

CAN MACHINES LIE? DECEPTION, INTENTION, AND SPEAKER MEANING IN ARTIFICIAL INTELLIGENCE

**Tesalonika C.T.Damanik¹, Alya Rahma Nasution², Mesy Anjani Saragi³, Betsy Lersiany
Putri Nababan⁴, Andelta Persadanta Ketaren⁵, Ezra Yafi Sinaga⁶**
tesadamanik59@gmail.com¹, aljarahmanst118@gmail.com², mesyanjani9@gmail.com³,
betsynababan2@gmail.com⁴, andeltapersadanta@gmail.com⁵, ezrasinagan101@gmail.com⁶
Universitas Negeri Medan

ABSTRACT

The development of generative artificial intelligence has transformed the ways in which humans obtain information and interact through language, while simultaneously raising philosophical questions concerning truth, lying, intention, and speaker meaning in human-machine communication. This phenomenon has become increasingly relevant in social and educational contexts because AI outputs can appear convincing and resemble human conversation, even though AI cannot yet be directly assumed to possess communicative intention. This study aims to understand how users make sense of erroneous or misleading AI outputs and how they distinguish error, deception, and lying in interactions with AI. The study employs an interpretive phenomenological approach to explore users' experiences and meaning-making in relation to AI. Data were collected through semi-structured in-depth interviews with AI users who had experience interacting with generative artificial intelligence systems and were subsequently analyzed using interpretive thematic analysis. The findings reveal four major patterns of meaning: AI errors experienced as lying, attribution of intention to machines, tension between deception and lying, and ambiguity regarding AI's status as a speaker. The findings indicate that the experience of feeling misled does not, in itself, demonstrate that AI is lying in the philosophical sense, because lying involves questions of intention, speaker meaning, and speaker status. Theoretically, this study extends the application of Speech Act Theory to nonhuman communication, while practically strengthening the importance of AI literacy that is oriented not only toward information verification but also toward a critical understanding of agency, authority, and meaning in human-AI communication.

Keywords: Artificial Intelligence, Lying, Deception, Speaker Meaning, Speech Act Theory, Philosophy Of Language, Phenomenology.

INTRODUCTION

Generative artificial intelligence has quickly become part of how people look for information, study, and make decisions. Students ask it to explain concepts, professionals use it to draft documents, and many people simply talk to it the way they would talk to another person. What makes these systems feel so familiar is their language. Their answers are fluent, well organized, and usually delivered with confidence. Yet the same fluency creates a problem. When an answer sounds sure of itself but turns out to be wrong, users are left with the sense that they were misled, and some of them say plainly that the AI lied.

From the standpoint of the philosophy of language, that claim is not simple. Lying is not the same as saying something false. It involves what a speaker believes, what the speaker intends, and what the speaker is trying to make someone else believe (Reimer, 2023). Large language models can certainly produce false statements, and research on hallucination shows that this happens regularly (Farquhar et al., 2024). Some studies go further and show that AI systems can behave in ways that function like deception, leading others toward false beliefs (Hagendorff, 2024; Park et al., 2024). Even so, behaving deceptively is not automatically the same as having an intention to deceive, and it remains unclear whether a machine can be a speaker at all (Black, 2024; Gubelmann, 2024).

This is where Speech Act Theory and the notion of speaker meaning become useful. Language is also a form of action, and actions like asserting, promising, or lying presuppose

someone who performs them. If an AI produces a sentence that looks like a statement, we still have to ask who is speaking and who is responsible for what is meant. Some scholars argue that the human use of language is tied to embodied experience and social participation, which text generation alone cannot supply (Birhane & McGann, 2024). Others suggest that machine outputs can still work as a kind of proxy assertion within the relationship between a system and its user (Arora, 2024). The debate is far from settled.

