İstanbul Üniversitesi Sosyoloji Dergisi,
IUSD 2026, 46 (1): 455–473

https://doi.org/10.26650/SJ.2025.46.1.1019

Submitted 02.11.2025
Revision Requested 09.04.2026
Last Revision Received 24.05.2026
Accepted 09.06.2026
Published Online 28.07.2026

İstanbul Üniversitesi Sosyoloji Dergisi

Review Article

Open Access

From Form to Ratification: Reassessing AI "Communication" After the Facebook Chatbot Episode

Tolga Yıldız [1]

[1] İstanbul University, Department of Psychology, İstanbul, Türkiye

Abstract This article revisits the 2017 Facebook negotiation-bot episode, widely reported as a case of artificial intelligence "inventing a language," to ask a narrower and more sociologically precise question: under what conditions should machine output count as language rather than as task-conditioned symbol coordination? Methodologically, the article is a narrative review and conceptual synthesis organized in line with the Scale for the Assessment of Narrative Review Articles (SANRA). It draws on purposively selected work in developmental psychology, philosophy of mind, sociology, and recent research literature on large language models (LLMs). The analysis proceeds in three steps. First, it shows that the Facebook case involved a reinforcement-learning negotiation agent and cannot be generalized to contemporary LLMs without serious qualification. Second, it argues that current debates about AI "understanding" are best read not as a choice between hype and dismissal but as disputes about different kinds of competence. Third, it develops a sociological distinction between form, coordination, and ratification. Contemporary AI systems can generate linguistically legible forms and sometimes support limited coordination, but language in the stronger sense depends on public uptake, normative correction, and institutional durability. The article's contribution lies in shifting the debate from inner-state attribution to the social conditions under which signs become language.

Keywords Chatbot · Large language models · Shared intentionality · Chinese Room · Anthropomorphism · Social construction of language

Citation: Yıldız, T. (2026). From form to ratification: reassessing AI "communication" after the Facebook chatbot episode. *İstanbul Üniversitesi Sosyoloji Dergisi*, *46*(1), 455-473. https://doi.org/10.26650/SJ.2025.46.1.1019

This work is licensed under Creative Commons Attribution-NonCommercial 4.0 International License.

2026. Yıldız, T.

Corresponding author: Tolga Yıldız tolga.yildiz@istanbul.edu.tr

Introduction

In 2017, a familiar headline circulated across global media outlets: Facebook chatbots had “invented their own language.” The story was irresistible. It condensed a diffuse set of technological anxieties into a single image: machines speaking among themselves, outside human supervision, perhaps already slipping beyond human control. Yet the underlying research was far less dramatic. The system in question was a negotiation model trained to maximize success in a bounded bargaining task, and the strange strings it produced were not evidence of a newly founded public language. They were a by-product of optimization under weak linguistic constraints (Lewis et al., 2017).

Still, the episode remains analytically useful. Its significance lies less in what the bots did than in what observers thought they had done. It reveals a recurring temptation in public and scholarly discourse alike: the tendency to treat language-like output as evidence of language in the fuller human sense. That temptation has only intensified in the age of large language models (LLMs), whose fluency far exceeds that of the 2017 negotiation agents. Once systems produce grammatically persuasive text, questions about meaning, understanding, communication, and even personhood return with renewed force.

This article begins from a narrower claim than celebratory or alarmist accounts usually allow. It does not argue that artificial systems are, in principle, forever incapable of richer communicative competence. That stronger ontological claim remains contested and cannot be settled here. Nor does it argue that contemporary AI should be dismissed as mere statistical mimicry without remainder. That dismissal misses too much. The inquiry pursued here is more circumscribed: what conditions must be met before machine output should be described as language rather than as a technically successful but socially thinner form of symbol coordination?

This problem is often blurred in two ways. Public discourse moves too quickly from fluent output to anthropomorphic inference. Scholarly debate, by contrast, often moves too quickly from the philosophical problem of “understanding” to broad conclusions about language itself. Recent work on LLMs has become more careful on this point, especially in distinguishing between formal and functional competence, between performance and grounding, and between useful behavioral

shorthand and stronger ontological commitments (Mahowald et al., 2024; Mitchell & Krakauer, 2023; Shanahan, 2024). Even so, what remains underdeveloped is a specifically sociological account of the threshold at which a sign system becomes language as a public social fact.

