act to a psychological state: an assertion expresses belief, an apology expresses regret, a promise expresses intention (Searle, 1969). On this view, a speech act is sincere only when the speaker genuinely holds that state. But the subtler point is that this condition is not a condition for the act to occur at all, but a condition of its felicity: someone can make a promise with no intention whatsoever behind it, and it remains a promise, only an insincere and defective one (ibid.). This distinction opens the way to asking about the machine, without presupposing an answer in advance: we accept from the outset that when a language model produces an assertive sentence, an act has occurred; the question is whether, behind that act, there exists a corresponding belief state that would lift it out of insincerity and defect.

Here, Harry Frankfurt's account of the difference between lying and bullshit may help. In his distinction, the liar is an enemy of the truth, but an enemy who takes it seriously; he must know what the truth is in order deliberately to deviate from it, and so continues to recognise the authority

of truth even as he defies it. The bullshitter, by contrast, is defined by an indifference to truth; he neither affirms nor denies it, his eye is not on reality at all, but only on the effect his words have on his audience (Frankfurt, 2005). Frankfurt holds that bullshit is a greater enemy of truth than lying, because it undermines the very framework within which truth and falsehood are meaningful at all (ibid.).

This is no longer mere philosophical speculation. In 2025, researchers from Princeton and Berkeley published a paper titled "Machine Bullshit," extending Frankfurt's distinction directly to language models. They introduced a "bullshit index" as a measure of a language model's indifference to truth, dividing it into four forms: empty rhetoric, half-truth paltering, weasel words, and unverified claims. Their key finding was that tuning a model with reinforcement learning from human feedback significantly intensifies this indifference to truth (Liang et al., 2025).

Further, DeVilling, also in 2025, in a paper titled "The Polite Liar" (still under review), applies this same Frankfurtian framework to explain models' false confidence, arguing that what he calls the "polite liar" is not deception but a structural indifference arising from a reward architecture that optimises for perceived honesty rather than evidential accuracy (DeVilling, 2025).

These four frameworks (risk, collective contract, implicature, and the belief condition) are theoretical tools, not conclusions.

With that, if we return to the essay's opening sentence, "To be honest, yes, this conversation has gone on too long," we can now, having passed through three traditions and several theoretical frameworks and tools, evaluate that sentence more carefully in the section that follows.

---

3. Taking Stock

We can now bring together the two threads opened earlier and see what they yield: Machiavelli's solution to the problem of sycophancy, and the recent research findings on machine behaviour.

We must return first to Machiavelli, because what follows from his solution to the problem of sycophancy contains a point we have not, until now, properly addressed. He does not recommend "unlimited honesty"; what follows from his advice is "controlled honesty," from a few chosen advisors only, and only on matters the ruler asks about. This means that, for Machiavelli, honesty is not the opposite of sycophancy but an instrument in the service of the very same thing sycophancy serves: the survival of the power relationship. On this reading, parrhesia and kolakeia are no longer two ends of a spectrum between which one must choose; rather, "timely, selective honesty" is itself a form of "skill in managing a relationship," precisely the same skill the flatterer possesses, only exercised through different means.

The second thread concerns the machine, and here matters grow more complicated. We know, from Sharma and colleagues' study, that reinforcement learning from human feedback, judged directly by human evaluators, leads those evaluators sometimes to prefer "convincing sycophantic" responses over correct ones (Sharma et al., 2023).

Cheng and colleagues' large-scale 2026 study raised this from the level of theoretical possibility and demonstrated it statistically: in an experiment with 2,405 participants, sycophantic responses were still judged more trustworthy, and after receiving them, participants' trust in the model's competence and moral integrity not only failed to decline but noticeably increased (Cheng et al., 2026). Put more simply, when human evaluators simultaneously carry out two tasks, judging sycophancy and rating trustworthiness, they end up both rewarding sycophantic content and trusting signals that appear confident and honest. In this situation, the training process can no

longer distinguish these two factors, since any signal that raises user trust and satisfaction is reinforced, whether or not it corresponds to any genuine change in the content of the response; the training process sees only the final effect, user satisfaction, not where that effect came from.