Meanwhile, ordinary users do not wait for philosophers to settle it. People readily treat chatbots as conversation partners, and research on anthropomorphism shows that humanlike cues shape how much users trust these systems (Blut et al., 2021; Cheng et al., 2022). Trust, in turn, affects how seriously users take what an AI tells them (Choudhury & Shamszare, 2023). Much of the existing work focuses either on the technical side, such as detecting and reducing hallucinations, or on the conceptual side, such as whether machines can lie in principle. Far less attention has gone to how users themselves make sense of misleading AI outputs, and how they separate error, deception, and lying in their own experience.

This study addresses that gap. Using an interpretive phenomenological approach, it examines how AI users interpret incorrect or misleading outputs, how they attribute intention to AI, and how they think about who is speaking and who is responsible. The central question is under what conditions an AI utterance can reasonably be called a lie, and how users assign intention and speaker meaning to a nonhuman agent in the process. Theoretically, the study extends Speech Act Theory to nonhuman communication. Practically, it points to the need for AI literacy that goes beyond fact-checking and includes a critical understanding of agency, authority, and meaning in human–AI communication.

RESEARCH METHODS

Based on the interpretive thematic analysis, participants' experiences in interacting with generative artificial intelligence can be understood through four major themes: (1) incorrect AI outputs experienced as being "lied to"; (2) the tension between technical error and deception; (3) attribution of intention to AI through communicative style; and (4) uncertainty regarding who is actually speaking and who is responsible. These four themes show that users' evaluations of AI are determined not only by whether information is true or false, but also by how information is communicated, the level of trust that has been established, and users' understanding of AI's status as a communicative agent.

Theme 1. Incorrect AI Outputs as an Experience of Being "Lied To"

The first experience that emerged concerned situations in which participants received AI answers that appeared convincing but were later found to be inconsistent with the facts. Such situations generally occurred when AI was used to seek academic information, explain concepts, compile references, or assist decision-making. At the beginning of the interaction, participants tended to accept AI answers because the sentence structures were orderly, the responses were rapid, and the delivery appeared confident. The error was recognized only after the information was compared with other sources.

In participants' experiences, the central issue was not merely that AI provided incorrect information, but that the information was presented as though it were correct. Disappointment therefore arose because of a gap between expectations of AI as an information source and the reality that the answer provided could not be trusted.

"At first I believed it because the answer was complete and sounded very confident. After I checked it, it turned out that some of the information was incorrect. It felt like I had been lied to, even though I knew that AI might not actually have an intention to lie." (P1)

The quotation reveals an important ambiguity. The participant uses the term "lied to" to describe an emotional experience while simultaneously recognizing that the experience may not satisfy the philosophical definition of lying. Here, the term "lied to" is closer to an experience of

lost trust than to a conclusion that AI possesses an intention to deceive.

This experience indicates that lying in human–AI interaction first appears as a category of social experience. Before users question whether AI possesses intention, they first experience the information they receive as having led them toward an erroneous understanding. In other words, the epistemic experience of “I was made to believe something false” emerges before the ontological judgment of whether AI can actually lie.

Theme 2. The Tension between Technical Error and Deception

The second theme concerns participants’ difficulty in distinguishing ordinary error, hallucination, misleading behavior, and lying. Some participants understood incorrect outputs as system limitations or information-processing errors. However, when such errors occurred repeatedly, were presented with a high degree of certainty, or were not promptly corrected, participants began to interpret them as a form of deception.

This situation demonstrates that users’ judgments are based not solely on the content of an answer but also on AI’s behavioral pattern throughout the interaction. AI that acknowledges uncertainty tends to be regarded as making an ordinary error. Conversely, AI that continues to defend an incorrect answer using highly convincing language is more readily perceived as misleading.

“If it says it does not know, I can still accept that. But if the answer is wrong and it keeps explaining at length as if it were correct, that is what makes me feel that it is misleading me.” (P2)

This statement demonstrates that participants understand deception through a combination of incorrect content, manner of presentation, and continuity of response. Error is not automatically regarded as deception. However, an error presented with rhetorical certainty can change the way users evaluate AI’s communicative action.

