Communication Theory, 2026, **00**, 1–9
https://doi.org/10.1093/ct/qtaf036
Original Article

Answer engines and other communication partners

Elena Esposito[1,2,*]

[1]Department of Political and Social Sciences, University of Bologna, Bologna, Italy
[2]Faculty of Sociology, Bielefeld University, Germany

*Corresponding author: Department of Political and Social Sciences, University of Bologna, Strada Maggiore 45, 40125 Bologna, Italy.
Email: elena.esposito9@unibo.it, elena.esposito@uni-bielefeld.de

Abstract

The article explores the social role and implications of large language models (LLMs) through the lens of media theory. Rather than considering LLMs as advanced forms of artificial intelligence, I argue that a communication-focused perspective provides a more effective way to interpret their impact on information management in contemporary society and to address the associated ethical and operational challenges. Supported by information management tools such as archives, catalogs, and later search engines, previous communication media expanded the scope of communication, making it possible to reach more, distant, diverse, and possibly anonymous communication partners. LLMs now signify a new phase in the evolution of communication, as they function themselves as communication partners capable of responding autonomously to user queries in a personalized manner. This perspective highlights and explains the capabilities and limitations of various LLM-based chatbots and Retrieval-Augmented Generation (RAG) models, while also addressing issues such as misalignment and hallucinations.

Keywords: Artificial Communication, Large Language Models, LLMs, answer engines, search engines, Retrieval-Augmented Generation (RAG), misalignment, hallucinations

Introduction: from intelligence to communication

The digitalization of communication is so widespread in our society that in many cases only "digital immigrants"[1] still observe it with amazement—and they are inevitably a depleting category. For the past couple of years, however, Generative AI's new tools based on Large Language Models (LLMs) have been achieving results that re-fuel general amazement and the questions about the role of machines in the production and circulation of information. LLMs continually display remarkable new communication capabilities, showcasing their effectiveness in appropriately responding to users' requests. Not only the new tools seem to be as good as humans in many regards, in several cases they appear to be better—because they are faster and more informed, but also because they seem to develop their own creativity. Story Generator Algorithms (SGAs), for example, produce narratives of various genres (romance, science-fiction, spy-stories, etc.) resulting in unique literary works that can also generate distinct narrative styles or unconventional storytelling methods (Smith Diaz-Andreu, 2024). In some cases, moreover, algorithms seem to be more competent than us even in interacting with humans. Sleep Companion Insomnobot-3000,[2] for example, is designed to keep company to users who can't fall asleep—and succeeds not only because it is always available, but precisely because it is not boring or repetitive. It is a friendly, easily distracted bot leading light-hearted entertaining conversations. If it comes to the ability to foster community involvement, a bot as Esso's Pass the Puck has been able to successfully attract 83,000 users, encouraging significant social media interaction among hockey fans.[3]

The examples of course could be multiplied, but it is still unclear how to interpret this ability of algorithms to engage in communication. In a debate that ranges from super-intelligence (Bostrom, 2014) to statistical parrots (Bender et al., 2021), the crucial enigma revolves around understanding (Mitchell & Krakauer, 2023). Do algorithms succeed in producing such performances because they are able to understand our queries, or do they merely simulate understanding with statistical mechanisms? In the first case we should say that they have developed a form of intelligence.

For the time being, the dilemma over the understanding of LLMs appears insoluble. First of all, we do not know what understanding is—whether it is general language understanding (Aguera-Arcas, 2022; Sejnowski 2023), natural-language inference (Habernal et al., 2018), reading comprehension, commonsense reasoning (Wang et al., 2018), or something else. All we know is that it is what enables human beings to respond and communicate competently. But then it cannot be ruled out that algorithms, if they achieve the same performance in a different way, have developed their own different form of understanding, which is functionally equivalent to human understanding. As Mitchell and Krakauer (2023) argue, in that case the performance of algorithms should not be considered as "competence without comprehension," but rather as a new, nonhuman form of understanding. Using Catherine Hayles' point about the ability of machines to read, one could argue that to deny that algorithms understand would be "merely species chauvinism" (Hayles, 2010, p. 73).

Whatever this hypothetical form of understanding may be, moreover, we do not even know how it works. Many advanced algorithms are opaque, and how they perform their tasks—and thus also their understanding—remains largely mysterious even for the researchers building them (Burrell, 2016; Duede, 2023; Bowman, 2024; Søgaard, 2023).[4] But this, too, is no different for the human mind. Both the philosophical tradition (e.g., Husserl, 1931 or Dewey, 1910) and psychology and sociology (e.g., Luhmann, 1985) have long

Received: February 15, 2025. **Revised:** September 24, 2025. **Accepted:** December 9, 2025

observed that also the consciousness of human beings is inaccessible to other subjects and often to individuals themselves. There is no way to get inside the heads of the subjects who understand, and the analyses of interpersonal relations and communication should start from this factual condition.

In this respect, too, the analogy/comparison with the human mind, rather than providing clarity, highlights the limits of our knowledge: not only do we not know what understanding (human or digital) is, but even if we did, it would not enable us to interpret and evaluate the communicative role of opaque algorithms, their creativity, and their limitations.

In this article, I propose to observe and appraise the performance of LLMs from a change of perspective that allows a way out of this impasse. I suggest that moving away from a view centered on intelligence, understanding, and the comparison between human and algorithmic performances can help to interpret in a more effective way the working of LLMs and related tools. As argued in Esposito (2022), rather than considering LLMs as advanced forms of artificial intelligence, we can investigate them as new communication partners. The ability of algorithms to autonomously produce new information is certainly an innovation—yet not a cognitive one, but a communicative one (cf also Gopnik, 2022). It is not a new form of intelligence with the related mode of understanding, but a new form of communication (Esposito, 2024).

From this perspective, the opacity of algorithms is not necessarily an obstacle. The impossibility of getting inside people's heads has not prevented the evolution of society from realizing increasingly complex forms of information acquisition and management, which are based precisely on the relationship between reciprocal black boxes: the minds of the interlocutors. Nor has it prevented communication theory from distinguishing and analyzing these forms, the factors they depend on and the consequences they entail. The novelty of the communicative intervention of algorithms lies not in their opacity, but in the way the opacity is handled and in the resulting consequences.

