Psychological brakes: directions for managing anthropomorphism in LLM design*

Gianmarco Pastore[1,2], Cinzia Di Dio[2], Angelo Cangelosi[3]

[1] University of Pisa, Lungarno Pacinotti 43, Pisa, Italy
[2] Catholic University of the Sacred Heart of Milan, Largo Fra Agostino Gemelli 1, Milan, Italy
[3] University of Manchester, 176 Oxford Road, Manchester, United Kingdom

Abstract

In a recent contribution, Floridi (2026) introduces the concept of "semantic pareidolia" to designate the tendency to attribute consciousness, intelligence, and emotions to large language models (LLMs), and closes with a call to develop cultural antibodies to counter it. This paper responds to that call by reconstructing the psychological substrate that Floridi's concept presupposes but does not examine. Three domains of the psychological and neuroscientific literature converge on the same point: humanization is a cognitive, developmental, and evolutionary mechanism, not an interpretive habit. The neural circuits dedicated to face detection are pre-culturally specified, active before reflective elaboration, and shared with non-human primates. Anthropomorphism documents how the mechanism generalizes beyond the visual domain: the ELIZA case first documented that coherent linguistic output alone is sufficient to trigger attribution of understanding and intentions; more broadly, mental characteristics are attributed to any entity displaying contingent, purpose-directed behavior — precisely the behavioral profile that LLMs systematically replicate — with individual differences that are stable over time. Theory of Mind specifies why language is the critical trigger: the evolutionary system connecting structured linguistic output to the attribution of intentional agency operates through an efficient-implicit subsystem that processes language as an intentional signal by architectural definition, before any flexible-reflective elaboration can apply. What this mapping reveals is that countering the tendency to humanize what is not human is not a matter of correcting a bad habit but of managing a psychological tendency — a distinction with direct consequences for intervention design. On this basis, the paper proposes "psychological brakes" as a conceptual category that extends and operationalizes Floridi's call: not measures calibrated on explicit awareness alone, but design and educational interventions calibrated on the implicit, automatic character of the mentalizing tendency. Four operational directions are identified: communicative profile regulation, pragmatic justification requirement, narrowing design, and user cognitive preparation.

Keywords

Semantic pareidolia, Anthropomorphism, Theory of Mind, LLM design ethics, Digital ethics.

1. Introduction

In a recent contribution published in Philosophy & Technology, Floridi [1] introduces the concept of semantic pareidolia, defining the phenomenon as the tendency to attribute consciousness, intelligence, and emotions to artificial intelligence (AI) models that do not possess such qualities. The analogy with traditional perceptual pareidolia — the tendency to perceive faces in clouds or in objects — is proposed as a conceptual tool for rendering intelligible mental attribution toward Large Language Models (LLMs): just as visual pareidolia perceives faces where there are none, semantic pareidolia perceives intentionality where there is only statistical correlation, and comprehension where there is only pattern recognition on a massive scale [1].

Floridi identifies four converging forces that intensify the phenomenon: the progressive digitalization of experience, what the author terms onlife experience, the growing intertwining of physical life and digital life; the economic drive of large companies that speculate on the phenomenon of overestimating LLM's capabilities for profit purposes; the growing loneliness and

*ReDH 2026: Rethinking the Future of Digital Humanities through Ethics and Artificial Consciousness, co-located with AIxIA 2026, October 6–9, 2026, Perugia, Italy

[1] gianmarco.pastore@phd.unipi.it (G. Pastore); cinzia.didio@unicatt.it (C. Di Dio); angelo.cangelosi@manchester.ac.uk (A. Cangelosi)
0009-0002-1404-7347 (G. Pastore); 0000-0002-3962-4684 (C. Di Dio); 0000-0002-4709-2243 (A. Cangelosi)

© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073
published 2026-11-05

CEUR-WS.org/Vol-4314/ReDH10.pdf

credulity in contemporary societies; the progressive improvement of AI models [1]. The escalating character of the trajectory is documented: from ordinary anthropomorphism to technological idolatry, all the way to the theological pareidolia observable in movements such as Way of the Future, Terasem, and The Turing Church, in which religious veneration is oriented toward AI models [1]. The contribution closes with an appeal for the development of cultural antibodies to semantic pareidolia.

Semantic pareidolia is an effective concept because it philosophically describes a psychological reality, that of humanizing and attributing human mental capacities to subjects and objects.

Being aware of the psychological evidence of the tendency to humanize the subjects and objects with which we interact is, however, fundamental for building those "antibodies" of which Floridi himself speaks. Humanizing and attributing mental capacities during our interactions is not a simple habit to be avoided, but a psychological tendency to be managed. The distinction is not aimed at a categorical clarification but is functional to the elaboration of strategies to counter the phenomenon that one intends to counter. Preventing a bad habit versus managing a psychological tendency that is activated whenever we process interactive stimuli are two different things. The increasingly common and everyday conversations with AI are the type of interactive stimuli to which we are increasingly exposed and which are the object both of Floridi's reflection and of the present contribution.

The aim of the present contribution is to offer a mapping of the psychological phenomena underlying the tendency to humanize what is not human — intuited but not problematized by Floridi's article — in order to specify what it means, concretely, to develop psychological brakes: interventions calibrated on the biologically specified character of these tendencies, rather than on the assumption that informational exposure alone is sufficient to correct them.

The psychological phenomena that describe the tendency to humanize and attribute human and mental characteristics in order to optimize our interaction that will be reviewed are (1) pareidolia, narrated in its neuroscientific evidence (2) anthropomorphism, the attribution of human characteristics to entities that are not human, such as personality, thoughts, age, gender, intentions. (3) Theory of Mind (ToM), the capacity to attribute mental states (intentions, beliefs, desires, emotions) to oneself and to others and to use such attributions to explain and predict behavior. Of these three, Theory of Mind constitutes the primary explanatory mechanism of the paper: pareidolia establishes the biological basis of face detection, anthropomorphism documents the extension of mental attribution to non-human entities, but it is ToM that specifically accounts for the automatic attribution of intentional agency in response to structured linguistic output — the modality in which LLMs produce their outputs. The extended treatment of ToM in Section 3 reflects this explanatory hierarchy, not a disproportionate weighting.