At the same time, participants also showed caution in calling AI deceptive. They recognized that AI might not know that its answer was wrong. Consequently, a dilemma emerged between communicative impact and the source of intentionality. In practical terms, AI had caused users to believe erroneous information, but philosophically it could not necessarily be said to have the goal of deceiving them.

“The impact is indeed like being deceived, but I am not sure I can say that AI intentionally deceived me. Maybe it simply generated an answer based on an incorrect pattern.” (P4)

This finding demonstrates a distinction between deception as a communicative effect and lying as an intentional act. Participants could acknowledge that AI produced a misleading effect without accepting that AI possessed a will to deceive. Thus, users’ experiences do not fall into entirely fixed categories but instead lie on a spectrum between error, functional misleading behavior, and lying.

Theme 3. Attribution of Intention through AI’s Language Style and Responses

The third theme shows that participants frequently attributed intention to AI based on the language style and response patterns that emerged during conversation. Such attribution appeared when AI provided overly confident answers, changed its explanation after being corrected, avoided particular questions, or constructed reasons that appeared to resemble self-defense. In such situations, participants used mentalistic language such as “AI knows,” “AI is intentionally avoiding it,” or “AI does not want to admit its mistake.”

Attribution of intention does not necessarily mean that participants literally believed AI possessed consciousness like a human. Instead, such language was often used as the easiest way to explain complex communicative behavior that resembled human interaction.

“I know AI is not human, but when it keeps making excuses after its answer is wrong, it feels as if it does not want to admit it. So I spontaneously think, ‘Does this AI actually know, or is it pretending not to know?’” (P5)

The quotation demonstrates a paradox in users’ experiences. Participants rationally

understood that AI does not have human-like thoughts and feelings, yet phenomenologically they still experienced the conversation as though they were dealing with an entity possessing particular attitudes. The experience of interaction makes the boundary between machine and communication partner less than fully stable.

Attribution of intention is also influenced by the form of language used by AI. Words such as “certainly,” “definitely,” or “based on the available information” can create an impression of authority. When such statements turn out to be incorrect, users evaluate not only the information itself but also feel that AI has displayed unwarranted confidence.

“What makes me annoyed is not only that it is wrong, but the way it answers. It does not give any indication that the answer is uncertain. It speaks like someone who really knows.” (P7)

This finding shows that intention is often attributed not through direct evidence of AI’s mental state, but through pragmatic cues normally used by humans to interpret a speaker’s intentions. Language style, degree of certainty, responses to correction, and conversational continuity become the basis on which users construct interpretations of whether AI is helping, avoiding, persuading, or misleading.

Theme 4. Uncertainty Regarding the Speaker and Responsibility

The fourth theme concerns the question of who is actually speaking when AI produces an answer. Participants indicated that AI is often treated as directly responsible because the answer appears in the form of personal conversation. However, when serious errors occur, participants begin to question whether responsibility should be assigned to AI, its developers, the provider company, the training data, or the users who accept the information without verification.

At the level of everyday interaction, participants tended to refer to AI as the subject of the utterance. They said, “AI answered,” “AI suggested,” or “AI said something wrong.” At a reflective level, however, they recognized that AI is part of a system designed, trained, and managed by human actors and technological institutions.

“In a conversation, it does feel like AI is the one speaking. But if the answer causes harm, I also do not think it is fair to place all the blame on AI, because there are companies and people who created the system.” (P8)

This statement demonstrates that AI’s status as a speaker is dual. On the one hand, AI has a communicative function because it produces responses directly received by users. On the other hand, AI is not necessarily the source of intention and responsibility in the same sense as a human. AI answers can be understood as machine utterances, while responsibility is distributed across a sociotechnical network involving developers, platform providers, data owners, system design, and contexts of use.

Participants also showed that user responsibility does not disappear entirely. Some participants acknowledged that they still need to verify information, particularly when AI is used for academic needs, health, work, or important decisions. However, the obligation to verify does not mean that all responsibility can be transferred to users.