At stake, I argue, is a threshold that syntax, fluency, and local task success alone cannot capture. Human language is not only a structured system of signs. It is a socially ratified practice. Signs become language when they are taken up by the public, stabilized through correction and reuse, inserted into relations of value and authority, and sustained across time. To clarify this point, the article develops a distinction between form, coordination, and ratification. This distinction is the article’s main contribution. It allows one to acknowledge the extraordinary fluency of current AI systems without collapsing that fluency into human-like communication.

Existing accounts have already drawn an important distinction between form and meaning and between performance and understanding. The contribution of this article is to add a different threshold to this discussion: the distinction between linguistic form, local coordination, and social ratification.

The argument unfolds in seven stages. It first explains the methodological scope of the review and the rationale for its conceptual architecture. It then revisits the Facebook negotiation-bot episode and specifies the limits of that case. Next, it situates the article in the current debate over LLM understanding. It then turns, in sequence, to developmental psychology, philosophy of mind, and sociology. The argument culminates in a more explicit account of what would count as stronger evidence of machine language and in a concluding reflection on the implications for the sociology of AI-mediated communication.

Method, Scope, and Analytical Position

This article is a narrative review and conceptual synthesis, not an empirical study. Because narrative reviews are particularly vulnerable to diffuseness, selective citation, and unmarked shifts in argumentative level, the article is organized in line with the Scale for the Assessment of Narrative Review Articles (SANRA) and therefore explicitly states its scope, search logic, selection criteria, analytical strategy, and limitations (Baethge et al., 2019).

The scope is deliberately bounded. The review centers on one emblematic public case, the 2017 Facebook negotiation bots, and on the subsequent scholarly debate over whether and in what sense contemporary LLMs can be said to understand or use language. The literature was assembled through purposive searches of major scholarly sources and publisher repositories, including *ACL Anthology*, *Proceedings of the National Academy of Sciences*, *Communications of the ACM*, *Trends in Cognitive Sciences*, *Open Mind*, *JSTOR*, and major university press catalogues. Search terms included combinations of Facebook chatbots, negotiation agents, LLM understanding, formal competence, functional competence, shared intentionality, Chinese Room, anthropomorphism, social construction of language, and sociology of AI.

Selection proceeded by three criteria. The first was theoretical centrality: a source was included when it provided concepts indispensable to the argument, such as shared intentionality, public meaning, institutionalization, symbolic capital, or semantic accountability. The second was topical relevance: recent peer-reviewed work was prioritized when it addressed the current question of LLM competence directly rather than AI in general. The third was argumentative necessity: sources were retained when they helped distinguish the question of language from adjacent questions about benchmark performance, generic AI ethics, or broad future-of-technology speculation.

Exclusion followed the same logic. This article does not attempt to survey the entire engineering literature on language models. It excludes papers whose primary contribution is technical benchmarking without direct bearing on the concept of language, and it excludes generalized ethical commentary not tied to the linguistic issue at stake. It also avoids treating media reports as equivalent to scholarly evidence, except when they are themselves the object of analysis, as in the Facebook case.

This review was designed to clarify a bounded conceptual problem rather than map the full technical literature on AI systems. Priority was therefore given to works published between 2020 and 2024 that addressed language, understanding, or public meaning directly, while earlier texts were retained only where they provided concepts indispensable to the argument. Sources were not included merely because they concerned AI in a broad sense; they were selected when they bore directly on the distinction between linguistic form, communicative competence, and socially ratified

meaning. This procedure increases conceptual focus, but it also means that the review is interpretive rather than exhaustive.

The analytical strategy is cumulative rather than additive. Developmental psychology is used to clarify how communicative competence emerges in human life. Philosophy of mind is used to distinguish formal performance from meaning and answerability. Sociology then reframes the problem by asking how signs become publicly stabilized, valued, and institutionally durable. Therefore, the argument moves from emergence to attribution and ratification.

Two points should be made here. First, the choice of developmental psychology, philosophy of mind, and sociology is deliberate rather than decorative. Linguistics and cognitive science remain present in the background, and much of the current LLM debate emerges from them, but they are not the organizing frames of this article because the issue here is not simply whether models capture grammar or solve tasks, but under what social and normative conditions output becomes language in a robust sense. Science and technology studies are likewise relevant, and selected STS work is incorporated where it sharpens the sociological argument. This article remains anchored in sociology because its center of gravity is public uptake, symbolic value, and institutional durability.