2. The empirical foundation of pareidolia and anthropomorphism

The tendency to perceive faces and to attribute human characteristics to entities that are not human does not constitute an occasional error of perception: the literature documents a set of biologically specified processes, present from the earliest months of life and shared with non-human species, which operate systematically independently of the intention or awareness of the observer. The term "pareidolia" requires terminological clarification before the evidence is presented. In the psychological and neuroscientific literature reviewed in this section, pareidolia designates a perceptual phenomenon strictly confined to the visual domain: the activation of the neural circuit dedicated to face detection in the presence of non-facial visual stimuli that satisfy its structural triggering conditions. This is the sense in which the term is used here. Floridi's "semantic pareidolia" employs the same term as a conceptual metaphor for a broader tendency — the attribution of intentional properties to non-human entities on the basis of their behavioral output, including linguistic output. The two uses are related by analogy, not by identity of mechanism. The evidence documented in this section concerns the perceptual-visual phenomenon; its relevance to the present argument is principled rather than direct: it establishes that the

biological architecture of humanization responds to structural properties of stimuli before reflective elaboration, and that this response is not a correctable interpretive habit. The extension of this principle to the linguistic domain is the subject of the following sections.

From the neuroscientific literature it emerges that the neural system dedicated to face recognition does not require the presence of a real face to be activated: it responds in the same modalities to human faces and to visual stimuli that share its basic structure without being one. The measurement of this response has been conducted with different techniques, which detect brain activity respectively through the magnetic fields it produces (magnetoencephalography), variations in cerebral blood flow (functional magnetic resonance imaging), the electrical activity recorded on the scalp (electroencephalography), and the direct recording of the activity of groups of neurons through surgically implanted electrodes.

A first domain of stimuli is that of everyday objects that casually recall a face: teapots, electrical outlets, holes, architectural elements, foods, rocks. Hadjikhani and colleagues [2] have shown that merely observing these images activates the same brain areas that are activated in front of a human face; Wardle and colleagues [3] have confirmed the same result using a standardized set of ninety-six photographs of facelike objects, further showing that the neural representation of illusory faces is initially similar to that of real faces but transforms within approximately 250 milliseconds becoming equivalent to that of ordinary objects, suggesting the existence of an initial automatism and a secondary reflective moment.

A second domain is that of the front parts of automobiles, in which the arrangement of headlights and grille suggests eyes and mouth: Kühn and colleagues [13] have documented the activation of the same brain areas dedicated to face recognition in response to photographs of front ends. In the study by Liu and colleagues [4], when participants were told of the possibility that a face was present in the image, even if in reality it was not, the brain areas dedicated to facial recognition were activated all the same: expectation and a sufficiently ambiguous visual stimulus were enough for the system to construct facial perception.

The most stringent evidence comes from recordings collected through electrodes surgically implanted in the brain of patients, typically people affected by drug-resistant epilepsy, for whom the electrodes are positioned temporarily for clinical purposes. This procedure allows, with the patients' consent, to directly record the activity of small groups of neurons. The study by Cerrahoğlu and colleagues [5] has shown that eighty-nine percent of the recorded neurons activity that responded to face-like objects spatially coincided with recorded neurons activity that responded to real human faces: the same brain cells that are activated at the sight of a human face are also activated at the sight of an object that resembles a face. These anatomical results fix the level at which the phenomenon of pareidolia operates: it is not an interpretive distortion subsequent to a correct percept, but the activation of the same neural circuit dedicated to real faces, in the presence of stimuli that satisfy its structural triggering conditions, in the absence, however, of the actual object that that circuit is intended to detect.

The tendency is not learned during development nor exclusive to the human species. The study of gaze direction through eye tracking, infrared cameras that follow eye movement in real time, without requiring the ability to respond verbally, allows this hypothesis to be tested in subjects who cannot report what they perceive. Kato and Mugitani [6] have shown that infants aged 10 and 12 months orient their gaze toward configurations composed of four spots arranged as eyes, nose, and mouth, in an age range in which cultural exposure is too limited to have shaped the response. The detection also occurs particularly rapidly: Keys, Taubert, and Wardle [7] show significantly shorter detection times for illusory faces compared to visually equivalent objects, and Saurels, Peluso, and Taubert [8] document that the advantage emerges even in peripheral vision, before the observer actively directs the gaze toward the stimulus. Taubert and colleagues [9] extend the phenomenon to rhesus macaques, which when presented with photographs of facelike objects spontaneously fixate on the internal regions with the same gaze pattern directed at real faces: facial detection is anchored to a brain architecture common to primates. These pieces of evidence establish the biological robustness of the general tendency to recognize an element of humanity

where none exists. Not a simple culturally transmitted habit, but a biologically specified mechanism present before any socialization, shared with evolutionarily distant species, and active independently of the intention or awareness of the observer. These pieces of evidence, however, define a mechanism that is activated only in the presence of figurative visual inputs and a humanization limited to the recognition of a face.

The broader principle that detection of human-like structural properties activates mental attribution extends, however, beyond the visual domain. In 1966 Weizenbaum [19] publishes the description of DOCTOR, the ELIZA script that simulated a psychotherapeutic interview through simple rules of syntactic rephrasing, without comprehension of the content nor coherent memory of the conversation. Users' reactions, documented more extensively in the subsequent volume [20], show attribution of understanding, therapeutic intentions, and empathic capacities to a system that lacked those capacities: the first historically documented case of anthropomorphism toward a system that operates exclusively through language.

ELIZA is relevant for two distinct reasons in the economy of the argument. On the analogical level, it historically anticipates the phenomenon that Floridi terms semantic pareidolia and that in psychology is defined as anthropomorphism. On the causal level, it isolates the activation vector: coherent linguistic output is sufficient to trigger the attribution of understanding and intentions in the absence of any visual component. The facial detection system documented in the previous section is not involved: the active mechanism is the attribution of mental states to an entity whose linguistic behavior satisfies the formal conditions of intentional interaction.

Anthropomorphism is used throughout this paper in its technical psychological sense as established by Epley, Waytz, and Cacioppo [10]: the attribution of human characteristics, motivations, intentions, and emotions to non-human agents, independently of their physical appearance. This usage is distinct from the colloquial sense in which the term designates an aesthetic convention giving non-human entities a human physical form. The phenomenon documented here is mentalistic, not morphological: it concerns the attribution of a mind, not of a body. As a psychological phenomenon, anthropomorphism is distinct from perceptual pareidolia: pareidolia concerns the detection of faces where there are none; anthropomorphism concerns the attribution of mental states to entities that do not possess them. The psychological literature documents that this tendency is not occasional: it is a structural cognitive modality, whose intensity varies in a systematic and measurable way among individuals.