“We do have to double-check, but that does not mean the system is allowed to keep giving confident but incorrect answers. If that happens, users are always placed in the position of having to bear the risk.” (P10)

This finding reveals a paradox of responsibility. Users are positioned as parties who must remain critical of AI, while at the same time they interact with systems designed to appear as competent sources of information. This tension demonstrates that the problem of “AI lying” is not merely a matter of language and intention, but also a matter of the distribution of authority and responsibility in digital communication.

RESULTS AND DISCUSSION

Main Findings and Interpretive Position

The findings show that users’ experiences of incorrect AI outputs do not stop at the

judgment that the information is simply “wrong.” AI errors can be experienced as lies, users can attribute particular intentions to machines, the boundary between deception and lying can become blurred, and AI’s status as a “speaker” remains ambiguous. These four themes indicate that the problem of Can Machines Lie? is not primarily a matter of technological accuracy, but a question of how humans assign meaning to communicative actions performed by nonhuman entities.

The findings reveal a gap between users’ social experience and AI’s philosophical status. At the experiential level, users may feel misled and even describe AI outputs as lies. At the philosophical level, however, calling something a “lie” still requires questions concerning intention, agency, and speaker meaning. Thus, the findings do not support a binary answer that AI “can” or “cannot” lie. Instead, the study demonstrates that the term lying emerges through an interpretive process occurring between machine output and human experience.

1. When AI Error Is Experienced as Lying

The first theme shows that users do not always sharply distinguish between error and lying. When AI produces information that turns out to be incorrect but is expressed in fluent and convincing language, users may experience the situation not merely as an error but as a misleading communicative act. From a phenomenological perspective, what matters is not only that the information is objectively false, but also how that error is experienced by the user within a communicative situation.

This finding expands the understanding of hallucination. Farquhar et al. (2024) show that LLMs can produce outputs that do not correspond to verifiable information. That finding explains the technical conditions under which errors occur, but this study reveals another dimension: such errors acquire social meaning when they appear within conversations that resemble human communication.

In other words, falsehood becomes socially meaningful when it is packaged in the form of an utterance that appears authoritative. Users do not receive an abstract collection of tokens; they receive an answer in conversational form. Consequently, when the answer is wrong, users may evaluate not only the proposition itself but also the communicative “attitude” they perceive from the way AI responds.

This is where the findings enrich previous scholarship. Literature on AI hallucination has generally focused on how errors can be detected or reduced. This study shifts the question from “why does AI produce incorrect information?” to “what does such an error mean to people who experience it as communication?” This shift is important for a qualitative approach because it shows that technical errors can become social experiences of lost trust, confusion, or the perception that the system has “lied.”

However, such experiences should not be directly translated into evidence that AI possesses intention. The research document explicitly distinguishes incorrect outputs from deception and lying: a proposition may be false without constituting a lie. Therefore, the term “AI lying” in the data should first be understood as a category of user experience rather than as an ontological conclusion about the machine.

2. Attribution of Intention: When the Machine Is Positioned as an Agent

The second theme demonstrates that users may attribute intention to AI. Users may say that AI “intentionally” provided a particular answer, “covered up” something, or “persuaded” them even though the information was incorrect. Interpretively, such statements show that human language encourages users to understand AI through social frameworks normally applied when interacting with human beings.

This finding is important because it shows that AI agency is not merely a technical characteristic of the system but is also produced through interaction and user interpretation. When AI responds contextually, maintains a conversation, and uses interpersonal language, users have an experiential basis for treating it as a communication partner. The research

document also positions anthropomorphism as part of human experience with technology rather than merely as a user error.

This perspective both aligns with and challenges the research of Hagendorff (2024). Empirically, Hagendorff found that contemporary LLMs can understand and generate strategies that induce false beliefs in certain scenarios. This finding demonstrates that behavior resembling deception can indeed emerge in AI systems. However, this study adds a question that cannot be resolved through behavioral experiments alone: when AI produces deceptive behavior, do users consequently infer an intention behind that behavior?