Second, the claim to originality is intentionally limited. The article does not claim novelty for the general skepticism that current AI systems lack human-like understanding; that line of argument is already well established, from Searle (1980) to Bender and Koller (2020). Nor does it claim to solve the metaphysical problem of machine consciousness. Its more modest contribution is to relocate the debate from inner-state attribution to the social threshold at which signs become ratified as language. The article is written from a sociology-of-knowledge orientation, and that orientation openly shapes what it can and cannot answer. It privileges public accountability over metaphysical adjudication.

One final point concerns the analytical position. This article proceeds from a sociology-of-knowledge perspective. It does not attempt to resolve the metaphysical question of whether artificial systems could, in principle, attain consciousness or understanding. Its concern is narrower: how claims about language are socially licensed, stabilized, and institutionalized. For that reason, the article treats philosophical accounts of meaning as necessary but not sufficient, and it approaches

AI “communication” primarily through the public life of signs rather than through speculative claims about inner states.

Like any narrative review, this article has limitations. It cannot claim exhaustive coverage. Its judgments about relevance are interpretive. Because it addresses moving literature, any synthesis remains time-bound. These limitations are not incidental; they are part of the method and should remain visible.

Facebook Negotiation Bots: What the Case Can and Cannot Bear

The research paper behind the 2017 controversy described an end-to-end negotiation system trained on dialogues about dividing objects with different values (Lewis et al., 2017). The task was not open-ended communication, storytelling, or world description. It was bounded bargaining. During training, some outputs departed from ordinary English and became repetitive or opaque. These strings were sufficient, under specific conditions, to sustain task performance, but they were not designed for public intelligibility and did not amount to the founding of a new linguistic community.

This distinction matters because the public narrative rapidly outran the research design. The experiment was not designed to test whether machines could find a language, much less whether they could evolve beyond human control. It was designed to see whether agents could learn useful negotiation strategies through data and self-play. The outputs became strange because the reward structure did not strongly penalize deviations from standard human readability. What appeared, from the outside, as “machine language” was, from the standpoint of the task, a local optimization.

It is equally important not to recast the original episode as if it involved a contemporary LLM. The system at issue was a reinforcement-learning negotiation agent built for a specific goal. Moving directly from that case to present-day transformer-based models blurs distinctions in architecture, training regime, and output profile. That blurring was one of the central weaknesses of much early commentary on the case, and it remains a temptation because chatbots often function in public language as an umbrella term covering very different systems.

At the same time, the case should not simply be discarded as outdated. It remains revealing because it exposes a recurrent mechanism of public over-attribution. Strange but patterned output

was read as a secret intention. Narrow task success was interpreted as autonomous communication. In this sense, the episode anticipates precisely the problem Shanahan (2024) identifies in contemporary discussions of LLMs: once systems produce text that resembles language, descriptions of them rapidly become philosophically loaded. Terms such as knows, believes, intends, and communicates migrate from pragmatic shorthand into metaphysical suggestions.

The Facebook episode is therefore best treated not as decisive evidence about what today’s LLMs can or cannot do but as a scene of misrecognition. It shows how quickly a legible or semi-legible form becomes a canvas for projections about agency and language. That lesson remains current.

From the 2017 Media Narrative to the Current Debate on LLM Competence

If the Facebook episode belongs to the era of media sensationalism, this debate over LLMs is more conceptually demanding. Recent scholarship has become notably more careful in distinguishing kinds of competence and levels of attribution.

Mitchell and Krakauer (2023) provide a useful map of the debate by showing that whether LLMs ‘understand’ is not a single, unified problem. Some claims concern behavioral competence: Can the system respond appropriately across contexts? Others concern grounding: does it connect linguistic forms to the world in a meaningful way? Still others concern social cognition, generalization, or human-like intelligence more broadly. Their intervention is valuable because it refuses the false choice between enthusiasm and dismissal. It shows that the debate is heterogeneous because the phenomenon itself is heterogeneous.

Mahowald et al. (2024) sharpened the issue by distinguishing between formal and functional linguistic competence. LLMs are often highly impressive at the former. They capture regularities of syntax, lexical distribution, genre, and many other features of linguistic form at a remarkable scale. But functional competence, using language in the world, in relation to shared context, goals, interlocutors, and practical situations, remains uneven. This distinction is particularly useful because it allows one to take current models seriously without granting too much too quickly. One need not say that models fully understand language in a human sense to acknowledge that they model important aspects of linguistic form extremely well.