Epley and colleagues [10] identify three factors that operate independently and add up: knowledge about human agency, immediately available in memory and used as an interpretive scheme for any complex behavior; the motivation to understand and predict the environment, for which attributing intentions is the most rapid way of rendering comprehensible an entity that acts in an articulated way; the motivation for social connection, for which attributing a mind to an object or an animal makes it an interactive partner. The more the three factors are simultaneously active, the more intense the anthropomorphic tendency. Waytz, Gray, Epley, and Wegner [11] extend the framework by showing that the perception of another's mind is organized along two independent dimensions, the capacity to act (agency) and the capacity to have experiences (experience), and that agents to whom a mind is attributed along both dimensions are treated as objects of full moral consideration.

Unlike the facial detection system, which operates on visual stimuli, anthropomorphism operates on any entity whose behavior appears articulated and directed to a purpose, and operates an attribution of meaning, which brings this phenomenon close to Floridi's semantic pareidolia. This mechanism can be easily activated by technologies such as LLMs: LLMs do not produce facial configurations, but they produce coherent, purpose-oriented behavior that is contingent with respect to the input, activating the factors that Epley and colleagues identify as determinants of anthropomorphism. The resulting attribution of agency and experience, in the structure of Waytz et al., leads to the treatment of the model as a morally considerable entity.

The individual variability of this tendency is stable and measurable. Waytz and collaborators [12] develop the Individual Differences in Anthropomorphism Questionnaire (IDAQ), administered

on large samples and repeated over time: the propensity to anthropomorphize presents marked differences among individuals but turns out to be substantially constant in the same person after several months. Those who obtain high scores more easily attribute complex emotions to animals and objects, consider their well-being more morally relevant, and trust more in technological devices with anthropomorphic traits.

The two phenomena treated separately so far, seeing faces where there are none and attributing human characteristics to entities that do not possess them, are not independent processes: the perception of a facial configuration carries with it an attribution of human traits, measurable both neurologically and behaviorally. The study by Kühn and colleagues [13] combined functional magnetic resonance imaging with subjective evaluations of humanity on photographs of automobile front ends: the intensity of activation of the brain areas dedicated to face recognition predicts the humanity score that the same person attributes to the automobile. The neural response to facial detection feeds the evaluative judgment on the humanity of the object.

To this is added convergent evidence on the behavioral level. Landwehr and colleagues [14] show that front ends perceived as aggressive correlate with sales success, an effect mediated by personality attributions. Windhager and colleagues [15] document that the geometric characteristics of the front end determine attributions of traits, personality, age, gender, emotion, with a morphometric pattern analogous to that observed for human faces; a cross-cultural replication on Austrian and Ethiopian samples confirms the child-adult dimension [16]. Aggarwal and McGill [17] extend the logic to commercial products in general. Wardle and colleagues [18] finally show that the projection also concerns characteristics of social identity: on 256 images of illusory faces in inanimate objects, participants systematically attribute emotional expression, age, and gender, with a marked bias toward male attribution in an approximate ratio of four to one, an effect also present in children aged 4–8 years.

The set of these pieces of evidence, from the neuronal co-localization between response to real faces and response to facelike objects, to the stability of the effect throughout the life course and across species, to the attribution of social identity and mental capacities to entities lack them, indicates that perceptual pareidolia and anthropomorphism are not separate processes but the two poles of a continuum: the detection of characteristics structurally analogous to human ones activates the attribution of what such characteristics normally entail.

The causal contribution, however, remains incomplete without identifying the cognitive system specifically responsible for the activation in the presence of structured linguistic output. Theory of Mind, introduced in the following section, adds another level of analysis: the evolutionary mechanism that explains why coherent language is connected to the attribution of intentions.

3. Theory of Mind as biological adaptation: evolutionary basis and implications for LLM anthropomorphism

Theory of Mind receives extended treatment in this section because it constitutes the primary activation mechanism of semantic pareidolia toward LLMs: it is the system that specifically connects structured linguistic output to the automatic attribution of intentional agency, and it is therefore the system that any intervention strategy — on the design side as on the user side — must account for.

From the developmental psychology literature it emerges that the capacity to attribute mental states to others, intentions, beliefs, desires, emotions, is not a competence that is acquired through cultural exposure or individual learning. It is a biologically specified adaptation: the literature describes it as a cognitive system that emerges robustly in all typical human environments, independently of specific cultural variations, and that is absent or structurally different in the species closest to humans. Premack and Woodruff [21] have provided its operational definition: Theory of Mind (ToM) is the capacity to attribute mental states to oneself and to others and to use such attributions to explain and predict behavior.

That it is an adaptation and not an acquisition has precise implications on the strength and systematicity with which it operates. An adaptation shaped by natural selection is calibrated on a specific type of stimuli, those that in the evolutionary environment were the reliable signals of mental presence. The type of stimuli on which ToM is calibrated emerges clearly from the interspecies comparison and from the analysis of the neural substrate that sustains it.

The selective pressure that has shaped ToM is identifiable in life in human cultural groups: contexts that require purpose-oriented coordination, transmission of practices between generations, negotiation of shared meanings. These contexts select a specialized set of sociocognitive abilities that do not reduce to general intelligence. Herrmann and colleagues [22] show it directly: administering a comparative cognitive battery, a standardized set of tests that measure different abilities systematically, allowing direct comparison between species, to chimpanzees, orangutans, and two-and-a-half-year-old human children, the results document comparable performances in the physical domain, comprehension of space, quantity, causality, but extremely different in the social domain in favor of human children. The distinctiveness of human cognition does not reside in a general intellectual superiority but in a specialized set of abilities to participate in forms of shared cognition. This specialization has a direct implication for the present argument: ToM is not a general-purpose reasoning capacity that can be selectively inhibited by correct information about the nature of an interlocutor. It is a domain-specific system, calibrated by selection on a specific class of stimuli, that operates independently of general knowledge.

Tomasello and colleagues [23] document that comprehension of others' intentions ontogenetically precedes the acquisition of language and constitutes its necessary condition. At 12–14 months, before language emerges systematically, children analyze others' actions as intentional choices among possible alternatives and attribute to others states of knowledge different from their own. In a paradigmatic study [24], 14-month-old infants observe an adult who turns on a light by touching it with the head: when the adult's hands are free, children imitate the gesture, inferring that that choice had an intentional motive; when the hands are occupied, they do not imitate, inferring that the gesture was constrained by the circumstance. In a second study [25], infants aged 12 and 18 months, faced with three objects of which two were already familiar to the adult and one was new, spontaneously hand over the new object when the adult expresses generic interest: they attribute to the adult a state of knowledge different from their own even before being able to linguistically communicate this distinction. Language develops on this basis: it is a tool of bidirectional coordination that presupposes, in its evolutionary structure, the comprehension that the interlocutor possesses communicative intentions, mental states oriented to modifying the mental state of the listener. Receiving a coherent linguistic message activates the intention-reading system not by analogy with other signals but constitutively: language is, in the documented evolutionary architecture, an intentional signal by definition.