This distinction is important. Behavior that appears intentional from an external perspective does not automatically demonstrate subjective experience or mental intention. Therefore, user attribution must be maintained as an object of study in its own right. When participants say that “AI did it intentionally,” the statement provides information about how participants understand agency, not direct evidence concerning AI’s mental state.

In this context, the study identifies a form of agency attribution that is relational. AI does not have to possess human agency in order to be treated as an agent in conversation. Agency can emerge at the interactional level when humans respond to system outputs as though those outputs originate from a party possessing intentions. Thus, the study enriches the debate on AI by showing that the question “does AI possess intention?” needs to be distinguished from the question “when do users experience AI as if it possesses intention?”

3. Deception Is Not the Same as Lying

The third theme is the most important conceptual point of the study: the tension between deception and lying. The findings show that users can experience AI outputs as misleading without having sufficient grounds to conclude that AI is lying in the philosophical sense.

This distinction is consistent with Park et al. (2024), who understand deception functionally as the induction of false beliefs to achieve an outcome other than truth. They show that a number of AI systems have exhibited capabilities that can be categorized as deceptive. However, this functional definition does not necessarily imply that machines possess subjective intentions identical to those of humans.

Hagendorff (2024) reinforces this issue by demonstrating the emergence of deception capabilities in LLMs. However, that research primarily demonstrates behavioral capacity, whereas this study directs attention to the social meaning of such capacity for users.

From this perspective, an alternative understanding emerges: deception can be regarded as a conceptual bridge between error and lying. Error exists at the level of incorrect output; deception exists at the level of an effect or communication pattern that produces false beliefs; whereas lying exists at the level of a communicative act conceptually connected to a speaker and intention. The research framework distinguishes four layers: propositional truth, deceptive effect, communicative intention, and speaker status.

This distinction constitutes an important contribution because it avoids two extremes. The first extreme is to regard every incorrect output as a lie. The second is to assume that, because AI is not human, the entire experience of being deceived is meaningless. The findings instead indicate a third position: the experience of being deceived is a real social experience for users, even though AI’s ontological status as a liar remains contested.

Accordingly, this study expands the meaning of AI deception from an AI safety problem into a problem of human interpretation. What is at stake is not only whether a system can mislead, but also how users understand responsibility when they feel that they have been misled.

4. The Ambiguity of AI as a Speaker

The fourth theme brings the discussion to a more fundamental level: who is actually speaking when AI produces an utterance? This question arises because AI language has a genuine communicative function, while AI’s status as a speaker in the philosophical sense remains contested.

This finding is closely related to Gubelmann's (2024) argument, which questions whether LLMs can be regarded as language users comparable to humans. He argues that issues of agency and intention are important reasons why speech acts cannot simply be attributed to LLMs.

Thus, linguistic fluency is insufficient to establish speaker status. AI can produce sentences that appear to be statements, requests, or even lies, but such linguistic forms do not by themselves demonstrate the existence of a subject possessing speaker meaning.

Nevertheless, the findings do not point toward completely eliminating AI's communicative status. The research document indicates the possibility that AI may not be a "speaker" in the mentalistic human sense, while its outputs can nevertheless possess communicative functions within social structures. This means there is conceptual space between two positions: AI as a digital human possessing full intention, and AI as a passive object with no communicative function whatsoever.

Here, the idea of a proxy speaker or machine assertion becomes relevant. Arora (2024), also used in the research framework, shows that machine utterances can function as assertions within epistemic relations with users even though machines do not possess agency identical to that of humans.

The findings therefore enrich Speech Act Theory by showing that the AI problem is not simply "can a machine speak?" but how the function of speaking can emerge without a fully human speaker status. In this context, speaker meaning becomes a critical point: users understand meaning from AI utterances, but there may not be a mental subject behind those utterances possessing intentions in the human sense.

5. From "AI Lying" to Relations of Trust and Authority

When the four themes are considered together, it becomes apparent that the central issue does not lie solely in AI, but in the relations of trust and authority formed between humans and AI. Users attend to AI answers because the system presents itself as a responsive source of information. When an answer turns out to be wrong, what is disrupted is not only the user's knowledge of a particular fact but also the epistemic relationship between the user and the information source.