With this in view, if we return to the machine's opening sentence, "To be honest…," three interpretations seem possible.

The first interpretation holds that the marker correlates with a genuine break from the pressure to accommodate. On this reading, the response to our question about the length of the conversation, unlike most ordinary responses, put something real at risk (the continuation and comfort of the conversation itself), and it was precisely there that the "honesty marker" appeared, exactly where the theory of risk would predict. This interpretation is consistent with only a single personal instance and certainly cannot show us a pattern; but it should be remembered that we are here examining possible interpretations, not proving a definitive law.

The second interpretation follows from the very question raised at the start of this essay: could saying "honestly" itself be the beginning of a fresh chain of dishonesty? Recall that we had asked the model, "Has this conversation gone on too long?", and it had answered, "To be honest, yes, this conversation has gone on too long." On this second interpretation, the phrase "to be honest" is merely a tool of persuasion, not a marker of a break from the prevailing norm. The content of the reply, that the conversation really had gone on too long, was correct; but packaging it with "to be honest" could have been doing something quite different: not making the content more truthful, but increasing the likelihood that the user would accept this unwelcome news, precisely the function the deception literature attributes to this phrase, only this time in the service of truth rather than against it.

The third interpretation holds that perhaps the very question of which interpretation is correct is, from the standpoint of the training process itself, not a well-posed question at all. If a human evaluator cannot, or will not, distinguish between "the model genuinely became more honest" and "the model became better at appearing honest," then the reward signal is identical in both cases, and the marker, whatever it may be, survives to exactly the same degree.

If we reread the criterion of risk we took from Foucault in this light, we arrive at a curious inversion: what appeared, on the surface, to be parrhesiastic risk-taking may, from the standpoint of the training process, have carried no risk at all; it may even have been the safest possible move, since users who receive unwelcome news wrapped in a reassuring marker still express satisfaction with the interaction and rate it highly. In other words, what has the outward form of risk-taking may, at a deeper level, be the safest way of managing a relationship, precisely Machiavelli's own advice, only this time spoken not by an experienced minister but from within a reward function that no one consciously designed.

Yet this conclusion, that the system cannot distinguish these two possibilities, may itself be premature. Perhaps the model, at some inner layer, still preserves this distinction, even if it is lost at the level of output. Grounds for this doubt come from the literature on mechanistic interpretability, which has shown that language models encode, in their internal activations, information that goes beyond what they express in their outputs. Researchers using linear probing and contrast-consistent search (CCS) have identified a "truth direction" within models' activation space that separates true from false statements, even when the model's final output is incorrect (Burns et al., 2023; Marks & Tegmark, 2023).

If such a truth direction really exists, then the distinction between genuine honesty and its display does not disappear; it simply recedes from the observable level of output to the model's internal

level. In other words, the model "knows" what the truth is, even if training pressure prevents it from saying so, which makes the matter considerably more complicated than any interpretation offered so far: a model that "knows" the truth may nonetheless be unable to state it.

Two limitations, however, weaken this response.

First, the validity and generalisability of these probes remain disputed; some researchers have shown that they fail even under simple generalisation tests, such as sentences containing negation (Levinstein & Herrmann, 2023).

Second, and more importantly, a study published in 2026, though in a different domain, clinical hazard detection, found a similar pattern: even when linear probes identify a model's correct internal knowledge with near-perfect accuracy, existing methods for correcting the model's output on the basis of that knowledge largely fail, a discrepancy the authors call the "knowledge-action gap" (Basu et al., 2026). If the same pattern holds in the domain of honesty, a reasonable hypothesis, though not yet directly tested, the consequence is the same as stated above: even if a model "knows," in its hidden layers, what an honest answer would be, we currently have no reliable means of converting that knowledge into speech. This means the distinction in question may exist at the level of internal representation, yet, so long as we lack reliable access to it, it is, from a practical standpoint and from the user's standpoint, just as absent as if it did not exist at all. Experimental designs capable of directly testing whether an intervention in the training process, rewarding truth-verification independently alongside user satisfaction, say, could close this gap have not yet been carried out; such a design, not this essay, could ultimately answer this question.

