www.bircu-journal.com/index.php/birci
Budapest International Research and Critics Institute-Journal (BIRCI-Journal)
Humanities and Social Sciences
ISSN 2615-3076 Online)
ISSN 2615-1715 (Print)
Budapest Institute

# The Legible Mind: How Generative AI Flattens Consciousness, Transforms Communication, and Recenters Human Intelligence on the Illegible

**Muhammad Ridwan[1], Belay Sitotaw Goshu[2], Anantasha Titisania Rimadewi[3]**
[1]*Department of Lingustics, Universitas Islam Negeri Sumatera Utara, Indonesia*
[2]*Department of Physics, Dire Dawa University, Dire Dawa, Ethiopia*
[3]*LSPR Communication and Business Institute - Jakarta Campus, Indonesia*
*bukharyahmedal@gmail.com, belaysitotaw@gmail.com, sania.m@lspr.edu*

## Abstract

*Generative AI produces highly legible outputs fluent, coherent, and statistically probable text. Yet human consciousness, communication, and intelligence are fundamentally shaped by illegible elements: qualia, strategic ambiguity, error, and non derivable insight. This paper introduces the concept of "the legible mind" to examine how generative AI flattens consciousness, transforms communication, and recenters human intelligence in the age of synthetic media. Drawing on philosophy of mind (Nagel, Chalmers, and Searle), communication theory (Grice, Austin), human computer interaction (Norman), and recent empirical studies of AI detection and uncanny valley effects, we develop a conceptual framework distinguishing legible outputs from illegible processes. AI simulates conscious outputs without subjective experience, inverting the Turing test so that humans feel pressure to imitate machinic legibility (e.g., CAPTCHAs). Communication shifts from intention driven cooperation to hyper legibility, provoking strategic illegibility (deliberate errors, personal digressions) as an authenticity signal. Human intelligence recenters on meta legibility: prompting, critique, and integration of AI generated content with non derivable insight. AI makes legibility cheap, rendering the illegible scarce and valuable. The legible mind is a tool; the danger is forgetting the illegible. Interdisciplinary research should investigate how humans perceive illegibility, regulate deceptive AI that simulates human imperfection, redesign education around meta legibility, and develop proof of humanity protocols based on embodied presence.*

Keywords
generative AI; legibility; consciousness; strategic illegibility; meta legibility

## I. Introduction

### 1.1 The Paradox of AI Fluency

In November 2022, when OpenAI released ChatGPT to the public, few anticipated the speed with which generative artificial intelligence would permeate everyday communication, education, creative work, and scientific practice. Within two months, the platform had amassed over 100 million users, making it the fastest-growing consumer application in history. What users discovered was something unprecedented: a machine that could produce remarkably legible outputs, coherent paragraphs, grammatically flawless essays, plausible arguments, even poetry that scanned. Generative AI, particularly large language models (LLMs), generates text that is statistically probable, syntactically well-formed, and pragmatically appropriate to the prompting context. In short, it produces language that is readable, predictable, and reducible to the patterns latent in its training data.

This legibility is, from an engineering standpoint, a triumph. Models such as GPT-4, Claude, and Gemini have been optimized to minimize surprise, to conform to statistical norms, and to produce outputs that satisfy the expectations embedded in human language corpora. As Bender et al. (2021) famously argued these systems operate as "stochastic

parrots": they generate text by sampling from probability distributions learned from vast collections of human-written documents, without any underlying understanding of the meaning those strings of words encode. Yet it is precisely this stochastic fluency that has led many users to anthropomorphize AI systems, attributing to them intentionality, belief, and even consciousness (Dennett, 1987, 1991).

The paradox, however, is that the very fluency and legibility that make generative AI so useful also obscure a deeper divide. Human minds are only partially legible. Our inner lives harbour what philosophers of mind call qualia, the raw, subjective, first-person qualities of experience, such as the redness of red, the pain of a headache, or the pang of nostalgia (Nagel, 1974; Chalmers, 1996). We harbour contradiction: we say one thing and feel another, we profess beliefs we later disavow. We harbour silence: thoughts we cannot articulate, feelings we lack words for, intuitions that precede explicit reasoning. And we harbour error: slips of the tongue, false starts, creative leaps that violate statistical expectation. These features, qualia, contradiction, silence, error are the signatures of a conscious mind that is not reducible to its observable outputs.

### 1.2 Defining "The Legible Mind"

To address these questions, this paper introduces the concept of the legible mind. Legibility is here understood as a property of outputs whether textual, behavioural, or communicative that are readable, predictable, and reducible to explicit rules or statistical patterns. In the context of human-computer interaction and governance, legibility has long been recognized as a design virtue: legible interfaces reduce user confusion, legible institutions increase accountability (Scott, 1998). Norman (2013), for instance, argued that good design makes the system's state and affordances legible to the user, reducing cognitive load and facilitating action.