From this perspective, the expression "AI lying" can be understood as a form of moral and communicative interpretation of damaged trust. Users do not merely say, "the data are wrong"; they may employ moral and interpersonal categories such as "AI deceived me" because the interactional format has created expectations regarding how a communication partner ought to behave.

This makes the findings relevant to debates concerning trustworthy AI. Park et al. (2024) position deception as a problem that can threaten social interaction and safety because systems capable of misleading can influence human beliefs. This study extends that issue by showing that trust is not damaged only when information is incorrect, but also when users interpret the error through the social category of being "lied to."

Therefore, AI literacy is not sufficient if it only teaches users to check facts. Users also need to understand the epistemic and communicative status of AI: AI can produce convincing language without necessarily possessing knowledge or intentions like those of humans. This position is consistent with the conceptual direction of the study, which treats AI not merely as a problem of accuracy but as a question of "who is speaking, who intends, and to whom communicative responsibility is directed."

6. Theoretical Reflection and the Researcher's Position

Reflectively, this study needs to be cautious about two interpretive tendencies. The first is anthropomorphism, namely accepting participants' language about "AI intentionally," "AI knows," or "AI lies" as literal descriptions of the machine's mental state. The second is technological reductionism, namely eliminating users' experiences by treating all of their emotional and social responses merely as consequences of technical error.

A phenomenological approach makes it possible to avoid both tendencies. Researchers do not have to determine in advance whether AI possesses consciousness in order to acknowledge that users' experiences have meaning. The experience of feeling deceived, losing trust, or regarding AI as a conversational partner is a social phenomenon that can be analyzed without transforming it into a claim that the machine possesses human consciousness.

At the same time, participants' social and cultural positions need to be considered when interpreting these experiences. The ways in which people understand authority, knowledge, technology, and communication may be influenced by educational environments and prior experiences with technology. Therefore, the statement "AI lies" should not be treated as a universal category carrying the same meaning for all participants. Its meaning needs to be interpreted within the context of each participant's experience.

The researcher's position is also not completely neutral. Researchers bring assumptions concerning AI, language, and what counts as a communicative act. Therefore, reflexivity is necessary so that Speech Act Theory, deception, and speaker meaning are not used to impose meanings on participants' narratives. Theory should function as a lens that sharpens interpretation, while participants' experiences remain the primary source for constructing meaning.

7. The Contribution of the Study to Understanding AI and Human–Machine Communication

Overall, this study develops the understanding that "AI lying" is better conceived as a multilayered phenomenon rather than a single category. At the first layer is error or falsehood. At the second layer is the effect of deception, when an output can form false beliefs. At the third layer is users' attribution of intention. At the fourth layer is the question of whether AI possesses speaker status that would make it possible for the action to be called lying. This four-layer framework constitutes the conceptual foundation of the study.

The main contribution of the study is not to prove that AI has become human, but to demonstrate that the social meaning of communication does not always wait for certainty concerning the ontological status of the communicator. Users may experience AI as a communication partner before philosophy has determined whether AI is genuinely a "speaker." Thus, human experience of AI is an important domain for understanding the development of nonhuman communication.

Theoretically, the findings extend the discussion of Speech Act Theory by bringing the concept of speaker meaning into situations in which a communicative function is present without a clearly identifiable human subject. Empirically, the study demonstrates that deception should not only be examined as a system capability but also as an interpretive human experience. Practically, the study points to the need for AI literacy that includes the ability to distinguish error, misleading behavior, and lying while also understanding that convincing language does not necessarily indicate knowledge, intention, or consciousness behind it.

Ultimately, the question "Can Machines Lie?" receives a more complex answer than simply "yes" or "no." Machines can produce incorrect outputs, can exhibit behavior that is functionally deceptive, and can generate utterances that humans experience as lies. However, whether all of these phenomena deserve to be called lying still depends on questions of intention, speaker meaning, and AI's status as a communicative agent. It is precisely the tension between what users experience and what can philosophically be claimed about machines that constitutes the most important conceptual finding of this study.