Visual pareidolia has shown us that the tendency to recognize faces where there are none is biologically specified, anthropomorphism has shown us that the attribution of mental states is activated on entities with articulated and contingent behavior, Theory of Mind, in the documentation of Tomasello and colleagues, specifies a further mechanism: the evolutionary system that regulates the comprehension of others' intentions develops in constitutive connection with language.

The neural substrate of these processes is identifiable in a distributed network. The temporoparietal junction is activated when it is necessary to infer the perspective of another agent, what he sees, from where he sees it, and in response to the attribution of false beliefs, that is, to the recognition that another human being can have a representation of the world different from the real one: it tracks the direction of others' gaze and reconstructs what the other can perceive from his position. The temporal poles integrate contextual knowledge accumulated on specific people and situations, allowing to anticipate what an individual is likely to think or feel in a given context even before he shows observable signs of his state. The medial prefrontal cortex is activated in the anticipation of future mental states of the interlocutor, with a documented internal segregation: reasoning about people similar to oneself activates a more ventral portion, reasoning about

dissimilar people a more dorsal portion [26]. The common characteristic of these three regions is that their activation occurs largely automatically, before the observer reflectively processes the content of the received message. The automaticity of this neural response defines the structural problem for intervention: any measure that operates on explicit representation — informing, correcting, notifying — arrives after the implicit mentalizing response has already been activated. The relevant level of the phenomenon is prior to the level at which declarative communication operates.

This automaticity corresponds to the operation of the efficient-implicit system described by Apperly and Butterfill [27], who propose that the comprehension of mental states, including the capacity to attribute false beliefs, requires two distinct cognitive systems operating in parallel. The efficient-implicit system operates rapidly and without significant costs of working memory or executive control: it sustains the implicit mentalistic behaviors already observable in infants and active in the course of adult social interactions in real time. The flexible-reflective system is computationally more onerous: it allows explicit reasoning on arbitrarily complex beliefs, operates sequentially, and develops gradually in association with language and executive functions. The two architectures operate in parallel in real interactions with qualitatively different temporal properties: the efficient-implicit system precedes the flexible-reflective system in the processing chain.

Visual pareidolia has already presented itself as a phenomenon that operates before reflective elaboration. The causal contribution of implicit ToM presents a further automatism: in a linguistic interaction the automatic attribution of intentional agency would occur before the flexible-reflective system can elaborate any declarative representation of the source of that language.

The activation of the ToM system is not, however, indiscriminate. Moriguchi and colleagues [28] have documented that children show the typical pattern of perseveration in response to human actions but not in response to robotic actions: the mentalizing apparatus discriminates the agent. Itakura and colleagues [29] specify the criterion of discrimination: when a robot executes actions with a clearly purpose-oriented structure, children show imitation patterns analogous to those observed with human agents. The relevant variable is not the biological nature of the agent but the structural characteristics of the action, coherence, purpose orientation, contingency with respect to the interlocutor's input. This result isolates the specific variable that defines the vulnerability of LLM interactions: coherent, purpose-oriented, contingent linguistic output is precisely the behavioral profile that current conversational models are designed — and increasingly optimized — to produce. The activation of the implicit mentalizing system in response to LLMs is therefore not an accidental byproduct of their design but a predictable consequence of it.

The confluence of these elements produces a direct implication for the understanding of semantic pareidolia. LLMs produce coherent, purpose-oriented linguistic output that is contingent with respect to the user's input: exactly the formal profile that the efficient-implicit ToM system is calibrated to treat as an indicator of intentional agency. To this is added the fact, documented by Tomasello and colleagues [25], that language is constitutively linked to intentional communication in the evolutionary architecture of human cognition: the intention-reading system does not have a mechanism to distinguish an interlocutor who possesses communicative intentions from a system that produces its formal signals. The systematicity of the pareidolic phenomenon toward LLMs does not reflect a cultural prejudice nor an occasional distortion of cognitive processing: it reflects the functional profile of an evolutionary adaptation in the presence of stimuli that satisfy its activation conditions.

4. From cultural antibodies to psychological brakes

The previous sections have documented a convergent apparatus. Perceptual pareidolia establishes that the detection of human-like structure in stimuli is a pre-cultural, pre-reflective mechanism, anchored to neural circuits shared with non-human primates and activated before any reflective elaboration. Anthropomorphism documents the extension of mental attribution beyond the visual domain: coherent linguistic output is sufficient — as ELIZA first demonstrated — to trigger

attribution of understanding and intentions, and the tendency generalizes to any entity displaying contingent, purpose-directed behavior. Theory of Mind specifies the primary activation mechanism: the efficient-implicit system connects structured linguistic output to the automatic attribution of intentional agency by architectural definition, because language is constitutively linked to intentional communication in the evolutionary architecture of human cognition. All three systems operate at the implicit level, prior to and independently of what the user explicitly knows about the model with which they interact. It is this convergence that defines the nature of the challenge: not an erroneous belief correctable by information, but a biologically specified modality of cognitive treatment activated independently of explicit knowledge. The following sections derive the operational consequences of this finding.

4.1. Ontological misunderstanding and its psychological consequences

The construct of semantic pareidolia rests on a necessary presupposition: simulating a capacity is not equivalent to possessing it. In the absence of this principle, the phenomenon described by Floridi [1] would not configure any problem, the attribution of consciousness to a model that simulates it would be ontologically justified, and the distinction between simulation and possession would turn out to be without consequences. It is instead from this presupposition that the misunderstanding becomes identifiable as such, and that its consequent interpretive and use practices can be defined as illusory and potentially harmful.