A related caution concerns claims about an abrupt breakthrough. Narratives of "emergent abilities" have sometimes been used to suggest that once models cross a certain scale, qualitatively new capacities suddenly appear. Yet this literature has itself become contested, with Schaeffer et al. (2023) showing that some obvious discontinuities may be artifacts of evaluation choices rather than evidence of deep thresholds. This does not imply that the scale is irrelevant. It implies only that threshold language requires stronger evidence than spectacle.

Messeri and Crockett (2024), writing about scientific research rather than everyday chatbot use, add another layer by analyzing what they call "illusions of understanding." Their argument matters here because it shifts attention from what systems compute to what humans infer. AI-generated output can produce a persuasive impression of coherence, grasp, or conceptual command more easily than it warrants. In that respect, the central problem is not only technical. It is epistemic and social. It concerns who authorizes a performance, under what conditions, and with what consequences.

A balanced account cannot rest on the skeptical literature alone. Functionalist traditions have long argued that mental and intentional states should not be tied too rigidly to biological substrates (Putnam, 1975). Dennett's (1987) intentional stance makes a more modest but still significant point: there are cases in which treating a system as if it had beliefs or intentions is an effective predictive strategy. These traditions do not prove that LLMs understand language in a human sense. But they do block an easy slide from "present systems are artificial" to "therefore nothing like meaningful competence can ever be attributed to them."

The best current discussions therefore proceed with more discipline. They neither enthrone LLMs as hidden minds nor reduce them to empty noise. What matters, then, is what exactly has been achieved and what kind of status that achievement licenses. That is the point at which developmental, philosophical, and sociological distinctions become necessary.

Developmental Grounds: Language Begins in Shared Orientation

A developmental perspective reframes the problem at its root. Human beings do not first acquire a stock of signs and only later add social meaning to them. Language emerges within relations of

interaction from the outset. For Vygotsky (1978), language is a cultural tool whose use reorganizes cognition itself. Children become speakers through participation with more competent others, correction, modeling, shared tasks, and gradual internalization. In this view, language is not simply stored in the individual mind as a code. It is a socially mediated means of action.

Tomasello (2008) provides a complementary emphasis through the concept of shared intentionality. Human communication develops through joint attention, cooperative motives, and the ability to treat another person as a partner in a common scene. This is more than signal production. A child learns that words can direct attention, repair misunderstanding, invoke absent objects, and coordinate experience with others that are themselves interpreters. The developmental achievement lies not simply in generating an acceptable string but in taking part in a shared world.

This matters for AI because current models traverse nothing like a comparable developmental path. They are trained on vast corpora of already externalized language rather than entering language through embodied participation in a small but intensely structured social environment. Cuskley et al. (2024) make this point sharply: even when LLMs resemble human language use at the output level, the conditions of their learning differ profoundly from those of human language acquisition and cultural evolution. Human learners develop through multimodal, emotionally saturated, socially scaffolded interaction. Text-only statistical training is not equivalent to developmental ecology.

That does not establish an impossibility thesis. One can imagine future systems situated within richer interactional environments. But it does mean that present-day models should not be too quickly described as language users simply because their output is fluent. The developmental literature contributes a criterion of shared orientation: language arises through coordinated attention, reciprocal adjustment, and learning under the expectations of others. This criterion does not settle every philosophical question, but it already prevents the collapse of language into mere string production.

Philosophical Clarification: From Fluency to Semantic Accountability

The philosophical question is often staged through Searle’s (1980) Chinese Room argument. A system may produce formally appropriate outputs without understanding what those outputs mean. Whatever one thinks of the argument in its strongest form, its enduring value lies in the distinction it enforces between successful symbol manipulation and semantic comprehension. Fluency alone is not yet a theory of meaning.

Wittgenstein (1953) pushes the problem in a different direction. Meaning is not a detachable substance hidden inside words; it is bound to use within forms of life. This point is sometimes invoked loosely in AI discussions, but its relevance remains strong. If meaning is inseparable from public practices of use, then asking whether an AI system “uses language” cannot be reduced to asking whether its outputs are well formed. The issue is whether those outputs stand within practices of correction, uptake, responsibility, and consequence.