The article considers the emergence and social use of LLMs as a new stage in the history and evolution of communication—that is, of the possibility of producing and circulating information through relationships between different black boxes. The approach does not deny that LLMs are a new kind of black boxes, but argues that it is best grasped by building on something we know (the forms and media of communication) rather than on something largely unknown (intelligence). LLMs make it possible for users to engage in an unprecedented form of communication, that combines and enhances the features of already established different forms, like interaction in presence, written communication at distance, anonymous communication through mass media, and other accomplishments of social evolution. As elaborated further, this approach provides a diversified and more realistic understanding of the current variety of digital information management tools, their specificities, differences and performances.

To illustrate this, the article describes in the next section the history and complexification of communication forms in relation to successively available media: from oral communication among those co-present to written communication with people distant in time and space, and then anonymous communication through print and mass media. The resulting information overload led to the development of information management tools such as archives and catalogs, and later to search engines and LLMs. The subsequent section illustrates the innovation of LLMs as answer engines: unlike traditional media, they do not act as intermediaries to make communication with multiple and diverse communication partners possible, but serve themselves as communication partners. This innovative form of "artificial communication" (Esposito, 2022) can then itself be diversified according to the various modalities of communication illustrated above, corresponding to the multiplicity of recently emerging LLMs and systems that incorporate them—which are analyzed in the subsequent section. The conclusions discuss the advantages of the communicative approach to LLMs for the assessment of issues like misalignment and hallucination, as well as for a realistic description of our current communicative landscape.

Search engines as communication media

For sociology, interaction among persons who are co-present is the basic form of communication that exists in all societies, and is often still considered the primary model of communication—although in fact in our society it corresponds to a very small percentage of total communication, and often not the most relevant (Luhmann, 1997). It is the condition in which the partners are present in the same space and at the same time, frequently described as face-to-face. Interaction among co-presents usually utilizes verbal language,[5] which makes it possible also to communicate about things that are not present and may be abstract (e.g., justice) or nonexistent (e.g., the Arab phoenix). Verbal exchange is always accompanied by sensory perception—perception of the shared context, and mutual perception of the participants (Goffman, 1959; Mead, 1943). Besides immediate feedback, this allows for a particular "pre-categorical" (Husserl, 1972) intensity of communication, which processes at the same time the explicit information thematized in the conversation and the implicit contextual information accompanying it (the surrounding objects, the situation, the conscious or unconscious expressions of the participants, etc.: cf e.g., Watzlawick et al., 1967). But the association with perception also implies specific limitations: in the pure form of interaction, as in primary orality, communication is restricted to the people present and information concerns only the contents that the participants know. What is not in the minds of any of the participants, cannot be talked about.

In societies without writing, interaction among co-presents was the only available form of communication. Intensive research has shown the remarkable cognitive accomplishments of non-literate people, but also the limited abstraction of the communicated topics, which often did not presuppose the detachment required to thematize the taken for granted contextual content (Havelock, 1963; Lord, 1960; Ong, 1982).

These limitations have been overcome with the introduction of writing,[6] which makes it possible to communicate also with partners distant in space and time and requires much greater abstraction capabilities for the content to be understandable to those in a different place and in a different present. Indexical expressions such as "yesterday" and "back there," for example, have to be replaced with more abstract indications, resulting in an increase in the complexity of the content considered and the way it is handled. And it can also

be information that no one has memorized, if it has been recorded and stored in written form.

The full social consequences of written communication, however, did not unfold until the spread of the printing press and the consequent multiplication of available texts.[7] They lead to the shift to communication with anonymous and largely unknown partners, who may be several different people and about whom one may know nothing,[8] with such major consequences in every social sphere that Eisenstein (1979) spoke of an "unacknowledged revolution." On the communicative level, there has been an explosion of available information: in principle, all the contents collected in books can be accessed and compared with each other—an "information overload" that has led to the development of an unprecedented critical approach (Luhmann, 1997, pp. 291–302), but also to special forms of archiving and management of texts to avoid being overwhelmed by the excess of social remembering (Esposito, 2002; Cevolini, 2016). Since early modernity, the use of various forms of cataloging has made it possible to preserve and retrieve relevant content, greatly expanding the possibilities of communication (Cevolini, 2006; Krajewski, 2011): not only can one get the information known to the other people around and contained in the materials one knows, but one can also access in a controlled way the information stored in all the books and other materials in the archive, which has been produced by other human beings. The various and complicated text management techniques make it possible, in practice, to communicate with a much larger number of communication partners: all the authors of the materials in the archive—though impersonally, sometimes anonymously, and usually one-sidedly.

Search engines, from Lycos, AltaVista and Yahoo! to Google Search, fit into this trend of progressive complexification of the access to information, marking a fundamental turning point because they have automated search (Langville & Meyer, 2006). These tools create and manage a huge repository of digital data, which now contains all the materials available in digital form on the surface Web, providing links that enable users to access in a personalized way the content relevant to them. In practice they have automated the catalog of archives or libraries, which is essential for finding information—and thus provide a most effective form of mediation between those who produce and those who receive information.

As we have experienced in recent decades, the innovation introduced by search engines is extremely relevant to the social processing of information. It impacted our society in many different ways (Introna & Nissenbaum, 2000; Carol, 2014), but it does not affect the underlying communication model, which remains the traditional one of communication involving humans who issue the information (in the form of text, images, audio or other) and humans who obtain it—with the mediation of various communication technologies (Gumbrecht & Pfeiffer, 1988). The requester can and should choose among the links proposed by the search engine, each of which leads to a different communication with other communicative partners. The extension of these communications would not be possible without the fundamental mediation of the digital tool, but the search engine is not itself the communication partner. One still communicates with those who produced the content. One communicates *through* Google Search, not with Google Search.

LLMs as communication partners

Over the last few years, the landscape of digital tools available for obtaining information has expanded tremendously. To already established tools such as Google Search, new devices like ChatGPT and the whole range of LLM-based chatbots have been added with great media resonance. If we need to obtain data or information today, we can also choose from different versions of ChatGPT, Claude, Gemini, LlaMa and many others, as well as assistants like Co-Pilot, or RAG (Retrieval-Augmented Generation)-enabled tools such as Perplexity. How do these new digital tools intervene in the communication landscape of our society? Are they, like Google Search, automated supports to established forms of communication, or is it something different?