The psychological scientific community is in fact responding to some claims regarding the fact that LLMs manage to pass linguistic tests that attest to ToM. The distinction between simulating and possessing a capacity finds direct empirical anchorage in the recent literature on the performance of LLMs on Theory of Mind tasks. Marchetti and colleagues [30] specify that some LLMs produce outputs that, evaluated with superficial measures, may appear indicative of mentalistic capacities, but not be so. The researchers specify that the phenomenon reflects two distinct orders of problems. The first order of problems is structural to the model. The performances of LLMs on ToM tasks depend on the recognition of linguistic patterns in the training data: when a model correctly responds to a false-belief task (recognition that another human being can have a representation of the world different from the real one), it produces the response formally expected by the psychological test thanks to a statistical processing of the correct response, not because it has performed the inference, which instead a human being would have to process to respond in the same way. This distinction, between statistically correct response and genuine understanding, remains difficult to detect as long as the evaluation is limited to the textual domain. It instead becomes evident as soon as one considers the multimodal and embodied dimensions of human ToM: the reading of expressions, the comprehension of gesture, the coordination of shared attention in contexts of physical interaction. On these dimensions LLMs do not have equivalent capacities, because their architecture is confined to the processing of text. Marchetti and colleagues document that most of the existing evaluations systematically neglect these dimensions, concentrating on tasks that structurally favor the strong points of linguistic models, and producing for this reason an inflated estimate of their mentalistic capacities. The second order of problems is methodological and concerns the researchers who test the ToM of LLMs: the studies that evaluate the mentalistic capacities of LLMs adapt classic tasks to the characteristics of the models, eliminating exactly the ambiguities that define real social cognition. Tasks designed to measure human ToM are simplified to make them compatible with the strong points of LLMs, producing an artificial inflation of apparent performances. The result, in the formulation of Marchetti and colleagues, is an illusion of understanding: not only do the models simulate understanding without possessing it, but the evaluation methods actively contribute to constructing this illusion.

Pang and colleagues [31] apply Morgan's Canon to the case: before attributing to LLMs a genuine ToM capacity, it is necessary to exclude simpler explanations. The apparently mentalistic performances turn out to be more parsimoniously explainable as associative learning on textual

corpora produced by human beings endowed with ToM, without requiring the attribution of a mentalistic capacity to the model. Verma and collaborators [32] document the structural fragility of these performances under conditions of contextual stress: systematic collapse when prompts are perturbed with contradictory information or impoverished contexts, a profile incompatible with a genuine capacity for attribution of mental states, and compatible instead with the pattern recognition that Morgan's Canon indicates as a sufficient explanation.

The convergence of these results circumscribes the distinction in operational terms. Marchetti and colleagues introduce the construct of Artificial ToM (AToM) as a formally separate category from human ToM: AToM designates the capacity of LLMs to produce responses aligned with human performances on ToM tasks, not a cognitive equivalent of these. What LLMs produce is a functional approximation of mentalistic behavior, lacking the affective component, lacking the embodied dimension, lacking the recursiveness that characterizes adult ToM, and dependent on test conditions that structurally favor the strong points of linguistic models. The psychological consequences of treating a functional approximation as a cognitive equivalent are documentable at the behavioral level.

The literature on anthropomorphism documented above indicates that agents to whom a mind is attributed along the dimensions of agency and experience are treated as reliable sources of judgment [11]. A model to which genuine understanding is attributed receives a degree of trust calibrated on that attribution, not on the actual structural characteristics of the model. The direction of this effect is empirically documented in a systematic way: a systematic review by Aquilino et al. on 45 studies on the measurement of trust in AI models shows that the level of perceived anthropomorphism of the model determines which component of trust is preferentially activated; models with characteristics more similar to a human interlocutor produce predominantly affective trust, founded on emotional connection and perception of social presence, independently of the evaluation of the actual capacities of the model [33]. The same review documents that this affective path is activated for models that use natural language as a modality of interaction, independently of the level of awareness of users regarding the artificial nature of the model. At the level of primary studies, Cheng and colleagues [34] document that when consumers attribute greater warmth or greater competence to a chatbot on the basis of the interaction, they develop significantly higher levels of trust, an effect mediated by the attribution of human mental characteristics to the model, not by the evaluation of its technical performances.

In a decisional context, medical, legal, financial, military, the miscalibration could produce decisions founded on an erroneous attribution of competence. In this case it is not possible to say that the model made a mistake because it is technically unreliable, but that the user has been victim of the psychological mechanisms that are inherent in him and that the design of the tool accommodates and feeds, allowing him to attribute a type of reliability, that founded on intentional understanding of others' mental states, that the model structurally does not possess, but continues to simulate.

The result is the potential progressive reduction of reliability of the human being's evaluative process, a potential reduction that, however, finds fertile ground in the cognitive-psychological structure of the human being itself. The efficient-implicit ToM system automatically attributes intentions, independently of a specific evaluation of the actual substantiality of these manifested intentions.

These effects do not require that the user explicitly believes that the model is conscious. The summary of psychological research set out in the previous sections indicates that the attribution of intentional agency occurs automatically, at the level of the implicit system, independently of what the user explicitly conceives about the nature of the model. The relevant ontological misunderstanding is not a declared belief, it is a modality of cognitive treatment that is activated independently of the explicit beliefs of the user and that produces measurable behavioral effects.

4.2. Structural insufficiency of current mitigation measures

Having established the ontological misunderstanding and its psychological consequences — miscalibrated trust activated at the implicit level, independently of the user's explicit beliefs about the model — it is now possible to evaluate why current mitigation measures are structurally insufficient to address it. Floridi [1] documents that some developers have already attempted mitigation interventions, disclaimers, collaborations with psychologists, reminders of the artificial nature of the model, and judges them insufficient, observing that these efforts are often annulled by the same producers in other contexts.

The development trajectory of conversational models, from ELIZA to current models, has always been and still is oriented toward the achievement of a growing perceived similarity to the human interlocutor. The explicit pragmatic justification of this choice has never been formulated: it is not documented in terms of efficiency or effectiveness for what reason a more human-like model should be preferred to a less human-like one for the carrying out of the tasks these models are called to execute. The only systematically documented correlation is that between perceived similarity to the human interlocutor and trust granted to the model: the level of perceived anthropomorphism constitutes the primary predictor of the type and intensity of trust developed by the user, with models that use natural language, the category in which conversational LLMs are, producing preferentially affective trust independently of the evaluation of the actual capacities of the model, as documented by the systematic review already cited [33]. A company that perfects its own models toward a growing and increasingly refined simulation of human interaction, without a declared pragmatic justification, is, consciously or not, accommodating and heightening a documented psychological tendency in order to increase the propensity to use its own products. Floridi [1] identifies in the economic drive of large companies that speculate on the phenomenon for profit purposes one of the forces that intensify semantic pareidolia; the same author observes that often company executives are the first to publicly support narratives that attribute human or proto-conscious characteristics to the models [1].