Here, however, caution is required. A fair discussion must acknowledge the force of counterposition. Functionalism argues that mental states are defined by their causal and behavioral roles rather than by their biological material alone (Putnam, 1975). Dennett’s intentional stance suggests that some forms of intentional description may be pragmatically warranted even when their metaphysical depth remains unclear. These views complicate any simple argument from a non-biological substrate to a non-meaningful output.

These counterpositions matter because they make easy skepticism untenable. If sufficiently rich behavioral performance can justify thin intentional descriptions, contemporary models cannot simply be dismissed as empty formal devices. Yet this concession does not by itself establish that a sign system has entered the public, norm-governed life of language. The issue, then, is not whether current models exhibit impressive competence; they clearly do. The issue is what kind of competence licenses claims about language as a social practice rather than as a successful patterning of outputs. That is precisely the point at which ratification becomes analytically necessary.

The claim advanced here is narrower. It does not assert that current AI systems can never, under any conceivable conditions, instantiate stronger forms of understanding. Nor does it equate the

absence of human biology with the total absence of any meaningful competence. What it insists on is more limited: competent performance alone does not settle the issue of semantic accountability.

By semantic accountability, this article means more than syntactic well-formedness and locally successful task completion. An utterance is semantically accountable when it is accountable to a shared world, public norms of correction, and consequences that exceed the immediate completion of a task or prompt. Human speakers can be asked what they meant, whether they misled, whether a concept was appropriately used, and whether a claim survives a later challenge. Model outputs can certainly be evaluated by users, institutions, and downstream systems. But that very point shows where accountability presently lies: not in a speaking subject who stands behind the utterance but in a hybrid arrangement of training data, interface design, prompting, institutional use, and human interpretation.

This is why the philosophical debate cannot be resolved by fluency alone. LLMs clearly exceed earlier systems in formal performance. They may even exhibit context-sensitive behaviors that make intentional description pragmatically tempting. But the stronger claim that they communicate as speakers does require a further step: the move from generated form to publicly answerable language. This step is not only philosophical. It is sociological.

A Sociological Theory of Ratified Language

If developmental psychology clarifies how language emerges and philosophy clarifies why fluency is insufficient, sociology asks a different question: how do signs become public, durable, and socially consequential? At this point, the article’s main contribution emerges.

Debates about AI “communication” become clearer when three levels are distinguished: form, coordination, and ratification.

Form refers to output that is recognizably language-like. At this level, a sequence is grammatically plausible, stylistically coherent, genre-appropriate, or distributionally aligned with human text. Contemporary LLMs are highly capable here. They generate sentences, paragraphs, summaries, instructions, and genres that often pass ordinary thresholds of legibility.

Coordination refers to the capacity of signs to support a bounded practical task. A signal does not need to be publicly meaningful in a rich sense to coordinate action locally. The Facebook negotiation bots reached something like this level: their outputs were enough, under narrow conditions, to sustain a bargaining routine. Human social life includes analogous cases: shorthand among experts, temporary working codes, local abbreviations, and even ephemeral in-group signals. But coordination alone does not yet amount to language in the stronger sociological sense.

Ratification is the crucial third level. A sign system becomes socially ratified when it is taken up by a public, stabilized through correction and reuse, attached to shared expectations, and inserted into relations of authority, legitimacy, and value. Mead (1934) helps identify the interactional dimension of this process: meaningful action is oriented not merely to a dyadic exchange but also to a generalized other. Berger and Luckmann (1966) help identify its institutional and temporal dimension: meanings become durable when they are externalized, objectivated, and internalized across actors and generations. Bourdieu (1991) adds a further layer: language is never mere neutral exchange; it is unequally valued practice, bound up with field position, symbolic capital, and the power to authorize legitimate speech.

These traditions are neither identical nor fully commensurable at the level of first principles. Meadian interactionism, phenomenological sociology, and Bourdieusian field theory differ in their assumptions about action, social order, and the level at which meaning is stabilized. This article does not treat them as parts of a seamless grand theory. It uses them selectively and at different analytical scales: Mead for interactional uptake, Berger and Luckmann for institutional sedimentation, and Bourdieu for the unequal valuation of legitimate expression within fields. Therefore, the point is not theoretical fusion but limited analytical articulation around the question of how signs become publicly consequential.