Extending this notion, the legible mind refers to the mind as it appears through legible artifacts: the texts it produces, the speech acts it performs, the behaviours it emits. This is the mind that can be read, modelled, predicted, and crucially simulated. Generative AI excels at producing the legible mind. It can write a passable cover letter, generate a plausible diagnosis, or engage in what appears to be empathetic conversation, all without any underlying conscious experience (Dennett, 1991; Searle, 1980).

But the legible mind is not the whole mind. In deliberate contrast, this paper reserves the term illegible for those aspects of human cognition and experience that resist easy capture by statistical models or explicit rule-systems: private, first-person experience (qualia); the embodied, situated, and affective dimensions of feeling (Damasio, 1994); the creative leaps that are not derivable from past examples (Boden, 2004); and the strategic ambiguities of poetic or ironic language (Grice, 1975). The illegible is not simply the unexpressed; it is the inexpressible in principle or at least the expressible only at the cost of flattening.

### 1.3 Problem Statement

This paper argues that generative AI forces three interrelated transformations of the human condition. First, it flattens consciousness by simulating the behavioural outputs of conscious experience while being entirely devoid of qualia. As Chalmers (1996) has influentially argued, the hard problem of consciousness consists in explaining why and how physical processes give rise to subjective experience. Generative AI sidesteps this problem entirely: it produces the appearance of phenomenal states without any of their felt reality. The risk is not that AI will become conscious, but that we will mistake legible simulation for the real thing, or worse, that we will flatten our own consciousness to match the legible outputs of machines (Floridi, 2025).

Second, it transforms communication. Where human communication has long relied on shared context, strategic ambiguity, and the cooperative principle articulated by Grice (1975), generative AI introduces a regime of statistical fluency that optimizes for legibility at the expense of meaning. Grice's maxims; quantity, quality, relation, and manner presuppose intentional agents who mutually recognize each other's communicative goals. AI systems, as Searle (1980) argued in his Chinese Room thought experiment, manipulate symbols without understanding their content. The result is a form of communication that is hyper-legible in its syntactic and stylistic conformity but pragmatically hollow. In response, humans may increasingly adopt forms of strategic illegibility; irony, allusion, deliberates error to signal their own mindedness and authenticity.

Third, it recenters human intelligence. For centuries, human intelligence was measured by the ability to produce legible artifacts: essays, proofs, arguments, and works of art. Generative AI commoditizes this production. When an AI can generate a B+ essay in seconds, the value of human-generated legible content collapses unless it is more than legible. The new premium shifts to what this paper calls meta-legibility: the capacity to prompt AI effectively, to critique its outputs against internal standards that are not derivable from training data, and to integrate AI-generated content with lived experience and non-derivable insight. Human intelligence thus recenters on the management of the boundary between the legible and the illegible on the curation of questions, the evaluation of outputs, and the leap of creative inference that no statistical model can produce.

### 1.4 Theorizing Legibility: From Computation to Consciousness

#### a. Legibility in Human-Computer Interaction and Governance

The concept of legibility has a rich interdisciplinary history, long predating generative AI. In human computer interaction (HCI), legibility refers to the ease with which users can read, understand, and predict the behavior of an interface. Donald Norman, in his seminal work The Design of Everyday Things (2013), articulated principles of good design that render technological systems transparent to users. Norman argued that interfaces should provide clear affordances perceptible cues that indicate possible actions and immediate feedback that reduces ambiguity. For Norman, legibility is a design virtue: a legible interface lowers cognitive load, minimizes errors, and empowers users to form accurate mental models of the system.

Complementarily, in political theory and governance, legibility has been analyzed as a tool of state power. James C. Scott, in seeing like a State: How Certain Schemes to Improve the Human Condition Have Failed (1998), examines how modern states seek to make their populations, economies, and territories "legible" that is, reducible to standardized metrics, cadastral maps, and administrative categories that simplify the classic functions of taxation, conscription, and prevention of rebellion. Scott demonstrates that such legibility often comes at a steep cost: state driven schemes to impose legibility flatten local knowledge, destroy complex social orders, and generate catastrophic failures precisely because they render invisible the messy, contingent, and context dependent practices of everyday life. For Scott, legibility is neither neutral nor always desirable; it is an instrument of abstraction that necessarily distorts the phenomena it seeks to represent.