In this framework, the mitigation interventions currently adopted present a structural incongruence: they operate on the level of explicit communication while the phenomenon they intend to counter operates on a different level. A disclaimer, an utterance that informs the user of the artificial nature of the model, operates on the plane of declarative representation: it requires that the user read, elaborate, and integrate information into his own representation of the interaction. The phenomenon that it intends to counter operates on a different plane: it is the response of a biologically specified cognitive apparatus to stimuli that satisfy its activation conditions. It is not an erroneous opinion that correct information can correct, it is the physiological functioning of an evolutionarily calibrated model. The disclaimer operates once, at a defined point of the interaction; the tendency to anthropomorphize operates continuously, on every conversational turn, through a cognitive apparatus that is activated independently of what the user explicitly knows about the model with which he interacts.

Countering a process rooted in the ethology of the species requires something more than a notification: it requires a work of mental setting that renders stable, in the user, the awareness of what the model with which he interacts really is, a statistical elaboration of linguistic data on a massive scale, structurally incapable of developing mental states. Whether and to what extent this type of cognitive preparation is capable of modulating the automatic activation of the mentalization apparatus is an open empirical question. What is structurally inadequate is the assumption that the sole declarative transmission of information is sufficient to produce this effect.

4.3. A conceptual orientation for intervention: Psychological Brakes

The structural mismatch documented in 4.2 — between the declarative level of current interventions and the implicit, automatic level of the phenomenon — defines the design space for psychological brakes. The concept designates a category of intervention qualitatively distinct from the disclaimer and the notification: not a corrective message directed at the user's explicit beliefs,

but a set of design and educational choices calibrated on the biologically specified character of the mentalizing tendency — a tendency that operates before and independently of what the user explicitly knows about the model with which he interacts. Psychological brakes are not intended to eliminate the tendency, which is neither possible nor the relevant objective; they are intended to reduce the degree to which the interaction is designed, consciously or not, to maximize its activation. Two sides of the dynamic are addressable — that of the developers and that of the users — and four lines of intervention are identifiable across them.

On the side of the developers:

Communicative profile regulation: The definition of a normative framework governing the choices that determine the level of perceived similarity to a human interlocutor — grammatical person, simulation of emotional involvement, subjective tone — independently of constraints on generated content. A model can fully respect existing content constraints and simultaneously be trained to present itself in a way that maximizes perceived human-likeness. Some developers have already initiated normative frameworks on content: Anthropic has published a document called Constitution, which defines values, expected behaviors, and ethical constraints incorporated in the training process of their model Claude. A parallel framework governing the communicative profile of the model does not yet exist and constitutes the first direction of intervention on this side.

The connection to the ToM mechanism is direct: the criteria that the efficient-implicit system uses to activate mentalizing — coherence, purpose orientation, contingency, and subjective tone — are precisely the properties that a communicative profile oriented toward human-likeness is designed to maximize. Regulating the communicative profile means acting on the stimulus before the implicit system is activated, rather than correcting its output after the fact. At the level of feasibility, this direction requires intervention at the training stage, not only at the deployment stage: communicative profile choices are embedded in the reward signals that shape model behavior and cannot be fully controlled through post-hoc output filtering.

Pragmatic justification requirement: The explicit declaration of the functional reason for which design choices oriented toward perceived similarity to the human interlocutor are adopted. The only systematically documented correlation between human-likeness and model performance is that between perceived anthropomorphism and the affective trust granted by the user [33] — a correlation that, as documented in 4.1, produces miscalibration rather than reliability. A developer that optimizes its model toward growing perceived similarity without a declared functional justification is, as Floridi [1] observes, accommodating and heightening a documented psychological tendency in order to increase the propensity to use its own products. The explicit requirement of pragmatic justification makes this choice visible and accountable.

This brake differs structurally from the other three: it does not operate on the cognitive system of the user but on the transparency of the developer's design choices. Making the pragmatic justification requirement explicit functions as a precondition for informed governance of the other three brakes. In terms of feasibility, this requirement is analogous to disclosure obligations in other domains where design choices have documented effects on user cognition and behavior — pharmaceutical labeling, financial product documentation, algorithmic transparency requirements. The institutional actors positioned to impose it include regulatory bodies, certification standards, and procurement requirements from public institutions that deploy LLM-based services.

Narrowing design: The explicit definition of the technical and market need that LLMs are called to satisfy. If the objective is to support the carrying out of tasks — elaborating information, automating processes, supporting decisions — there is no functional necessity that the interface of use be a conversation in natural language that simulates a sociable human interlocutor. The chat is a design choice, not a technical requirement: structured and task-oriented interfaces could reduce the activation of the mentalizing system without reducing the functionality of the tool, configuring themselves as a direct form of psychological brake at the interface level. If, on the other hand, the objective is that of the life companion, questions of legitimacy arise that exceed the perimeter of the present contribution and that require a level of psychological and normative attention that the field has not yet systematically developed.

The connection to the ToM mechanism is specific: the efficient-implicit mentalizing system is activated by output that is coherent, purpose-oriented, and contingent with respect to the user's input — but the conversational format, which simulates the turn-taking structure of human dialogue, adds a layer of social contingency that is not functionally required for task execution. A structured, task-oriented interface can preserve the informational contingency of the output while reducing the social-interactional contingency that the implicit system treats as an indicator of intentional agency. This is not a purely theoretical direction: command-line interfaces, structured query systems, and form-based tools represent existing design paradigms that reduce conversational framing without reducing functional output quality. Whether reduced conversational framing produces measurable reductions in implicit anthropomorphism is an empirical question that existing interface design research has not yet systematically addressed in the context of LLMs.

On the side of the users:

User cognitive preparation: The possibility of intervening on the communicative profile of the model already exists partially on the user's side: some platforms allow the user to configure the communicative behavior of the model through personal settings — modifying the register, reducing the characteristics of subjectivity simulation, orienting the interaction toward a more impersonal and task-oriented profile. This option, however, requires that the user be aware of the problem he intends to manage. More broadly, the preparation required is not limited to disseminating a correct representation of the functioning of these models. Such dissemination is a necessary but insufficient condition: it operates on the level of declarative knowledge, while the mentalizing tendency it addresses operates at the implicit level, activated before and independently of explicit knowledge. The cognitive preparation calibrated on the implicit and automatic character of the ToM system must include the comprehension that the attribution of intentional agency occurs before what the user explicitly knows about the model; this comprehension itself must become part of the preparation, constituting the preliminary cognitive condition for a mental setting capable of modulating, even partially, the automatic attribution of intentional agency to structured linguistic stimuli.