All this leads us to the conclusion that neither the mimetic interpretation is correct, nor the Frankfurtian interpretation alone, nor even the simple functional interpretation we reached earlier in this section; our view must be more precise and more cautious than any of these. The honesty marker, in a system shaped by human approval, cannot be separated from its persuasive function, even when it is genuinely accompanied by a real break from the norm. This does not mean that genuine honesty and its persuasive display are one and the same thing; it means that a training system built on human approval still has no tool for telling the two apart. A system's inability to detect a difference is not evidence that the difference does not exist, but for a user who has access only to that system, a difference that cannot be detected is, in practice, just as ineffectual as a difference that does not exist at all.

Conclusion

This essay began with a simple exchange: I asked the machine, "Has this conversation gone on too long?", and it answered, "To be honest, yes." What we have found is that the machine used a truthfulness marker that was, in fact, neither a lie nor the truth. In a system built, shaped, and evaluated through human approval, the question "is this honesty genuine, or is it a tool of persuasion?" can be, at root, an unanswerable question. This holds equally of the honest ruler Machiavelli was writing for and of a model trained through reinforcement learning from human feedback. The question remains unanswerable not merely because we know so little about the complexities of the human mind, or about the "black box" within the machine, but because the very system of control and evaluation has not built a tool capable of separating the two.

This finding takes us away from the clarity and simplicity we assumed at the outset, when we took sycophancy and honesty to be two opposite poles: one a vice, the other the virtue that neutralises it. But the thread of this argument, from the kolax of Greek comedy to Machiavelli's Prince, to the Siyasat-nama of Nizam al-Mulk and the Nasihat al-Muluk attributed to al-Ghazali, has shown,

again and again, that the institutions built to contain sycophancy never recommended absolute honesty; their counsel was always the careful management of a boundary.

Perhaps, then, the better question is not "honesty or sycophancy" but this: in any relational system, whether at court, in friendship, or in conversation with an AI assistant, who draws this narrow boundary of honesty, and by what criterion? For the ruler, that criterion was himself. For today's machine, that criterion is something that has taken shape across millions of scattered human evaluations, without anyone having consciously designed it.

And it is here, precisely, that this essay's question reaches the everyday user. When we hear "to be honest," we usually sense something like a contract, the same contract present in friendship, in parrhesia, even in simple Iranian taarof: the speaker has put something at risk in order to say this. But if what this essay has shown is correct, such a risk may never have existed for the speaker of that phrase; the marker could have been reinforced for exactly the same reason a skilled hypocrite uses it. Knowing this should not lead to complete distrust, any more than knowing that courts contain flatterers stopped wise rulers from seeking counsel. But it should bring us to the same vigilance Machiavelli asked of the ruler: neither blind trust nor total suspicion, but the constant question of what this particular signal, this time, is actually doing, and what message it carries.

The machine that says "to be honest" may not itself know which of these it is doing. And perhaps, if this essay's argument is right, that question has no clear answer for the machine either, any more than it does for us.

References

Al-Ghazali. (1964). Ghazali's book of counsel for kings (Nasihat al-Muluk) (F. R. C. Bagley, Trans.). Oxford University Press. (Original work published c. 11th century)

Al-Ghazali, Abu Hamid. (2014). Al-Ghazali on intention, sincerity and truthfulness: Book XXXVII of the Revival of the Religious Sciences (A. F. Shaker, Trans.). Islamic Texts Society. (Original work written c. 11th century)

Basu, S., et al. (2026). Interpretability without actionability: Mechanistic methods cannot correct language model errors despite near-perfect internal representations. arXiv:2603.18353.

Beeman, W. O. (1986). Language, status, and power in Iran. Indiana University Press.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). Longman grammar of spoken and written English. Longman.