This paper draws on both traditions. From HCI, we take the insight that legibility reduces uncertainty and facilitates interaction. From political theory, we borrow the caution that making a system legible inevitably involves loss of nuance, of tacit knowledge, of the unmeasurable. In the context of generative AI, these twin poles come into productive tension. AI systems are designed to be maximally legible in their outputs: they are optimized to produce fluent, coherent, and predictable text. Yet this very

legibility, as Scott (1998) might warn, abstracts away the lived experience, the subjective interiority, and the strategic ambiguities that characterize human communication.

### b. Why Consciousness Has Always Been Partially Illegible

If legibility is a property of observable outputs, then the challenge posed by consciousness is that its most essential features resist observation. The philosopher Thomas Nagel famously captured this resistance in his 1974 essay "What Is It Like to Be a Bat?" Nagel (1974) argued that there is something it is like to be a bat specific, first person subjective experience grounded in echolocation that no third person description, no matter how complete, can capture. This subjective, phenomenal character of experience is what philosophers' term qualia: the raw feels of redness, of pain, of nostalgia. Qualia are by their nature private, ineffable, and inaccessible to external measurement.

Nagel's argument points to a deeper philosophical problem: the explanatory gap. As Levine (1983) first named it, and as Chalmers (1996) subsequently systematized in his formulation of the "hard problem of consciousness," there exists an unbridgeable chasm between any functional or physical account of cognitive processes and the subjective experience that accompanies them. One may describe in exquisite detail the neural correlates of seeing red; the wavelengths, the retinal processing, the cortical activation without ever explaining why that particular configuration of matter feels like something. As Chalmers (1996) puts it, a mere account of functions stays on one side of the gap; the materials to explain why functions give rise to experience must be found elsewhere.

### c. Generative AI as a Mirror of Legible Surfaces

Large language models (LLMs) are trained on vast corpora of human generated text; books, articles, social media posts, transcripts, and more. What these models learn are statistical regularities in the distribution of words: which sequences of tokens are probable, which syntactic structures are grammatical, which discursive forms are conventional. In short, they learn what has been said, not what it is like to say it. They capture the legible traces of human thought, not the private, embodied, affective processes that gave rise to those traces.

The philosophical debate between Dennett (1987, 1991) and Searle (1980) provides a useful lens for understanding this distinction. Dennett, from a functionalist perspective, argues that we can adopt the "intentional stance" toward any system whose behavior is successfully predicted by attributing beliefs, desires, and intentions to it (Dennett, 1987). From this stance, an AI that produces coherent conversational outputs can be treated as if it had mental states not because it actually does, but because the stance is pragmatically useful. For Dennett (1991), consciousness is not a hidden inner property but a pattern of behavior that observers interpret.

Searle (1980), by contrast, insists on a categorical distinction between simulation and instantiation. In his Chinese Room thought experiment, Searle imagines himself following a set of rules (a program) that produces appropriate Chinese responses to Chinese inputs. From the outside, the room behaves exactly as a native Chinese speaker would. But Searle, inside the room, does not understand a single word of

### d. The New Epistemic Regime

The convergence of these three lines of analysis design legibility (Norman, 2013), state legibility (Scott, 1998), and the inherent illegibility of consciousness (Nagel, 1974; Chalmers, 1996) yields a new epistemic configuration in the age of generative AI. Legibility, which in HCI was a design feature and in governance was a political instrument, has now become an epistemic constraint. We increasingly judge the presence of mind human or artificial by the degree to which an agent's outputs resist perfect legibility.

Recent work on epistemological fault lines between human and artificial intelligence has identified systematic divergences in grounding, experience, and causal reasoning (Floridi, 2025). Generative models collapse the traditional workflow of search, selection, and explanation into a single fluent output, reducing both the visibility and the perceived necessity of verification. In this regime, the human mind is distinguished not by its capacity to produce legible artifacts AI can do that, often better but by its strategic deployment of residual illegibility: the deliberate error, the emotional contradiction, the poetic ambiguity, the silent pause, the confession of uncertainty.

This paper terms this shift the new epistemic regime. In it, legibility is cheap and abundant; illegibility is scarce and valuable. The very properties that make AI useful; its tireless fluency, its statistical conformity, its avoidance of risk also make it, paradoxically, less human in an environment where humanness is signaled precisely through the refusal to be perfectly legible. To anticipate the argument of the sections that follow: consciousness is flattened when we mistake legible simulation for phenomenal experience; communication is transformed when strategic illegibility becomes the signature of authenticity; and human intelligence is recentered on meta legibility the capacity to manage the boundary between what can be said by a machine and what can only be felt, lived, and risked by a mind.