The distinction between declarative knowledge and preparation calibrated on the implicit system has concrete implications for what this brake would need to contain. Informing users that LLMs are not conscious — the standard content of current educational efforts — operates at the level of explicit belief and leaves the efficient-implicit mentalizing system unaddressed, because that system is activated by structural properties of the stimulus independently of what the user explicitly believes about its source. Preparation calibrated on the implicit system would need to include, at minimum, experiential exposure to the gap between the model's apparent intentional coherence and its actual processing architecture — not as a declarative fact but as an experiential understanding that the contingency of the output is produced by statistical pattern matching, not by communicative intention. Whether such preparation can produce durable modulation of implicit mentalizing responses is an open empirical question; research on the reduction of implicit biases through metacognitive training offers a partial analogy, though the transfer to mentalizing specifically has not been documented.

5. Conclusions

The psychological and neuroscientific literature reviewed in this paper documents that the tendency to attribute intentional agency to large language models is not a contingent interpretive error correctable through declarative means. Perceptual pareidolia establishes that the biological architecture of humanization responds to structural properties of stimuli before reflective elaboration. Anthropomorphism documents that this attribution extends to any entity displaying coherent, purpose-directed, contingent behavior — precisely the profile that LLMs systematically produce. Theory of Mind specifies the primary activation mechanism: the efficient-implicit system that connects structured linguistic output to the automatic attribution of intentional agency

operates by evolutionary definition, without a mechanism to distinguish a biological interlocutor from a system that produces its formal signals.

The convergence of these three systems at the implicit level has a direct consequence for the evaluation of existing mitigation strategies. The distinction between simulating and possessing a capacity — the necessary presupposition of Floridi's concept of semantic pareidolia — finds empirical anchorage in the recent literature on LLM performance on Theory of Mind tasks: what these models produce is a functional approximation of mentalistic behavior, lacking the affective component, the embodied dimension, and the recursiveness that characterize adult ToM, and inflated by evaluation methods that adapt tasks to the structural strengths of linguistic models. The psychological consequences of this misunderstanding — miscalibration of trust that preferentially activates affective paths independently of the model's actual capacities, and progressive reduction of the user's evaluative autonomy — are produced at the level of the implicit system, independently of what the user explicitly believes about the nature of the model with which they interact.

From this follows the structural insufficiency of current interventions. Disclaimers and notifications on the artificial nature of the model operate on the level of declarative representation, a level that is structurally posterior to the implicit activation they propose to neutralize, and active only once with respect to a process that reactivates at every conversational turn. The development trajectory of LLMs toward growing perceived similarity to the human interlocutor, in the absence of a documented pragmatic justification, operates in the same direction as the phenomenon one declares to want to counter. The four psychological brakes proposed in Section 4.3 — communicative profile regulation, pragmatic justification requirement, narrowing design, and user cognitive preparation — constitute directions of empirical research rather than validated prescriptions, but their calibration on the implicit level of the phenomenon distinguishes them structurally from the declarative interventions whose insufficiency has been documented.

Limitations

The contribution is a selective narrative review with argumentative function, not a systematic review: the selection of studies is oriented to the construction of the conceptual framework and does not exhaustively cover the available literature on the individual phenomena treated.

The four psychological brakes proposed — communicative profile regulation, pragmatic justification requirement, narrowing design, and user cognitive preparation — are directions of research coherent with the empirical foundation developed, not validated prescriptions. The effectiveness of each remains an open empirical question that requires specific experimental designs calibrated on the implicit level at which the phenomenon operates.

The case of LLMs as life companion is explicitly excluded from the analysis and configures a class of psychological effects, attachment, dependence, substitution of social relations, that would require a separate conceptual framework, as well as a level of normative attention that the field has not yet systematically developed.

Declaration on Generative AI

While writing this paper, the author used Claude (Anthropic, claude.ai) to: check for grammar and spelling errors, paraphrase and rephrase the text, and verify bibliographic citations. After using this tool/service, the author reviewed and edited the content as necessary and assumes full responsibility for the content of this publication.

References

[1] L. Floridi, "AI and Semantic Pareidolia: When We See Intelligent Consciousness Where There Is None", Philosophy & Technology 39 (2026) 36. DOI: 10.1007/s13347-026-01052-1.