From a sociological standpoint, the real issue is not whether a system emits a patterned signal or whether that signal succeeds in a local task. It is whether the signal enters a public world of norm-governed uptake. Ratification requires more than recognition. It requires a social process through which forms acquire durability, correctability, and differential value.

This framework clarifies both the Facebook case and the status of present-day LLMs. The Facebook bots achieved, at most, local coordination. Their strings were not carried into a public interpretive community, did not become durable conventions, and were not stabilized through collective correction. Once the experiment ended, the strings vanished. They did not become a language because no social body ratified them as such.

Contemporary LLMs occupy a more complex position. Unlike the 2017 negotiation bots, they are trained on enormous corpora of already ratified human language. Their fluency is therefore dependent on prior public stabilization achieved by human communities. Models combine forms that have already been sedimented. In many settings, they also facilitate coordination. They help users draft emails, summarize meetings, translate documents, answer questions, and generate code. But even here, the ratifying force remains external. Human institutions, professions, audiences, editors, managers, teachers, bureaucracies, and publics decide whether an output is adopted, circulated, corrected, credited, or discarded.

One further qualification is needed here. Ratification should not be treated as a single all-or-nothing threshold operating uniformly across all domains. Some forms of field-specific and partial ratification may already be observable. Professional software development is one example, where AI-generated code has achieved routinized uptake within workflows, even if its legitimacy remains dependent on human review, integration, and responsibility. Certain professional writing practices, including some academic ones, offer another example: AI-generated prose may be signed, submitted, revised, and circulated under human names, even where institutional norms remain unsettled or partially concealed. These cases do not yet amount to full ratification in the sense used here because the output still lacks autonomous standing as a recognized speaking position. Instead, they show that ratification may be uneven, field-specific, and scalar rather than simply absent or complete.

This point is easy to miss because AI systems are often described as if they were solitary actors. A sociological analysis yields a different picture. What appears as “machine communication” is more accurately understood as an assemblage: training corpora produced by human labor; interfaces shaped by firms; prompts authored by users; evaluation criteria imposed by institutions; and

outputs interpreted within preexisting fields of value. Suchman (2007) warned against accounts of human-machine interaction that obscure the performative work through which agency is configured. Woolgar (1985), in an earlier sociology of AI, made a related point: AI is revealing not because it straightforwardly reproduces human intelligence but because it exposes how the categories of intelligence, expertise, and agency are rhetorically and institutionally constructed.

Seen this way, the sociological claim is not simply that AI systems either speak or fail to speak. They increasingly mediate linguistic practice while remaining dependent on human ratification. Their outputs can redistribute symbolic capital, alter the value of writing labor, and reshape who counts as a competent speaker or author in specific fields. But these effects do not show that machines have independently founded a language community. They show that machine-generated forms have entered human institutions and are now being processed within them.

Reading through Bourdieu, the issue becomes more precise. If linguistic legitimacy is field-dependent, AI output matters not simply because it circulates. It matters because, under specific institutional conditions, it can be converted into recognized competence. In some fields, fluency may begin to function as a proxy for expertise; in others, assisted production may devalue embodied or credentialed authorship by obscuring the labor on which legitimate speech traditionally rests. What matters, then, is not merely that AI-generated language enters circulation but that it may reconfigure the terms under which symbolic capital is accumulated, delegated, or contested.

This perspective also clarifies the problem of anthropomorphism. Weizenbaum (1976) recognized, in responses to ELIZA, how readily humans infer mindedness from dialogic responsiveness. Epley (2014) provides a broader account of this tendency. The sociological question, however, is not merely why individuals anthropomorphize. It is how such attributions become organized, rewarded, and normalized within platforms, professions, and institutions. Which settings encourage users to treat machine output as authored speech? Which infrastructures conceal the labor, data, and policy apparatus behind a fluent interface? Which organizations benefit when generated text is received as if it carried speakerly authority?

The idea of ratification does two things here. Analytically, it prevents the collapse of the linguistic form into full communication. Empirically, it directs attention to the institutions through which AI

output becomes consequential. A sentence generated by a model acquires social weight not because the model privately means it, but because courts, publishers, schools, firms, bureaucracies, and publics may decide to treat it as credible, useful, or authoritative. The sociological threshold is crossed not inside the model but in the social life of the output.

What Would Count as Stronger Evidence of Machine Language?