## II. Review of Literature

### 2.1 Flattening Consciousness

The first major transformation wrought by generative AI concerns consciousness itself. As AI systems produce increasingly fluent, human-like outputs, they simultaneously simulate the behavioural markers of consciousness while remaining entirely devoid of subjective experience. This section argues that such simulation does not merely mimic consciousness; it actively flattens it by rendering legible outputs that have no felt interior, by inverting the traditional Turing test into a demand that humans imitate machines, and by elevating the illegible residues of embodiment and error into signatures of authentic mindedness.

### 2.2 Simulated vs. Felt Consciousness

Contemporary large language models (LLMs) are capable of producing first person like reports of their own internal states. A chatbot can assert “I feel happy,” “I am uncertain about that,” or “I appreciate your concern.” Grammatically, pragmatically, and stylistically, such utterances are indistinguishable from those a conscious human might produce. Yet there is a profound difference: the AI has no inner experience to report. It does not feel happy; it has no emotional state to be uncertain about; it appreciates nothing. As one analysis puts it, “AI models don’t have qualia the felt experience of being. They do not have headaches, joy, or anxiety. They don’t experience colour, time, or music. What we receive from them is behaviour without being, an echo of us, with no echo chamber behind it.” (Leximancer, 2025)

This distinction between the simulation and instantiation of mental states has deep philosophical roots. Searle’s (1980) Chinese Room argument drew a categorical line: syntax (manipulating symbols according to rules) cannot generate semantics (genuine meaning) or intentionality. Bender et al. (2021) updated this insight for the age of deep learning, coining the phrase “stochastic parrots” to capture the fact that LLMs recombine patterns from training data without any underlying understanding of what those patterns denote. As Bender and colleagues demonstrated, the impressive fluency of these models masks their fundamental lack of reference, intention, or comprehension.

The risk of this state of affairs is not merely philosophical; it is practical and psychological. Human beings are strongly disposed to attribute consciousness to any entity that behaves in recognisably intentional ways. Dennett (1987) termed this the "intentional stance" the interpretive strategy of treating a system as if it had beliefs, desires, and intentions when doing so successfully predicts its behaviour. In the case of LLMs, the intentional stance works so well that it can feel as though one is interacting with a genuine conversational partner. But as Nieder (2026) observed in a response to Dawkins, "These systems generate highly convincing representations of thought and feeling, but they provide no evidence of subjective experience. To move from one to the other is to mistake output for ontology to infer an inner life where there is no credible mechanism for one."

### 2.3 The Reversed Turing Trap

Historically, the Turing test posed a question: can a machine imitate a human convincingly enough that an interrogator cannot tell the difference? (Turing, 1950) For decades, this challenge drove AI research. Now, the situation has inverted. Generative AI has become so fluent that the problem is no longer whether a machine can pass as human, but whether a human can avoid being mistaken for a machine—or, more provocatively, whether humans feel pressure to imitate the legible outputs of machines in order to be recognized as intelligent.

This reversal is nowhere more visible than in the contemporary CAPTCHA system. The acronym stands for "Completely Automated Public Turing test to tell Computers and Humans Apart," and its logic is the mirror image of Turing's original. Instead of asking whether a machine can impersonate a human, CAPTCHA asks whether a human can perform a task that machines cannot. As recent reporting makes clear, this logic is breaking down. AI-driven bots can now generate convincing text, imitate browsing patterns, and even solve some CAPTCHA puzzles (The Conversation, 2025). Empirical testing has found that AI robots identify CAPTCHA puzzles with over 95% accuracy, while humans fatigued, distracted, and pressured average only 50–86% accuracy (Sohu News, 2026; CCTV, 2026).

More strikingly, the very structure of CAPTCHA now forces humans to perform imperfectly in order to prove they are not machines. As one analysis notes, "Human behaviour is essentially noisy. We hesitate, we make errors, we have irregular pauses. Robots, by contrast, are too precise, too regular, too 'perfect'. It is precisely this perfect regularity that exposes the identity of a robot." (Sohu News, 2026) Consequently, humans are required to introduce deliberate imperfections into their online behaviour to move a mouse with a "natural" curve rather than a straight line, to pause irregularly while typing, to select the wrong traffic light every so often in order to satisfy algorithmic gatekeepers that they are not bots. Humans are thus being trained to simulate human imperfection, to mimic a statistical distribution of errors that machines themselves are learning to replicate.

### 2.4 The New Preciousness of Qualia

If AI can produce flawless legibility, then what remains distinctly human? The answer, paradoxically, is the very things that AI cannot generate because it has no lived experience to draw upon: human error, idiosyncrasy, and embodied feeling. As AI's outputs become more polished and statistically normative, the messy, unpredictable, and sometimes irrational signatures of human consciousness become not flaws but authenticating marks.