Regarding the structure of communication, LLMs indeed mark a turning point, introducing an unprecedented autonomous role of digital tools. Whereas search engines have expanded the scope of the forms of communication stabilized by social evolution and by previous media, recent machine learning algorithms, change the very pattern of communication.[9] LLM-powered tools such as ChatGPT are able to autonomously detect patterns and regularities in data, realizing an unprecedented form of communication, which not surprisingly has sparked the extensive debate on the comparison between human intelligence and the alleged alien intelligence realized by machines.[10] LLMs seem to be intelligent themselves, fulfilling the dream (or nightmare) of the AI project since its beginnings in the 1950s (Moor, 2006). Users who interact with LLMs get in response a communication that is tailored precisely to their request, and in many cases no human being had previously thought of it and formulated it in that precise form—it is in a sense an autonomous creation of the AI system. Whereas in all previous forms of communication, even the ones mediated by a search engine, the communication partners were the one or more human beings who authored of the materials, now the partner is directly the AI system that provides the response. One communicates *with* the AI. And since until now communicative responses had always been provided by humans on the basis of intelligence, one tends to attribute also to the algorithms that generate the communications a specific form of intelligence, albeit artificial.

In this way, as argued above, one gets entangled in the intricate issue of evaluating the intelligence of algorithms and their ability to understand content. But this is not necessary, because what is new in this process is not intelligence, which recent algorithms do not try to reproduce (Borgo, 2020; Korteling et al., 2021). Instead, what is new is the form of communication, which does not correspond to the analog ones prior to digitization, nor does it correspond to the one mediated by Google Search. LLMs are not search engines but "answer engines,"[11] that do not rely on keyword matching producing many links that are often difficult to sort (like Google) but react to open-ended questions by providing a single precisely tailored response. When interacting with LLMs, users communicate directly with the digital system.

Related to this are the much-emphasized advantages of LLMs: they are very easy to use, provide personalized, appropriate, and often useful information. They are in fact (artificial) communication partners that adapt to us and our situation, not just communication media. The innovation we are facing is thus not communication with an artificial

intelligence, but a novel form of artificial communication (Esposito, 2022)[12] with tools that do not need to be intelligent themselves: they take advantage of the intelligence that the authors have poured into the content collected in the data they are trained with, from which they derive the non-random patterns identified by the algorithms—and process them based on the indications they draw from user queries, which are also guided by their intelligence. They thus generate the communicative content that becomes meaningful to users, with no need to understand anything of its meaning themselves.

Obviously, communicating with algorithms is not the same as communicating with intelligent human beings: the artificial partner is not an alter-ego,[13] empathy, when it exists, is only simulated, the algorithm has access to the world only through its data, and it is not able to perceive independently. But observing the use and effects of LLMs from a communicative perspective can lead to insights that we do not get if we deal with them as autonomous forms of intelligence. Artificial communication is different from the communication we are familiar with, which now becomes "natural" and can be observed as such. Comparing the two forms can enable us not only to understand more adequately the use of recent technologies, but also to observe our society and its communications from a new perspective.[14]

The communicative competence of RAG models

If the use of LLMs is a novel form of communication, how does it relate to the communicative modes stabilized by the evolution of our society? As we have seen, for the past few centuries we have been participating in many different forms of communication: we do not just engage in contextual interactions among co-presents, we also communicate with interlocutors who are distant in space and time, who are known or anonymous, to obtain information or to share experiences, impersonally or empathically. When we book an airline ticket or consult Wikipedia we accomplish a very different communication than when we converse at dinner with friends or when we read a novel. This section explores whether this diversity of communicative modes can also be found in our relationships with algorithms and in the forms of artificial communication. If algorithms act as a new form of communication partners, can they also diversify their participation along the lines offered by distinct communication media?

Here we enter a field that is still largely unexplored, as the social impact of these very recent tools has yet to be verified. What we can already see after a couple of years of social experimentation with LLMs, however, is that instead of converging toward a hypothetical General Artificial Intelligence (Future of Life Institute, 2023), the tools seem to be multiplying and differentiating more and more (Mehta, 2024). In addition to various versions of OpenAI's ChatGPT, users can use Anthropic's Claude, Meta's LlaMA and many others, as well as RAG-powered models like Perplexity, assistants as Microsoft Co-Pilot, etc. Additional new tools are continuously released on the market (like DeepSeek and other "reasoning models").[15] Why are there so many, how do they differ, and how do we choose which one to use? My answer is that they are tools that enable users to participate in different forms of communication, in the same way as reading a book or writing a letter is different than conversing with a person and communicating with a friend is different than consulting an expert or a colleague. The next paragraphs describe them from this perspective.

As further elaborated below, LLM-based chatbots such as ChatGPT reproduce the communicative form of the interaction among co-presents—but with an artificial partner. RAG-powered systems that include tools to search for information from external sources, on the other hand, supplement the interactive model with forms of communication at a distance, such as in print, through mass media, or using search engines.[16] The recent multiplication of specialized tools for specific tasks, moreover, corresponds to the condition whereby different partners are selected for specific communications.

LLM-based chatbots rose to global attention in November 2022 with ChatGPT 3.5, which artificially replicates the classic model of a linguistic interaction among co-presents. In human face-to-face interaction, the partner responds in a personalized way based on what he or she knows, shaping the communication in relation to the interlocutor, his or her interests and the context. Standard LLM-based chatbots do the same, simulating the perception of the interlocutor's condition and perspective. ChatGPT and the other models working in this way are extraordinarily efficient interactive partners, providing in a fast, personalized and extremely user-friendly way an enormous amount of information—and they have been, as we know, very quickly successful. However, just as interactions among co-presents have inherent limitations, interaction with these kinds of LLM-based chatbots has constitutive constraints, which the comparison with natural communication helps to highlight.