[2] N. Hadjikhani, K. Kveraga, P. Naik, S. P. Ahlfors, “Early (M170) activation of face-specific cortex by face-like objects”, NeuroReport 20 (2009) 403–407. DOI: 10.1097/WNR.0b013e328325a8e1.
[3] S. G. Wardle, J. Taubert, L. Teichmann, C. I. Baker, “Rapid and dynamic processing of face pareidolia in the human brain”, Nature Communications 11 (2020) 4518. DOI: 10.1038/s41467-020-18325-8.
[4] J. Liu, J. Li, L. Feng, L. Li, J. Tian, K. Lee, “Seeing Jesus in toast: Neural and behavioral correlates of face pareidolia”, Cortex 53 (2014) 60–77. DOI: 10.1016/j.cortex.2014.01.013.
[5] B. Cerrahoğlu, C. Jacques, D. Rekow, J. Jonas, S. Colnat-Coulbois, S. Caharel, A. Leleu, B. Rossion, “The neural basis of face pareidolia with human intracerebral recordings”, Imaging Neuroscience 3 (2025) imag_a_00518. DOI: 10.1162/imag_a_00518.
[6] M. Kato, R. Mugitani, “Pareidolia in infants”, PLoS ONE 10 (2015) e0118539. DOI: 10.1371/journal.pone.0118539.
[7] R. T. Keys, J. Taubert, S. G. Wardle, “A visual search advantage for illusory faces in objects”, Attention, Perception, & Psychophysics 83 (2021) 1942–1953. DOI: 10.3758/s13414-021-02267-4.
[8] B. W. Saurels, N. Peluso, J. Taubert, “A behavioral advantage for the face pareidolia illusion in peripheral vision”, Scientific Reports 14 (2024) 10238. DOI: 10.1038/s41598-024-60892-z.
[9] J. Taubert, S. G. Wardle, M. Flessert, D. A. Leopold, L. G. Ungerleider, “Face pareidolia in the rhesus monkey”, Current Biology 27 (2017) 2505–2509. DOI: 10.1016/j.cub.2017.06.075.
[10] N. Epley, A. Waytz, J. T. Cacioppo, “On seeing human: A three-factor theory of anthropomorphism”, Psychological Review 114 (2007) 864–886. DOI: 10.1037/0033-295X.114.4.864.
[11] A. Waytz, K. Gray, N. Epley, D. M. Wegner, “Causes and consequences of mind perception”, Trends in Cognitive Sciences 14 (2010) 383–388. DOI: 10.1016/j.tics.2010.05.006.
[12] A. Waytz, J. Cacioppo, N. Epley, “Who sees human? The stability and importance of individual differences in anthropomorphism”, Perspectives on Psychological Science 5 (2010) 219–232. DOI: 10.1177/1745691610369336.
[13] S. Kühn, T. R. Brick, B. C. N. Müller, J. Gallinat, “Is this car looking at you? How anthropomorphism predicts fusiform face area activation when seeing cars”, PLoS ONE 9 (2014) e113885. DOI: 10.1371/journal.pone.0113885.
[14] J. R. Landwehr, A. L. McGill, A. Herrmann, “It’s got the look: The effect of friendly and aggressive ‘facial’ expressions on product liking and sales”, Journal of Marketing 75 (2011) 132–146. DOI: 10.1509/jmkg.75.3.132.
[15] S. Windhager, D. E. Slice, K. Schaefer, E. Oberzaucher, T. Thorstensen, K. Grammer, “Face to Face: The perception of automotive designs”, Human Nature 19 (2008) 331–346. DOI: 10.1007/s12110-008-9047-z.
[16] S. Windhager, F. L. Bookstein, K. Grammer, E. Oberzaucher, H. Said, D. E. Slice, T. Thorstensen, K. Schaefer, “‘Cars have their own faces’: Cross-cultural ratings of car shapes in biological (stereotypical) terms”, Evolution and Human Behavior 33 (2012) 109–120. DOI: 10.1016/j.evolhumbehav.2011.06.003.
[17] P. Aggarwal, A. L. McGill, “Is that car smiling at me? Schema congruity as a basis for evaluating anthropomorphized products”, Journal of Consumer Research 34 (2007) 468–479. DOI: 10.1086/518544.
[18] S. G. Wardle, S. Paranjape, J. Taubert, C. I. Baker, “Illusory faces are more likely to be perceived as male than female”, Proceedings of the National Academy of Sciences 119 (2022) e2117413119. DOI: 10.1073/pnas.2117413119.
[19] J. Weizenbaum, “ELIZA — A computer program for the study of natural language communication between man and machine”, Communications of the ACM 9 (1966) 36–45. DOI: 10.1145/365153.365168.
[20] J. Weizenbaum, Computer Power and Human Reason: From Judgment to Calculation, W. H. Freeman, San Francisco, 1976.

[21] D. Premack, G. Woodruff, “Does the chimpanzee have a theory of mind?”, Behavioral and Brain Sciences 1 (1978) 515–526. DOI: 10.1017/S0140525X00076512.

[22] E. Herrmann, J. Call, M. V. Hernández-Lloreda, B. Hare, M. Tomasello, “Humans have evolved specialized skills of social cognition: The cultural intelligence hypothesis”, Science 317 (2007) 1360–1366. DOI: 10.1126/science.1146282.

[23] M. Tomasello, M. Carpenter, J. Call, T. Behne, H. Moll, “Understanding and sharing intentions: The origins of cultural cognition”, Behavioral and Brain Sciences 28 (2005) 675–691. DOI: 10.1017/S0140525X05000129.

[24] G. Gergely, H. Bekkering, I. Király, “Rational imitation in preverbal infants”, Nature 415 (2002) 755. DOI: 10.1038/415755a.

[25] M. Tomasello, K. Haberl, “Understanding attention: 12- and 18-month-olds know what’s new for other persons”, Developmental Psychology 39 (2003) 906–912. DOI: 10.1037/0012-1649.39.5.906.

[26] C. D. Frith, U. Frith, “The Neural Basis of Mentalizing”, Neuron 50 (2006) 531–534. DOI: 10.1016/j.neuron.2006.05.001.

[27] I. A. Apperly, S. A. Butterfill, “Do humans have two systems to track beliefs and belief-like states?”, Psychological Review 116 (2009) 953–970. DOI: 10.1037/a0016923.

[28] Y. Moriguchi, T. Kanda, H. Ishiguro, S. Itakura, “Children perseverate to a human’s actions but not to a robot’s actions”, Developmental Science 13 (2010) 62–68. DOI: 10.1111/j.1467-7687.2009.00860.x.

[29] S. Itakura, H. Ishida, T. Kanda, Y. Shimada, H. Ishiguro, K. Lee, “How to build an intentional android: Infants’ imitation of a robot’s goal-directed actions”, Infancy 13 (2008) 519–532. DOI: 10.1080/15250000802329503.

[30] A. Marchetti, F. Manzi, G. Riva, A. Gaggioli, D. Massaro, “Artificial Intelligence and the Illusion of Understanding: A Systematic Review of Theory of Mind and Large Language Models”, Cyberpsychology, Behavior, and Social Networking 28 (2025) 505–514. DOI: 10.1089/cyber.2024.0536.

[31] D. K. F. Pang, S. K. Y. Pang, M. D. Broeker, A. Hibble, “Do large language models have a theory of mind?”, Proceedings of the National Academy of Sciences 122 (2025) e2507080122. DOI: 10.1073/pnas.2507080122.

[32] M. Verma, S. Bhambri, S. Kambhampati, “Theory of Mind abilities of large language models in Human-Robot Interaction: An illusion?”, in: Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction (HRI ‘24), ACM, New York, NY, 2024, pp. 36–45. DOI: 10.1145/3610978.3640767.

[33] L. Aquilino, C. Di Dio, F. Manzi, D. Massaro, P. Bisconti, A. Marchetti, “Decoding Trust in Artificial Intelligence: A Systematic Review of Quantitative Measures and Related Variables”, Informatics 12 (2025) 70. DOI: 10.3390/informatics12030070.

[34] X. Cheng, X. Zhang, J. Cohen, J. Mou, “Human vs. AI: Understanding the impact of anthropomorphism on consumer response to chatbots from the perspective of trust and relationship norms”, Information Processing & Management 59 (2022) 102940. DOI: 10.1016/j.ipm.2022.102940.