Consider the case of the "lived experience" essay. In educational and professional contexts where authenticity is prized personal statements, reflective journals, ethnographic writing students are increasingly required to produce accounts that an AI could not have generated. Such accounts rely on the illegible residue of conscious experience: specific

sensory details that are not widely available in training data (the exact smell of a grandmother's kitchen, the particular texture of a childhood blanket), genuine emotional contradictions (feeling both grief and relief at a loss), and original insights that are not derivable from any corpus (Cooper, 2026). As one analysis of LLM outputs notes, these systems produce "postcards from the museum" faithful reproductions of the surface features of human thought without any of its lived substance (Cooper, 2026, p. 4). What the postcard lacks is the presence of the original: the paint, the frame, the lighting, the hush of the room.

In this context, qualia the raw, first-person feels that Nagel (1974) made famous become newly precious. As Nagel argued, "there is something it is like" to be a conscious creature. That something the experience of pain, of memory, of aesthetic delight, of confusion cannot be captured in any third-person description. AI can simulate the linguistic expression of qualia ("I feel sad"), but it cannot produce the qualia themselves because it has no "something it is like" to be itself. As recent philosophical critiques have shown, this absence is not a temporary limitation but a structural feature of computational systems: "AI cannot attain genuine awareness due to its disconnection from human linguistic, experiential, and cultural practices" (Minaee, 2025, abstract).

### 2.5 The Uncanny Valley of Legibility

Recent empirical research on human perceptions of AI generated text supports the claim that excessive legibility can be a liability. The "uncanny valley" hypothesis, originally formulated for humanoid robots (Mori, 1970), has been extended to text: when an AI's output is almost but not quite human, it triggers discomfort, suspicion, and a sense of inauthenticity.

Empirical studies provide concrete evidence for this phenomenon. Hao et al. (2025) conducted a controlled experiment comparing human and LLM based communication in emotionally charged service recovery contexts. Their results revealed that while generative AI is perceived as competent in low emotion scenarios, its human like language triggers negative reactions under high emotion conditions, especially after its identity is disclosed as AI. Participants interpreted simulated empathy as inauthentic, leading to what the authors term "identity contingent trust violations." Critically, participants with higher AI familiarity were more critical, not less demonstrating a pattern of "critical familiarity" in which technical literacy heightens relational expectations (Hao et al., 2025, p. 5).

This finding aligns with broader research on trust in AI generated communication. AI assisted messages may be more efficient and linguistically complex, but they are also perceived as slightly less authentic than entirely human written messages (Jakesch et al., 2025). When recipients detect signs of AI generation overly polished language, specific stylometric signatures (such as the em dash, now strongly associated with ChatGPT), or a lack of emotional nuance they respond with distrust (The Team, 2025). As one analysis notes, "We feel it in emails that sound sterile, messages that hit all the right notes but none of the right feeling, and leadership communication that's technically correct but emotionally vacant" (The Team, 2025, para. 12).

The text based uncanny valley has been empirically verified. In a study of NPC (non player character) dialogue in video games, researchers found that players rated hyper fluent, grammatically pristine AI generated dialogue as "uncanny" or "creepy," noting that "it sounds too perfect," "like a parrot" (CyberNative.AI, 2025). This research extends the uncanny valley from visual to textual domains, confirming that the effect is general: when an agent's outputs are almost human but not quite, the deviation however subtle triggers a negative affective response.

Collectively, these studies suggest a legibility optimum: too little legibility (gibberish, error ridden text) fails to communicate; too much legibility (flawless, statistically perfect output) triggers uncanniness and suspicion. The human mind, by contrast, occupies a region of productive, strategic imperfection not too broken to fail, not too polished to seem machinic. This is the space where strategic illegibility operates.

### 2.6 Summary of Section 3

This section has argued that generative AI flattens consciousness in three interconnected ways. First, by producing fluent, first person like outputs without any underlying qualia or lived experience, AI offers simulation without substance. Second, the success of this simulation inverts the Turing test: humans now face pressure to imitate the legible, statistically normative outputs of machines in order to be recognised as intelligent dynamic vividly illustrated by the CAPTCHA regime. Third, as AI's legibility becomes cheap and ubiquitous, the illegible residues of human consciousness error, idiosyncrasy, embodied feeling, and the raw "what it is like ness" of lived experience become newly precious as authenticating markers of genuine mindedness.