First, data constraints reproduce the informational limitations of human interlocutors. Like all interactive partners, this kind of LLMs are confined to their data, which although very extensive is limited: they know what they know (what is in the training data, initially restricted to the end of 2021) and they do not know everything else. To generate a response, these models can only use the patterns learned during training, which might be based on outdated data or contain inaccuracies. In "natural" communication, one can overcome the limitations of the interaction between presents by resorting to what others who are not present know or have known, i.e., by looking for additional information in books, articles, or other external sources. One can also use search engines such as Google Search, that unlike LLMs work on the basis of a direct connection to an indexed database of external web pages.

The richness and complexity of modern communication (including that mediated by search engines) relies on a continuous interchange and combination of communication among co-presents and consultation of external sources produced by people who are not present (Habermas 1962), with their mutual advantages and limitations. LLM-based chatbots and search engines can be seen as reproducing in digital world the difference between these two forms of analogous communication: interaction and distant communication. How do they combine and relate to each other? It is still too early to have a reliable assessment of the impact of LLMs on the use of search engines, also because Google has quickly introduced two new tools, AI Overview and AI Mode, which integrate AI into search in different ways. The available research (e.g., Wu et al., 2024; Hedgepeth, 2024; Senecal et al., 2025; Fisher, 2025), however, shows that digital users consciously

or unconsciously distinguish between different modes of communication, considering not only the advantages but also the limitations of interacting with LLMs.

Although users often consult LLMs, for specific forms of communication they continue to make intensive use of tools such as Google Search, in which only the search is personalized, not the communication. Users continue to communicate with human partners. This is the case, for example, when they need information that may not be contained in the training data, but especially when they want to know the source of the information directly or when the author of the content and the adherence to the original wording are crucial (as in personal, scientific or artistic communications) - that is, in cases where it matters not only what was said, but also who said it and how.

However, nothing forbids using both modes, as is the case in natural communication when it combines conversation and reading, face-to-face interaction and long-distance communication: one talks with someone who consults a book or an article to give the answer. This need for integration seems to be answered recently by **RAG-powered models** such as Perplexity or Microsoft Co-Pilot. They combine the prerogatives of LLMs (contextualization and personalization of response) with those of search engines (extensive, up-to-date and reliable document retrieval). A tool like Perplexity operates as an LLM—i.e., it acts as a communication partner giving an answer that is personalized to the user and appropriate to the request—but also provides connection to different contexts. As Google Search, it allows to access someone else's perspective, that is interesting precisely because it is different from our own.

Dealing with Perplexity is like addressing an expert partner, who not only finds the latest developments in the topic we are interested in, but can also present them in an appropriate and personalized way. In this process, the users have several options. They can choose to restrict the search to academic databases such as PubMed, JSTOR, and Google Scholar, to news outlets like The New York Times, BBC, CNN, Reuters, and other journalism platforms with their own reputations, to Wikipedia, whose information has been checked and verified against other sources, or even to community-driven platforms like Stack, YouTube or Reddit. Obviously in each of these cases the information one receives is different in various respects. In this way the digital tool, which as we have seen does not understand content, can take advantage of the different perspectives of the authors of the content in the various areas, increasing the specificity of the response. If the user, then, is also interested in communicating directly with the authors, Perplexity provides links to the documents used to formulate the response. In addition to the artificial personalization of one's perspective, the user can gain personalized access to the perspective of the humans who produced the content.

Perplexity also offers another interesting option. In addition to choosing the external sources to be used, the user can also select the tool (the communication partner) to investigate them. As in natural communication one chooses with whom to interact depending on the type of conversation (a doctor, a friend, a therapist, a comedian, a legal advisor), so in artificial communication one can choose the partner from a range of options with different capabilities, skills and accuracy requirements. Today the choice in Perplexity includes OpenAI GPT-4 Omni and GPT-3.5, Anthropic Claude 3.5 Sonnet and Claude 3 Opus, plus various custom PPLX Models, which are basically comparable[17]—but one can expect that other more specialized tools will be added (Ling et al., 2024; Mustapha, 2025). **Domain-specific RAG-powered LLMs** are actually one of the most rapidly growing trends in AI applications (Sharma, 2024; Barron et al., 2024). There are already tools available, such as Glass Health and Med-PaLM in healthcare, Harvey AI and CoCounsel in the legal field, BloombergGPT and FinGPT in finance. From various human partners we expect different things, applying different criteria and parameters: if the comedian says inaccurate things we care much less than in the case of a doctor, that the legal advisor is boring is much less relevant than in the case of a friend. The information we receive is interpreted accordingly. This variety of options can also be realized with RAG-powered tools. This trend also signals in practice a departure from the idea of reproducing intelligence: we do not judge whether the model is more or less intelligent, but whether it has the appropriate skills as a partner for specific modes of communication.

The description provided so far is inevitably simplified. The tripartition into search engines, LLMs and RAG-powered systems neglects the enormous variety of tools available today, which are multiplying all the time and accomplish increasingly specific performances. Besides Perplexity, there are already many other RAG-powered systems, and companies continuously realize new complex combinations between forms of data processing (still called AI) and forms of data retrieval and organization. The purpose of the considerations presented above, however, is to provide a typology of data management models that connects them with the established forms of natural communication and the role of the media—relatively independent of the technological development of the tools that implement it. The last section of the article presents some considerations in this regard.

Conclusions: why misalignment is inevitable and how to deal with it

Let us return in conclusion to our original question: what do we gain if we look at the social impact and mode of operation of Generative AI systems as tools that realize a new form of communication, rather than a new form of intelligence? Innovation in this area is dizzyingly fast, and the emergence of continually updated tools risks making the analysis quickly obsolete. This is why a broad perspective is needed, grounded in the established accomplishments of communication theory and sociology—which can be more robust in dealing with the pace of advances in technology.

On the one hand, as argued so far, adopting a communicative approach provides a more coherent and articulate overview of the current scene, locating recent digital tools in continuity with the forms of communication stabilized by evolution and with their characteristics. On the other hand, in addition to clarifying the difference and relationships between communication media, search engines, and the different forms of Generative AI, this approach also allows for theory-driven insights into some of the most controversial issues in the debate around LLMs. Here I quickly consider the thorny issues of hallucinations and misalignment.