Both celebratory and skeptical accounts of AI tend to leave the evidential threshold vague. If the current model output does not yet justify calling it language in the robust sense, what would move the argument?

In this view, five stronger criteria follow.

First, a sign system would need to show stable cross-situational persistence. One-off odd strings or locally useful compression are not enough. The forms would need to recur across contexts without being reducible to transient optimization noise.

Second, there would need to be public uptake beyond the originating system boundary. Here, a distinction is necessary between uptake as adoption and uptake as ratification. Adoption means that outputs are widely used, repeated, or incorporated into practice. Ratification requires more: such forms acquire recognized legitimacy, become subject to stable norms of correction and attribution, and are granted standing within a community as more than convenient instruments. No language is founded merely because a model or pair of models produces a pattern, or even because that pattern is reused at scale. The pattern would need to be learned, reused, normatively stabilized, and recognized by a community as carrying a socially valid standing beyond the immediate technical loop.

Third, the system would need to participate in repair and negotiated correction. Human language is not only produced; it is continuously fixed, challenged, clarified, and recalibrated. Stronger claims about machine language would require evidence not only of output generation but also of sustained participation in mutual repair under shared norms.

Fourth, the sign system would need to acquire institutional insertion and symbolic value. This does not mean state recognition or formal codification in a maximal sense. It means entry into

social settings where signs matter, where they are differentially valued, and where their use is tied to recognizable expectations and consequences.

Fifth, there must be some form of diachronic continuity. Language persists. It outlives single interactions and becomes available for future uptake. Without this temporal durability, one has coordination or performance but not yet a socially ratified practice.

These criteria do not settle the philosophical question of machine consciousness and are not intended to. They are sociological criteria. They specify what stronger evidence would have to look like before claims about machine “language” become more than metaphors, shortcuts, or marketing language.

Discussion

Three recurring mistakes become easier to avoid once the argument is framed in these terms.

The first is technical flattening. Not all AI systems are the same. No negotiation agent, a rule-based customer-service bot, and a transformer-based LLM should be treated interchangeably simply because each may be described colloquially as a chatbot. The distinction between form and coordination helps prevent such flattening.

The second is philosophical inflation. It is tempting either to attribute full understanding to current models or to deny them any meaningful competence whatsoever. The distinction between form, coordination, and ratification opens a narrower middle path. It permits strong recognition of present-day performance without converting that performance into an unwarranted ontology of machine subjectivity.

The third is sociological thinness. Much AI discourse still stops at the level of cognition or computation. But for sociology, the more urgent question concerns public uptake: how claims about AI speech are authorized, where they circulate, whose labor they obscure, and how they reshape hierarchies of expertise, authorship, and legitimacy. Once this shift is made, the central object of inquiry is no longer the fantasy of machines secretly inventing languages. It is the social process through which institutions begin to treat machine-generated text as if it were equivalent to public language.

This matters in concrete ways. Universities, publishers, administrative systems, and media organizations increasingly encounter generated texts not as abstract philosophical problems but as practical issues of classification, trust, and authorship. In such settings, the operative issue is rarely whether the model ‘really understands.’ What matters is whether the output will be accepted, corrected, sanctioned, or endowed with authority. The sociology of AI language begins there.

This suggests not that ratification is wholly absent but that it is socially uneven. Some domains already exhibit stabilized forms of adoption and delegated legitimacy, whereas others continue to treat AI output as derivative, assistive, or normatively suspect.

Conclusion

The Facebook chatbot episode was never, in any strong sense, a case of AI inventing a language. It was a case of bounded optimization misread through a culturally potent narrative about autonomous machines. But dismissing the episode as a mere media misunderstanding would miss its deeper value. It revealed how quickly language-like output invites anthropomorphic and social over-attribution.

Since then, LLMs have made that temptation much stronger. Their fluency is real. Any serious analysis must begin by acknowledging that. Yet fluency is not the same as communication in the full human sense, and local task success is not the same as public language. The central contribution of this article has been to name this difference more precisely through a three-level distinction between form, coordination, and ratification.

Human language is not only a system of signs. It is a socially ratified practice grounded in shared orientation, semantic accountability, institutional durability, and unequal regimes of value. Contemporary AI systems can generate linguistic forms and sometimes sustain limited coordination. But they do not, by themselves, establish the public processes through which signs become a durable social fact. Their outputs enter the language only through human uptake, correction, and institutional recognition.