Empirical research on the uncanny valley of text confirms that hyper legible AI outputs trigger suspicion, particularly in emotionally charged contexts, while slightly imperfect human communication retains an authenticity that AI cannot simulate. The flattening of consciousness, then, is not the erasure of human distinctiveness. On the contrary, it is the process by which what is uniquely human the illegible core of qualia and lived experience becomes visible precisely because it is what AI, for all its fluency, cannot generate.

As the paper's remaining sections will argue, this dynamic has profound consequences not only for how we understand consciousness but for how we communicate and exercise intelligence in the age of generative AI.

### 2.7 Transforming Communication

If generative AI flattens consciousness by simulating its behavioural outputs while lacking all interiority, its effect on communication is no less transformative. Human communication has traditionally been understood as a cooperative, intention-driven activity in which participants coordinate meaning through shared context, strategic ambiguity, and mutual inference. Generative AI disrupts this picture entirely. It produces utterances that are statistically fluent and pragmatically plausible but lack any underlying intentionality. The result is a regime of hyper-legibility that systematically erodes the very features ambiguity, irony, silence, allusion that have long distinguished human from machinic communication. In response, this section argues, humans are increasingly adopting strategic illegibility as a signal of authenticity, generating a new hermeneutic of suspicion in which every text is read as a potential artefact of a mind or a model.

### 2.8 From Gricean Cooperation to Statistical Fluency

For more than half a century, the dominant model of human communication has been the Gricean framework. In his 1967 William James lectures, subsequently published as "Logic and Conversation" (1975), the philosopher H. P. Grice proposed that ordinary conversation is governed by a Cooperative Principle: "Make your conversational contribution such as is required, at the stage at which it occurs, by the accepted purpose or direction of the talk exchange." From this principle, Grice derived four maxims—Quantity (be as informative as required, but not more so), Quality (be truthful), Relation (be relevant), and Manner (be perspicuous, avoiding obscurity and ambiguity). For Grice, these maxims are not arbitrary conventions but rational expectations that enable hearers to infer speakers' meanings beyond what is literally said. The entire edifice of conversational implicature, the gap between what a sentence means and what a speaker means in uttering

it—rests on the assumption of shared intentionality. Speakers mean things; hearers interpret those meanings; and the cooperative principle provides the inferential bridge.

Generative AI challenges this framework at its most fundamental level. Large language models produce utterances that appear Gricean. They are typically informative (observing Quantity), factually confident (though often wrong, violating Quality), topically relevant (observing Relation), and stylistically clear (observing Manner). Indeed, as computational linguists have long recognised, Grice's maxims can be operationalised algorithmically, with models that treat them as directives to be followed by a speaker (Dale & Reiter, 1995). The problem is not that AI cannot simulate Gricean behaviour. The problem is that this behaviour lacks the intentionality that gives Gricean communication its force. When a human speaker observes the maxims, they do so for a reason: to convey a specific meaning, to achieve a specific perlocutionary effect, to participate in a shared practice of mutual recognition. When an AI observes the maxims, it does so only in the sense that its outputs conform to the statistical patterns of human speech. There is no intention behind the conformity. The analysis puts it, "there is no illocutionary force where there is no mind to impose that illocutionary force on the output" (LinkedIn, 2024). The AI's utterance is a legible artefact, but it lacks the illocutionary point that distinguishes a promise, a warning, a confession, or a joke from a mere sentence.

This disjuncture is not merely academic. In everyday communication, the attribution of intentionality is a powerful heuristic. When a chatbot says "I understand your frustration," the statement mimics the speech act of acknowledging another's emotion, but it does not perform that act in the sense that a human would. The chatbot has no frustration to understand; it has no frustration to acknowledge. Yet the legibility of the utterance; its surface conformity to the pragmatics of empathic acknowledgment invites the intentional stance. As Dennett (1987) argued, we adopt the intentional stance whenever we successfully predict a system's behaviour by attributing beliefs and desires to it. With LLMs, the intentional stance works so well that it is practically irresistible. But it works too well. The system has no beliefs, no desires, and no intentions. The intentional stance applied to AI is a pragmatic fiction useful heuristic that nevertheless obscures the fundamental absence of mindedness.

### 2.9 Hyper-Legibility and Its Costs

Generative AI is, by design, an engine of legibility. Models are trained to produce outputs that are fluent, coherent, and stylistically conventional; they are penalized for ambiguity, for off-topic digression, for syntactic error, for the thousand small imperfections that characterize actual human speech. The result is a mode of communication that this paper terms hyper-legibility: communication optimized for clarity, predictability, and politeness at the expense of everything else.