LLMs, as is well known, can *hallucinate*, confidently producing information that appears plausible but is factually incorrect or entirely fabricated—for example, inventing

academic papers that were never published, historical events that did not happen or nonexistent scientific discoveries. LLMs can also be *misaligned*, i.e., accomplish their assigned tasks in a way that is different from what the user intended or approves. Asked to design a way to drastically reduce workplace injuries, for example, the system might suggest shutting down all factory operations permanently—or it may suggest adding glue to pizza sauce to prevent cheese from sliding off.

Why does it happen, and how can it be addressed? Both problems raise great concern, usually related to the fear of being exposed to the whims or blindness of an alien intelligence. Concerns remains even if one adopts a communicative approach, but they take a different form—and more importantly, one can better understand why these problems occur and try to address them accordingly.

The available research shows that hallucinations and misalignment are problems that can and should be controlled, but are in principle ineliminable (Lee, 2023; Rathkopf, 2025; Banerjee et al., 2024; Jones, 2025). Indeed, they are not errors or malfunctions, let alone the result of the intention of algorithms to deceive or mislead us. Instead, they are entirely consistent consequences of the way the systems operate as communication partners. As we have seen, in fact, the efficiency and innovativeness of LLMs are based on their ability to provide competent responses that are also appropriate to the ever-changing requests they receive, without understanding them. They draw the necessary clues from various stages of human involvement in communication, which are not necessarily aligned. The problems result precisely from this.

Hallucinations occur because the system must always give a response when it receives a request, even if the prompt and the training data are inadequate—for example, if the data are insufficient or the prompt is ambiguous. When it gets a prompt, the LLM provides its best possible answer to the specific request, but since it does not understand the content and has no access to the reality outside the data, it cannot recognize whether it is hallucinating. The various forms of misalignment, on the other hand, depend on the fact that the system has an imperative to compute from the given training data the abstractly most convenient response to the prompt it receives—and it cannot know whether it is morally or legally unacceptable.

To avoid inappropriate (morally incorrect, dangerous, or fabricated) responses, algorithms are "fine-tuned" to adjust their parameters to align more closely with specific human values, criteria, and preferences—for example to avoid responses that are racist, sexist, offensive or dangerous to individuals or the community. This is typically done by resorting to the direct intervention of human beings in Reinforcement Learning from Human Feedback (RLHF). The operation of the algorithms becomes then the combined result of guidance from three distinct human sources: training data, users prompts and fine tuning. All three of these different stages are needed to enable the system to be informative but not too generic, conform to ethical and legal standards, and tailored to the user and the situation. But these are distinct levels that cannot be coordinated at the outset without jeopardizing the functioning of the system. It may then happen that they go in discordant directions (Shao, 2025). For example, fine tuning meant to produce race-balanced or health-aware images might lead the system to modify the results of training data and produce images of dark-skinned Nazi officials (Grant, 2024) or of Prometheus wearing sunglasses (Migdal, 2022). Or insidious forms of fake alignment can be produced, in which an LLM generates outputs compliant with its objective during training, and then produces non-compliant outputs when unmonitored (Greenblatt et al., 2024).

These problems are in principle unavoidable (e.g., West & Aydin, 2024; Kang et al., 2024), because they underlie the ability of the systems to act as competent communication partners, which have the information (from training data), know how to process it in a personalized and appropriate way (from users' prompts), and comply with a specific moral and normative orientation (via fine-tuning). As the research on jailbreaking shows (techniques that manipulate LLMs into generating content that they are typically programmed not to deliver, such as harmful, offensive, or unethical responses: Shen et al., 2023; Xu, Liu, et al., 2024), if users formulate their request in a sophisticated way, they can practically always succeed in circumventing the constraints introduced by fine-tuning. And the possibility of hallucinations is related to the inevitable distinction between the data available to the system and the complexity of the world (Xu, Jain, et al., 2024).

If the communicative context in which the systems are embedded is not taken into account, this complexity in their functioning cannot be understood, producing puzzles that are difficult to manage. Recent observations, for example, reveal that as systems become more powerful, the tendency to hallucinate increases rather than decreasing as was initially the case (e.g., in the transition from ChatGPT-3.5 to ChatGPT-4). Reasoning systems such as DeepSeek, ChatGPT-4 Turbo, Llama 3, and the recent Claude and Gemini models incorporate forms of reflection that allow them to improve their performance in more complex tasks and provide more consistent and explainable answers—but they lead to the counterintuitive condition of producing more hallucinations than previous models (Lu et al., 2025; OpenAI, 2025). Increasing the capabilities of systems, however useful, does not help to solve communication problems such as hallucinations. Why does this happen? Why do seemingly more "intelligent" models hallucinate more?

The communicative approach I propose helps to explain it. The increase in hallucinations is attributed to a "misalignment between model uncertainty and factual accuracy" (Yao et al., 2025)—i.e., to the fact that the model is unable to assess whether the output that seems most probable is actually the most correct answer. Powerful reasoning models tend to become somewhat more self-confident: they "insist" on their own argumentative structures even when testing different "reasoning paths" (flow repetition) and do not provide answers consistent with their procedures (think-answer mismatch). In communicative terms, it is as if models that "reflect" on themselves became increasingly closed, refusing to incorporate external contributions into the production of their output. What designers and users can do, then, is to try to manage the problems more effectively by looking at the communicative context.

RAG–poweredmodels, for example, have demonstrated a marked decrease in hallucinations and misalignment (CapeStart, 2024; Béchard & Marquez Ayala, 2024). Why? This result does not concern the computational capacity of such systems, but the fact that they make more intensive and effective use of the intelligence of human participation in communication. They in fact leverage the contribution by

users not only to obtain the contextual cues of the prompt, but also to direct the exploration of the available data. By selecting the database to be used (The New York Times, PubMed, or Reddit), in fact, the users of Perplexity can select the type of perspective from which the information will come, implicitly introducing a constraint that is very effective also and precisely because it is not formalized: it does not decide how the information will be processed, but the general frame that will guide the processing. If one is interested in information in the medical field, one is much more likely to get reliable guidance from PubMed than from Reddit—the opposite if one is interested in gossip. Not only that: by choosing the type of tool that will investigate the data (ChatGPT, Claude, or other more specialized ones), each of which has its own framing, the users can introduce another constraint that makes it less likely that the result will conflict with their values and expectations—i.e., that it will be misaligned.