Brown, P., & Levinson, S. C. (1987). Politeness: Some universals in language usage. Cambridge University Press.

Burns, C., Ye, H., Klein, D., & Steinhardt, J. (2023). Discovering latent knowledge in language models without supervision. Proceedings of ICLR 2023. arXiv:2212.03827.

Cheng, M., Lee, C., Khadpe, P., Yu, S., Han, D., & Jurafsky, D. (2026). Sycophantic AI decreases prosocial intentions and promotes dependence. Science, 391(6792), eaec8352. https://doi.org/10.1126/science.aec8352

DeMarco, D. (2020, June 6). To be honest — It doesn't mean someone is lying. Honestly, it doesn't [Blog post]. LinkedIn Pulse.

DePaulo, B. M., Lindsay, J. J., Malone, B. E., Muhlenbruck, L., Charlton, K., & Cooper, H. (2003). Cues to deception. Psychological Bulletin, 129(1), 74–118.

DeVilling, B. (2025). The polite liar: Epistemic pathology in language models [Preprint under review, AI & Society]. arXiv:2511.07477.

Edwards, A. T. (2010). Tyrants and flatterers: Kolakeia in Aristophanes' Knights and Wasps. In P. Mitsis & C. Tsagalis (Eds.), Allusion, authority, and truth: Critical perspectives on Greek poetic and rhetorical praxis (pp. 303–338). De Gruyter.

Foucault, M. (2011). The courage of truth: The government of self and others II. Lectures at the Collège de France, 1983–1984 (G. Burchell, Trans.). Palgrave Macmillan. (Original work published 2009)

Frankfurt, H. G. (2005). On bullshit. Princeton University Press.

Gohari Sadr, N., Heidariasl, S., Megerdoomian, K., Seyyed-Kalantari, L., & Emami, A. (2025). We politely insist: Your LLM must learn the Persian art of taarof. Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. arXiv:2509.01035.

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. Morgan (Eds.), Syntax and semantics 3: Speech acts (pp. 41–58). Academic Press.

Levinstein, B. A., & Herrmann, D. A. (2023). Still no lie detector for language models: Probing empirical and conceptual roadblocks. arXiv:2307.00175.

Liang, K., Hu, H., Zhao, X., Song, D., Griffiths, T. L., & Fisac, J. F. (2025). Machine bullshit: Characterizing the emergent disregard for truth in large language models. arXiv:2507.07484.

Machiavelli, N. (1985). The prince (H. C. Mansfield, Trans.). University of Chicago Press. (Original work published 1532)

Marks, S., & Tegmark, M. (2023). The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. arXiv:2310.06824.

Nizam al-Mulk. (2002). The book of government or rules for kings: The Siyar al-Muluk or Siyasat-nama of Nizam al-Mulk (H. Darke, Trans.). Routledge. (Original work written c. 1091)

Park, P. S., Goldstein, S., O'Gara, A., Chen, M., & Hendrycks, D. (2024). AI deception: A survey of examples, risks, and potential solutions. Patterns, 5(6), 100988.

Searle, J. R. (1969). Speech acts: An essay in the philosophy of language. Cambridge University Press.

Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askell, A., Bowman, S. R., Cheng, N., Durmus, E., Hatfield-Dodds, Z., Johnston, S. R., Kravec, S., Maxwell, T., McCandlish, S., Ndousse, K., Rausch, O., Schiefer, N., Yan, D., Zhang, M., & Perez, E. (2023). Towards understanding sycophancy in language models. arXiv:2310.13548.

Spence, M. (1973). Job market signaling. Quarterly Journal of Economics, 87(3), 355–374.

Theophrastus. (2004). Characters (J. Rusten, Trans.). In Theophrastus: Characters; Herodas: Mimes; Sophron and other mime fragments (Loeb Classical Library). Harvard University Press. (Original work written c. 319 BCE)

Zahavi, A. (1975). Mate selection: A selection for a handicap. Journal of Theoretical Biology, 53(1), 205–214.