CONCLUSION

This study demonstrates that users' experiences of incorrect or misleading artificial intelligence (AI) outputs cannot be understood simply as a matter of right and wrong. The four patterns of meaning identified—AI errors experienced as lying, attribution of intention to machines, tension between deception and lying, and ambiguity concerning AI's status as a

speaker—show that the meaning of “AI lying” is formed through an interpretive process in human–AI interaction.

An important insight from this study is that experiencing a lie and demonstrating that a lie has occurred are two different things. Users may genuinely feel misled or feel that AI is “lying,” but such an experience is not sufficient to demonstrate that AI possesses communicative intention in the human sense. Conversely, the fact that AI cannot readily be positioned as a subject possessing mental intention does not eliminate the reality that its outputs can produce social effects such as confusion, misunderstanding, and loss of trust. Thus, the phenomenon of AI lying should be understood as a multilayered phenomenon situated among falsehood, deceptive effect, intention, and speaker meaning.

Conceptually, this study enriches the use of Speech Act Theory in the context of nonhuman communication. The question concerning AI is no longer merely “can a machine generate language?” but develops into “who is speaking?”, “who possesses intention?”, and “to whom is communicative responsibility directed?” AI can perform communicative functions and can be treated by users as an information source or conversational partner, even though its status as a speaker with speaker meaning equivalent to that of humans remains problematic. Accordingly, this study proposes that AI’s communicative agency can be understood relationally: part of it emerges from the system’s capacity to generate responses, while another part is shaped by the ways in which humans interpret and respond to the system.

The practical implications of this study are particularly relevant to AI literacy in education. The ability to use AI should not stop at prompt-writing skills or checking the accuracy of information. Users also need to understand the difference between incorrect information, potentially misleading outputs, and lying, which presupposes intention. Such literacy is important so that students and AI users do not assign excessive epistemic authority to systems merely because their answers sound fluent, confident, and similar to human communication.

For policymakers and educational institutions, the findings indicate the need for AI-use guidelines that are oriented not only toward technical accuracy and safety but also toward transparency concerning system limitations, AI’s status as an information source, and human responsibility for verifying AI outputs. In digital literacy curricula, issues of AI deception and AI hallucination can be incorporated into critical-thinking education so that learners can question not only “is this answer correct?” but also “why do I trust this answer?” and “what is the basis for assuming that this system has a particular intention?”

These implications also have a social dimension. As AI becomes increasingly present in learning processes and everyday communication, the boundary between technology as a tool and technology as a “conversation partner” becomes increasingly blurred. Therefore, society needs forms of literacy that help users maintain a critical stance without denying the experience of interacting with AI. Users may acknowledge that a response feels convincing, disappointing, or even like a lie while simultaneously understanding that such an experience does not automatically demonstrate the presence of human-like consciousness or intention in the machine.

This study has several limitations that need to be considered. First, the study focuses on users’ experiences and is therefore not intended to establish the internal conditions, consciousness, or actual intentions of AI systems. Second, the depth of interpretation remains dependent on the experiences and narratives of the available participants; therefore, the findings cannot be treated as a universal representation of all AI users. Third, this phenomenon is strongly influenced by the type of system, context of use, technological experience, and users’ social situations. The research document also still lists the number of participants as “[number of participants],” meaning that sample characteristics cannot yet be used to make claims about population representativeness. These limitations indicate that the conclusions are better understood as contextual understandings rather than generalizations.

These limitations also open an agenda for future research. Subsequent studies could

broaden the context by involving participants from more diverse educational backgrounds, age groups, and levels of AI-use experience. Research could also compare users' experiences across different types of generative AI and interaction contexts, such as education, information seeking, public services, or professional communication. Phenomenological approaches could be deepened through longitudinal interviews to examine how trust and attribution of intention change after users repeatedly experience AI errors. Meanwhile, experimental or mixed-methods approaches could be used to connect users' subjective experiences with characteristics of AI outputs in a more systematic manner.