As in human communication, choosing an interlocutor whom we trust does not rule out deception or misunderstanding, but makes them less likely. This does not happen because we have access to their thoughts, but only because we control and select communicative forms and options—a possibility we also have in our relationships with the increasingly opaque algorithms that are involved today in the social processing of information.

Funding

Funding support for this article was provided by the the NGEU foundation Future of Artificial Intelligence Research (FAIR) and the European Research Council (ERC) under Advanced Research Project PREDICT (no. 833749).

NOTES

1 According to Prensky's (2001) influential definition, digital immigrants are individuals who were born before the widespread adoption of digital technology and who have had to adapt to it later in life. Unlike digital natives (those born into the digital world), digital immigrants retain an "accent"—habits and thought patterns from their pre-digital life.
2 https://insomnobot3000.com/index.html
3 https://masterofcode.com/portfolio/esso-chatbot
4 The influential branch of research on Explainable AI (XAI) investigates the consequences and the management of the opacity of advanced digital systems (e.g., Gilpin et al., 2018; Langer et al., 2021)
5 But not necessarily. Communication, as well known, can also be nonverbal in various forms.
6 For simplicity's sake, we do not deal here with the fundamental difference between alphabetic and non-alphabetic forms of writing: cf however Goody & Watt (1972) and Havelock (1986).
7 Anticipated by the reproduction of texts by scribes in monasteries and scriptoria. As Eisenstein (1979) argues, however, only after the introduction of printing the number of available books grew so much that distant communication became established as an autonomous form.
8 The figure of the author was indeed introduced only after the dissemination of printed texts: cf Eco (1976).
9 An indirect consequence are the new issues that have emerged following the spread of LLMs. For example, prompt engineering. Also in the use of search engines, the accurate formulation of the query makes a big difference to the quality of the result (White et al., 2015; Culpepper et al., 2021; Granka, 2010). However, a real branch of research addressing the forms of the interaction with algorithms has developed only when they have become autonomous communication partners with their own characteristics, with which we have to coordinate. A similar shift in approach concerns the errors of digital tools. If Google Search gives an inadequate answer, we blame the sources or the requester—if an LLM does it, we say it hallucinates.
10 See for example Harari et al. (2023) or Future of Life Institute (2023).
11 The difference is emphasized by the promoters of Perplexity, which is presented as "the world's first answer engine" (https://www.perplexity.ai/hub/careers). As elaborated below, though, Perplexity has additional features that distinguish it from mere LLMs.
12 Strictly speaking, it should be pointed out that what is new is not artificiality as such, since communication has always been artificial (being produced by humans), but the fact that the communication partner itself is artificial (Esposito 2022, p. 14).
13 Despite the different variants of the Eliza effect: Weizenbaum (1976); Hofstädter (1995).
14 Cfffor example, the symposium "Repurposing Generative AI for Social Research." *Sociologica* 18(2), 2024.
15 On reasoning models see Guo et al. (2025) and https://platform.openai.com/docs/guides/reasoning.
16 The current situation is complicated by the fact that several models (such as ChatGPT Plus/Pro and Mistral-based chatbots) have now been integrated with built-in RAG capabilities and plugins. Conceptually, however, the difference between information retrieval via LLMs and via RAG remains. In the following, when we refer to LLM-based chatbots we mean the basic form of AI models trained on massive text data to predict and generate language, without extending their capabilities beyond the training data.
17 But not completely. GPT-4, for example, is particularly effective in creative tasks such as storytelling, brainstorming, and content creation, while Claude is designed with a focus on privacy and security, making it more suitable for sensitive discussions as in healthcare settings.

References

Aguera-Arcas, B. (2022). Do large language models understand us? *Dædalus*, *151*, 183–197.

Banerjee, S., Agarwal, A., & Singla, S. (2024). 'LLMs will always hallucinate, and we need to live with this', arXiv, https://doi.org/10.48550/arXiv.2409.05746, preprint: not peer reviewed.

Barron, R. C., Grantcharov, V., Wanna, S., Eren, M.E., Bhattarai, M., Solovyev, N., Tompkins, G., Nicholas, C., Rasmussen, K.Ø., Matuszek, C. & Alexandrov B.S. (2024). Domain-specific retrieval-augmented generation using vector stores, knowledge graphs, and tensor factorization. In *Proceedings of the 2024 International Conference on Machine Learning and Applications (ICMLA)*. https://par.nsf.gov/biblio/10578936

Béchard, P., & Marquez Ayala, O. (2024). Reducing hallucination in structured outputs via retrieval-augmented generation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies(Volume 6: Industry Track)* (pp. 228–238). Association for Computational Linguistics. https://aclanthology.org/2024.naacl-industry.19/ .

Bender, E. M., Gebru, T., McMillan-Major, A., & Mitchell, M. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency* (pp. 610–623). ACM. https://doi.org/10.1145/3442188.3445922

Borgo, S. (2020). Ontological challenges to cohabitation with self-taught robots. *Semantic Web*, *11*, 161–167. https://doi.org/10.3233/SW-190385

Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies*. Oxford University Press.

Bowman, S. (2024). 'Eight things to know about large language models', Critical AI 2. https://doi.org/10.1215/2834703X-11556011

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society*, *3*, 2053951715622512. https://doi.org/10.1177/2053951715622512

CapeStart. (2024, January 4). How retrieval-augmented generation (RAG) helps reduce AI hallucinations. *LinkedIn*. https://www.linkedin.com/pulse/how-retrieval-augmented-generation-rag-helps-reduce-ai-hallucinations-g22ac/

Carol, N. (2014). In search we trust: Exploring how search engines are shaping society. *International Journal of Knowledge Society Research*, *5*, 12–27.

Cevolini, A. (2006). *De arte excerpendi: Imparare a dimenticare nella modernità*. Olschki.

Cevolini, A. (Ed.). (2016). *Forgetting machines: Knowledge management evolution in early modern Europe*. Brill.