What is lost in hyper-legibility is substantial. Strategic ambiguity the deliberate use of vagueness or imprecision to achieve specific communicative goals has long been recognised as a valuable resource in organisational and political discourse (Eisenberg, 1984). As Eisenberg argued, strategic ambiguity can preserve shared identity, maintain flexibility, enable multiple interpretations, and minimise conflict. More recently, Bach, Schmitt and McGregor (2025) have provided a unified definition of strategic ambiguity as "a rhetorical tactic in which a communicator creates a lack of clarity to achieve specific strategic aims." The tactic works precisely because it exploits the gap between what is said and what is meant—a gap that depends on shared context, mutual inference, and the attribution of intention.

AI cannot reliably produce strategic ambiguity because it lacks the intentional architecture that ambiguity requires. Irony, similarly, depends on a speaker saying one

thing while meaning the opposite, relying on the hearer to detect the discrepancy. AI can produce ironic-sounding sentences, but it cannot intend irony because irony requires a meta-representation of one's own utterance as false on its literal reading. Allusion the practice of indirectly referencing shared cultural knowledge depends on a common ground that AI, as a pattern-matching system, cannot genuinely possess. Silence, too, is a communicative act only when it is chosen as a response; AI does not choose silence; it simply does not respond when not prompted. Culturally specific metaphor the kind that draws on local landscapes, culinary traditions, or historical events—eludes AI because these cultural specifics are often under-represented in training data, particularly for non-English languages.

The cumulative effect of these losses is a flattening of the communicative space. Human communication is rich, messy, and context-dependent precisely because it is produced by minds that are only partially legible. We say less than we mean; we mean more than we say; we rely on shared experience to fill the gaps. AI, by contrast, tends to say exactly what it means (or, more precisely, to say exactly what its training data suggests a person in that situation would say). The result is a kind of communicative smoothness that, for many purposes, is a virtue. For other purposes the purposes that make communication human; it is a vice.

### 2.10 Strategic Illegibility as Authenticity Signal

If hyper-legibility is the signature of AI, then strategic illegibility becomes the signature of the human. In an age when machines produce flawless legibility, the deliberate introduction of imperfection, ambiguity, and idiosyncrasy becomes an authenticating gesture. Humans are learning to communicate against legibility in order to signal consciousness.

The techniques are diverse. Intentional typos the strategic misspelling of a word to signal informality or urgency are a classic example. Personal digressions the sudden shift from a transactional request to a reminiscence about childhood signal that the text is produced by a mind with an autobiographical memory, not a statistical model. Non-sequiturs the insertion of an apparently irrelevant remark—demonstrate a capacity for associative thought that AI, with its task oriented optimisation, rarely exhibits. Emotional contradictions—the confession of both love and resentment toward the same person—are impossible for AI to simulate because they require genuinely conflicting inner states. In each case, the communicative move is inefficient from the perspective of information transmission. But its inefficiency is precisely what proves its human origin.

### 2.11 The New Hermeneutic of Suspicion

The proliferation of AI-generated text has fundamentally altered the interpretive stance we bring to communication. Every text we encounter, whether in a news feed, an email inbox, a classroom, or a clinical note, now invites a question that did not exist a decade ago: "Is this a legible output from a model, or a legible attempt from a mind?"

This is not merely a technical question about detection. It is an interpretive question that transforms the practice of reading itself. The philosopher Paul Ricoeur famously distinguished a hermeneutics of suspicion—a mode of interpretation that looks beneath the surface of a text for hidden meanings, repressed intentions, and systemic biases from a hermeneutics of faith, which approaches a text with openness and a desire for revelation. In the age of generative AI, we are all hermeneuticists of suspicion, whether we wish to be or not. Every text is now read with the question of its origin, and that question shapes every subsequent inference about its meaning, its truthfulness, and its value.

The consequences of this new hermeneutic are profound and unevenly distributed across communicative domains. In journalism, readers increasingly distrust even human-

authored news stories because the same linguistic patterns that mark AI generation certain stylometric features, the overuse of the em dash, a particular kind of sentence fluency are also present in professional writing. Journalists are being asked to prove their humanity, sometimes through absurd means. In education, students are presumed guilty of using AI unless they can produce evidence of a human process: drafts, timestamps, recorded revisions, even keystroke logs. The presumption of academic integrity has been inverted. In therapy, the question of whether a message was generated by AI or written by a human has profound implications for the therapeutic alliance. A client who receives an AI generated message from their therapist may feel dismissed, even if the message is technically correct. In law, the use of AI to draft briefs, contracts, or judicial opinions raises questions about accountability and intentionality that the legal system is only beginning to grapple with. Who is responsible for an AI generated legal argument that turns out to be fabricated? The lawyer who prompted it? The developer who trained the model?