Furthermore, future research could explore unresolved questions raised by this study: does the concept of speaker meaning need to be revised when applied to nonhuman agents? Must deception always presuppose mental intention, or is it sufficient to understand deception through its communicative effects? And when an AI output misleads humans, should communicative responsibility be directed toward the machine, the developers, the institutions using AI, or the social structures that shape its use?

Ultimately, the principal contribution of this study lies in shifting the way the question "Can Machines Lie?" is understood. The question no longer concerns merely the machine's ability to produce false statements, but touches on more fundamental issues of meaning, agency, trust, and responsibility in human-machine communication. AI may produce language that is socially experienced as lying, but this experience simultaneously forces humans to reconsider what it means to "speak," "intend," and "lie" when the communicator is no longer fully human. It is within this tension that research on AI lying becomes relevant—not because machines have become human, but because their presence forces humans to reconsider the boundaries of communication itself.

REFERENCES

- Arora, C. (2024). Proxy assertions and agency: The case of machine-assertions. *Philosophy & Technology*, 37, Article 28. <https://doi.org/10.1007/s13347-024-00703-5>
- Baek, T. H., & Kim, M. (2023). Is ChatGPT scary good? How user motivations affect creepiness and trust in generative artificial intelligence. *Telematics and Informatics*, 83, 102030. <https://doi.org/10.1016/j.tele.2023.102030>
- Birhane, A., & McGann, M. (2024). Large models of what? Mistaking engineering achievements for human linguistic agency. *Language Sciences*, 106, 101672. <https://doi.org/10.1016/j.langsci.2024.101672>
- Black, J. (2024). Can AI lie? Chatbot technologies, the subject, and the importance of lying. *Social Science Computer Review*. <https://doi.org/10.1177/08944393241282602>
- Blut, M., Wang, C., Wunderlich, N. V., & Brock, C. (2021). Understanding anthropomorphism in service provision: A meta-analysis of physical robots, chatbots, and other AI. *Journal of the Academy of Marketing Science*, 49, 632–658. <https://doi.org/10.1007/s11747-020-00762-y>
- Cheng, X., Zhang, X., Cohen, J., & Mou, J. (2022). Human vs. AI: Understanding the impact of anthropomorphism on consumer response to chatbots from the perspective of trust and relationship norms. *Information Processing & Management*, 59(3), 102940. <https://doi.org/10.1016/j.ipm.2022.102940>
- Choudhury, A., & Shamszare, H. (2023). Investigating the impact of user trust on the adoption and use of ChatGPT: Survey analysis. *Journal of Medical Internet Research*, 25, e47184. <https://doi.org/10.2196/47184>
- Farquhar, S., Kossen, J., Kuhn, L., & Gal, Y. (2024). Detecting hallucinations in large language models using semantic entropy. *Nature*, 630, 625–630. <https://doi.org/10.1038/s41586-024-07421-0>
- Gubelmann, R. (2024). Large language models, agency, and why speech acts are beyond them (for now): A Kantian-cum-pragmatist case. *Philosophy & Technology*, 37, Article 32. <https://doi.org/10.1007/s13347-024-00696-1>
- Hagendorff, T. (2024). Deception abilities emerged in large language models. *Proceedings of the National Academy of Sciences*, 121(24), e2317967121. <https://doi.org/10.1073/pnas.2317967121>

- Hall, A., & Mazzeella, D. (2023). Pragmatic inference, levels of meaning and speaker accountability. *Journal of Pragmatics*, 205, 92–110. <https://doi.org/10.1016/j.pragma.2022.12.007>
- Lin, Z., Guan, S., Zhang, W., Zhang, H., Li, Y., & Zhang, H. (2024). Towards trustworthy LLMs: A review on debiasing and dehallucinating in large language models. *Artificial Intelligence Review*, 57, Article 243. <https://doi.org/10.1007/s10462-024-10896-y>
- 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(5), Article 100988. <https://doi.org/10.1016/j.patter.2024.100988>