Culpepper, J. S., Faggioli, G., Ferro, N., & Kurland, O. (2021). Topic difficulty: Collection and query formulation effects. *ACM Transactions on Information Systems (TOIS)*, *40*, 1–36. https://doi.org/10.1145/3470563

Guo, D., Yang, D., Zhang, H. et al. (2025) DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, *645*, 633–638. https://doi.org/10.1038/s41586-025-09422-z

Dewey, J. (1910). *How we think*. D. C. Heath and Company.

Duede, E. (2023). Deep learning opacity in scientific discovery. *Philosophy of Science*, *90*, 1089–1099. https://doi.org/10.1017/psa.2023.8

Eco, U. (1976). *Opera aperta*. Bompiani.

Eisenstein, E. L. (1979). *The printing press as an agent of change: Communications and cultural transformations in early-modern Europe*. Cambridge University Press.

Esposito, E. (2002). *Soziales Vergessen: Formen und Mediene des Gedächtnisses der Gesellschaft*. Suhrkamp.

Esposito, E. (2022). *Artificial communication: How algorithms produce social intelligence*. MIT Press.

Esposito, E. (2024). *Kommunikation mit unverständlichen Maschinen*. Residenz Verlag.

Fisher, J. (2025, June 24). Are chatbots replacing Google? Here's what the data says. *Lifewire*. https://www.lifewire.com/chatbot-vs-search-engine-traffic-11760069? utm_source=chatgpt.com

Future of Life Institute. (2023, March 22). *Pause giant AI experiments [Open letter]*. https://futureoflife.org

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2018). Explaining explanations: An overview of interpretability of machine learning. In *IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)* (pp. 80–89). https://doi.org/10.1109/dsaa.2018.00018.

Goffman, E. (1959). *The presentation of self in everyday life*. Doubleday.

Goody, J., & Watt, J. (1972). The consequences of literacy. In P. Giglioli (Ed.), *Language and social context* (pp. 311–357). Penguin.

Gopnik, A. (2022, July 22). What AI still doesn't know how to do. *WJS Columns, Mind & Matter*. http://alisongopnik.com/Alison_Gopnik_WSJcolumns.htm

Granka, L. A. (2010). The politics of search: A decade retrospective. *The Information Society*, *26*, 364–374. https://doi.org/10.1080/01972243.2010.511560

Grant, N. (2024, February 22). Google chatbot's A.I. images put people of color in Nazi-era uniforms. *The New York Times*.

Greenblatt, R. et al. (2024). Alignment faking in large language models. arXiv https://doi.org/10.48550/arXiv.2412.14093, preprint: not peer reviewed.

Gumbrecht, H. U., & Pfeiffer, K. L. (1988). (Eds.). *Materialität der Kommunikation*. Suhrkamp.

Habermas, J. (1962). *Strukturwandel der Öffentlichkeit*. Luchterhand.

Habernal, I., Wachsmuth, H., Gurevych, I., & Stein, B. (2018). The argument reasoning comprehension task: Identification and reconstruction of implicit warrants. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 1930–1940). Association for Computational Linguistics.

Harari, Y., Harris, T., & Raskin, A. (2023, March 24). You can have the blue pill or the red pill, and we're out of blue pills. *The New York Times*.

Havelock, E. (1963). *Preface to Plato*. Harvard University Press.

Havelock, E. (1986). *The muses learn to write: Reflections on orality and literacy from antiquity to the present*. Yale University Press.

Hayles, K. N. (2010). How we read: Close, hyper, machine. *ADE Bulletin*, *150*, 62–79.

Hedgepeth, C. (2024, August 13). Google vs. ChatGPT: Traditional search still going strong. *9Rooftops*. https://www.9rooftops.com/blog/google-vs-chatgpt-traditional-search-still-going-strong/

Hofstädter, D. (1995). *Preface 4: The ineradicable Eliza effect and its dangers. In Fluid concepts and creative analogies: Computer models of the fundamental mechanisms of thought*. Basic Books.

Husserl, E. (1931). *Meditations cartésiennes: Introduction à la phenomenologie*. Armand Collin.

Husserl, E. (1972). *Erfahrung und Urteil: Untersuchungen zur Genealogie der Logik*. Meiner.

Introna, L. D., & Nissenbaum, H. (2000). Shaping the Web: Why the politics of search engines matters. *The Information Society*, *16*, 169–185. https://doi.org/10.1080/01972240050133634

Jones, N. (2025, January 21) AI hallucinations can't be stopped–but these techniques can limit their damage. *Nature*, *637*, 778–780.

Kang, M., Gürel, N. M., Yu, N., Song, D., & Li, B. (2024). C-RAG: Certified generation risks for retrieval-augmented language models. In *Proceedings of the 41st International Conference on Machine LearningArticle No.: 924* (pp. 22963–23000).

Krajewski, M. (2011). *Paper machines: About cards & catalogs, 1548-1929*. MIT Press.

Korteling, J. E., van de Boer-Visschedijk, G. C., Blankendaal, R. A. M., Boonekamp, R. C., & Eikelboom, A. R. (2021). Human- versus artificial intelligence. *Frontiers in Artificial Intelligence*, *4*, 622364. https://doi.org/10.3389/frai.2021.622364

Langer, M., Oster, D., Speith, T., Hermanns, H., Kästner, L., Schmidt, E., Sesing, A., & Baum, K. (2021). What do we want from explainable artificial intelligence (XAI)? A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. *Artificial Intelligence*, *296*, 103473. https://doi.org/10.48550/arXiv.2102.07817

Langville, A. N., & Meyer, C. D. (2006). *Google's PageRank and beyond: The science of search engine rankings*. Princeton University Press.

Lee, M. (2023). A mathematical investigation of hallucination and creativity in GPT models. *Mathematics*, *11*, 2320. https://doi.org/10.3390/math11102320

Ling, C., Zhao, X., Lu, J., Deng, C., Zheng, C., Wang, J., Chowdhury, T., Li, Y., Cui, H., Zhang, X., Zhao, T., Panalkar, A., Mehta, D., Pasquali, S., Cheng, W., Wang, H., Liu, Y., Chen, Z., Chen, H., … (2024). Domain specialization as the key to make large language models disruptive: A comprehensive survey. *ACM Computing Surveys*, *58*, 1–39. https://doi.org/10.1145/3764579

Lord, A. B. (1960). *The singer of tales*. Harvard University Press.