For sociology, this is where the real work begins. The decisive question is not whether machines secretly speak, but how societies authorize machine output, redistribute communicative authority,

and renegotiate the boundary between assistance, authorship, and public speech. That is where the study of AI and language becomes fully sociological.

In practical terms, the distinction matters wherever generated text is treated as if it were authored speech. In education, it bears on authorship, learning, and the legitimacy of assisted writing. In publishing and journalism, it bears on credibility, attribution, and editorial responsibility. In administrative and bureaucratic settings, it bears on how generated language may acquire institutional force without a corresponding speaking subject who can answer for it. The point is not that AI text is socially inert but that its authority is mediated and institutionally conferred.



Peer Review Externally peer-reviewed.
Conflict of Interest The author has no conflict of interest to declare.
Grant Support The author declared that this study has received no financial support.

Author Details Tolga Yıldız
[1] İstanbul University, Department of Psychology, İstanbul, Türkiye
0000-0001-9293-2013 tolga.yildiz@istanbul.edu.tr

References

Baethge, C., Goldbeck-Wood, S., & Mertens, S. (2019). SANRA - a scale for the quality assessment of narrative review articles. *Research Integrity and Peer Review, 4* (1), Article 5. https://doi.org/10.1186/s41073-019-0064-8.

Bender, E. M., & Koller, A. (2020). Climbing towards NLU: On meaning, form, and understanding in the age of data. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 5185–5198). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.463.

Berger, P. L., & Luckmann, T. (1966). *The social construction of reality: A treatise in the sociology of knowledge*. Anchor Books.

Bourdieu, P. (1991). *Language and symbolic power* (J. B. Thompson, Ed.; G. Raymond & M. Adamson, Trans.). Harvard University Press.

Cuskley, C., Woods, R., & Flaherty, M. (2024). The limitations of large language models for understanding human language and cognition. *Open Mind, 8*, 1058–1083. https://doi.org/10.1162/opmi_a_00160.

Dennett, D. C. (1987). *The intentional stance*. MIT Press.

Epley, N. (2014). *Mindwise: How we understand what others think, believe, feel, and want*. Alfred A. Knopf.

Lewis, M., Yarats, D., Dauphin, Y., Parikh, D., & Batra, D. (2017). Deal or no deal? End-to-end learning of negotiation dialogues. In M. Palmer, R. Hwa, & S. Riedel (Eds.), *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (pp. 2443–2453). Association for Computational Linguistics. https://doi.org/10.18653/v1/D17-1259.

Mahowald, K., Ivanova, A. A., Blank, I. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2024). Dissociating language and thought in large language models. *Trends in Cognitive Sciences, 28* (6), 517–540. https://doi.org/10.1016/j.tics.2024.01.011.

Mead, G. H. (1934). *Mind, self, and society* (C. W. Morris, Ed.). University of Chicago Press.

Messeri, L., & Crockett, M. J. (2024). Artificial intelligence and illusions of understanding in scientific research. *Nature, 627* (8002), 49–58. https://doi.org/10.1038/s41586-024-07146-0.

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* (13), e2215907120. https://doi.org/10.1073/pnas.2215907120.

Putnam, H. (1975). *Philosophical papers: Vol. 2. Mind, language and reality*. Cambridge University Press.

Schaeffer, R., Miranda, B., & Koyejo, S. (2023). Are emergent abilities of large language models a mirage? In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, & S. Levine (Eds.), *Advances in Neural Information Processing Systems* (Vol. 36, pp. 55565–55581). Curran Associates.

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences, 3* (3), 417–457. https://doi.org/10.1017/S0140525X00005756.

Shanahan, M. (2024). Talking about large language models. *Communications of the ACM, 67* (2), 68–79. https://doi.org/10.1145/3624724.

Suchman, L. A. (2007). *Human-machine reconfigurations: Plans and situated actions* (2nd ed.). Cambridge University Press.

Tomasello, M. (2008). *Origins of human communication*. MIT Press.

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Harvard University Press.

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

Wittgenstein, L. (1953). *Philosophical investigations* (G. E. M. Anscombe, Trans.). Blackwell.

Woolgar, S. (1985). Why not a sociology of machines? The case of sociology and artificial intelligence. *Sociology, 19* (4), 557–572. https://doi.org/10.1177/0038038585019004005.