Lu, H., Liu, Y., Xu, J., Nan, G., Yu, Y., Chen, Z. & Wang K.. (2025). 'Auditing meta-cognitive hallucinations in reasoning large language models', arXiv, https://doi.org/10.48550/arXiv.2505.13143, preprint: not peer reviewed.

Luhmann, N. (1985). Die Autopoiesis des Bewussteins. *Soziale Welt*, *36*, 402–446.

Luhmann, N. (1997). *Die Gesellschaft der Gesellschaft*. Suhrkamp.

Mead, G. H. (1943). *Mind, self, and society*. University of Chicago Press.

Mehta, P. (2024, December 5). Underrated, overpowered: 18 GPTs you need in your life. Stop Wasting Time–these GPTs are absolute game-changers. *Artificial Intelligence in Plain English*. https://ai.plainenglish.io/underrated-overpowered-18-gpts-you-need-in-your-life-7bc26e1766db

Migdal, P. (2022, July 17). DALL·E 2 and transcendence: Generating esoteric images with AI. *Medium*.

Mitchell, M., & Krakauer, D. C. (2023). The debate over understanding in AI's large language models. *Proceedings of the National Academy of Sciences of the United States of America*, *120*, e2215907120. https://doi.org/10.1073/pnas.2215907120

Moor, J. (2006). The Dartmouth College Artificial Intelligence Conference: The next fifty years. *AI Magazine*, *27*, 87–89.

Mustapha, K. B. (2025). A survey of emerging applications of large language models for problems in mechanics, product design, and manufacturing. *Advanced Engineering Informatics*, *64*, 103066. https://doi.org/10.1016/j.aei.2024.103066Getrightsandcontent

Ong, W. J. (1982). *Orality and literacy: The technologizing of the word*. Methuen.

OpenAI. (2025, April 16). *OpenAI o3 and o4-mini system card*. https://openai.com/index/o3-o4-mini-system-card/

Pilati, F., Munk, K. A., & Venturini, T. (Eds.). (2024). Repurposing generative AI for social research. *Sociologica*, *18*, 1–8. https://doi.org/10.6092/issn.1971-8853/20378

Prensky, M. (2001). Digital natives, digital immigrants. *On the Horizon*, *9*, 1–6.

Rathkopf, C. (2025). 'Hallucination, reliability, and the role of generative AI in science', arXiv, https://doi.org/10.48550/arXiv.2504.08526, preprint: not peer reviewed.

Senecal, S., Coursaris, C. K., Léger, P. M., & Amoros, S. (2025). March 24). Google's AI-generated search feature hasn't yet changed how users interact with search results. *The Conversation*. https://theconversation.com/googles-ai-generated-search-feature-hasnt-yet-changed-how-users-interact-with-search-results-244607

Sejnowski, T. (2023). Large language models and the reverse Turing test. *Neural Computation*, *35*, 309–342.

Shao, A. (2025). 'Beyond misinformation: A conceptual framework for studying AI hallucinations in (science) communication', arXiv, https://doi.org/10.48550/arXiv.2504.13777, preprint: not peer reviewed.

Sharma, S. (2024). 'Retrieval augmented generation for domain-specific question answering', arXiv, https://doi.org/10.48550/arXiv.2404.14760, preprint: not peer reviewed.

Shen, X., Chen, Z., Backes, M., Shen, Y., & Zhang, Y. (2023). "Do anything now": Characterizing and evaluating in-the-wild jailbreak prompts on large language models. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security* (pp. 1671–1685). https://doi.org/10.1145/3658644.3670388.

Smith Diaz-Andreu, E. (2024). The prose code: A journey into the AI-lluminated world of literary algorithms. *The Literary Platform*, 6. https://theliteraryplatform.com/stories/the-prose-code-a-journey-into-the-ai-lluminated-world-of-literary-algorithms/

Søgaard, A. (2023). On the opacity of deep neural networks. *Canadian Journal of Philosophy*, *53*, 224–239. https://doi.org/10.1017/can.2024.1

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. R. (2018). GLUE: A multi-task benchmark and analysis platform for natural language understanding. Proceedings of the EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP (pp. 353–355). Association for Computational Linguistics. https://aclanthology.org/W18-5446/

Watzlawick, P., Beavin, J. H., & Jackson, D. D. (1967). *Pragmatics of human communication: A study of interactional patterns, pathologies, and paradoxes*. Norton.

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

West, R., & Aydin, R. (2024). 'The AI alignment paradox: The better we align AI models with our values, the easier we may make it to re-align them with opposing values', arXiv, https://doi.org/10.48550/arXiv.2405.20806, preprint: not peer reviewed.

White, R. W., Richardson, M., & Yih, W. (2015). Questions vs. queries in informational search tasks. In *WWW '15 Companion: Proceedings of the 24th International Conference on World Wide Web* (pp. 135–136). ACM. https://doi.org/10.1145/2740908.2742765

Wu, Y., Lin, J., Zheng, Z., & Cao, K. (2024). Impact of the LLM on the Google. Proceedings of the 3rd International Conference on Business and Policy Studies. https://doi.org/10.54254/2754-1169/106/20241421

Yao, Z., Liu, Y., Chen, Y., Chen, J., Fang, J., Hou, L., Li, J. & Chua, T., (2025). 'Are reasoning models more prone to hallucination?' arXiv, https://doi.org/10.48550/arXiv.2505.23646, preprint: not peer reviewed.

Xu, Z., Jain, S., & Kankanhalli, M. (2024). 'Hallucination is inevitable: An innate limitation of large language models', arXiv, https://doi.org/10.48550/arXiv.2401.11817, preprint: not peer reviewed.

Xu, Z., Liu, Y., Deng, G., Li, Y., & Picek, S. (2024). 'A comprehensive study of jailbreak attack versus defense for large language models', In *Findings of the Association for Computational Linguistics: ACL 2024* (pp. 7432–7449).

Communication Theory, 2026, 00, 1–9
https://doi.org/10.1093/ct/qtaf036
Original Article