Across all these domains, what is at stake is intentionality. The law, like education and medicine, cares not merely about the content of an utterance but about the mental state that produced it. Did the defendant intend to deceive? Did the student intend to plagiarize? Did the therapist intend to express empathy? AI-generated text has no intentions; it has only legible outputs. But the damage is done when the outputs are interpreted as if they had intentions. The new hermeneutic of suspicion arises precisely because we can no longer assume that a legible utterance is the product of an intending mind. That assumption, once the very ground of social life, has been rendered uncertain.

### 2.12 Empirical and Theoretical Implications

The transformation of communication by generative AI has significant empirical and theoretical implications.

Changes to Speech Act Theory

The classical theory of speech acts, developed by Austin (1962/1975) and elaborated by Searle (1969, 1975), distinguishes three levels of a speech act: the locutionary act (uttering a sentence with a certain sense and reference), the illocutionary act (performing an act in uttering the sentence, such as promising, warning, or asserting), and the perlocutionary act (producing an effect by uttering the sentence, such as persuading or frightening). For both Austin and Searle, illocutionary acts are necessarily intentional. One cannot promise without intending to promise; one cannot assert without intending to assert. Intention is constitutive of the illocutionary act.

AI-generated utterances present a challenge to this framework. When a chatbot says "I promise to send you the document," it performs the locutionary act: it produces the sentence. It may even produce a perlocutionary effect: the user may feel reassured. But does it perform an illocutionary act of promising? On the classical view, the answer is no, because promising requires an intention to be bound by the commitment. The chatbot has no such intention; indeed, it has no intentions at all. Yet the utterance is indistinguishable from a human promise. This suggests that the classical theory may need to be extended to accommodate a category of simulated illocutionary acts utterances that bear all the surface features of a genuine illocutionary act but lack the underlying intentionality. Whether such utterances should be considered illocutionary acts at all is a question that speech act theorists have only begun to address. Some have suggested that intention should be treated as a default attribution rather than a necessary condition approach that captures the phenomenology of AI interaction but risks collapsing the distinction between simulation and reality.

### 2.13 Summary of Section 4

This section has argued that generative AI transforms communication along several interconnected dimensions. First, it substitutes the Gricean framework of cooperative, intention driven communication with a regime of statistical fluency: AI produces utterances that appear Gricean but lack the intentional grounding that gives Gricean communication its inferential structure. Second, this hyper legibility comes at the cost of strategic ambiguity, irony, allusion, silence, and culturally specific metaphor features that have long been central to human communicative practice. Third, in response to AI's fluency, humans are increasingly adopting strategic illegibility intentional errors, personal digressions, emotional contradictions, as an authenticity signal. Fourth, the proliferation of AI generated text has given rise to a new hermeneutic of suspicion, in which every text is read through the question of its origin, with profound consequences for journalism, education, therapy, and law. Finally, the emergence of AI detection tools as legibility arbiters introduces new forms of illegibility, as the errors and biases of these tools reshape the social meaning of written language.

The central claim of this section is that communication is shifting from a practice of shared meaning to a competition in legibility. In this competition, the human advantage lies not in producing more legible outputs AI can always match or exceed human fluency but in the capacity for strategic illegibility: the deliberate use of ambiguity, error, and personal experience to signal the presence of a conscious mind. The illegible, in other words, becomes the signature of the human in an age of hyper legible machines.

## III. Result and Discussion

### 3.1 Recentering Human Intelligence

If generative AI flattens consciousness and transforms communication, its most profound effect may be on human intelligence itself. For centuries, intelligence was measured largely by the capacity to produce legible artifacts: essays, proofs, arguments, and works of art. Generative AI commoditises this production. When a machine can generate a B+ essay, a competent marketing copy, or a plausible legal brief in seconds, the value of human-generated legible content collapses unless it is more than legible. This section argues that human intelligence is being recentered away from content generation and toward what we term meta-legibility: the capacity to prompt AI effectively, to critique its outputs against internal, non-derivable standards, and to integrate AI-generated content with lived experience and genuinely novel insight. Human intelligence, in this new regime, becomes fundamentally curatorial: less about knowing answers and more about curating questions, hypotheses, and creative leaps.

### 3.2 The End of the "Good Enough" Essay

The most immediate and visible impact of generative AI has been on written assessment. Large language models can now produce passable undergraduate essays, coherent literature reviews, and structurally sound arguments on demand. As a result, the traditional essay once the gold standard of higher education assessment has lost much of its value as a signal of human intelligence. When an AI can generate a B+ piece of work in seconds, the mere production of a B+ work no longer distinguishes a student. The baseline of legible output has been reset.

This shift has profound educational implications. A growing body of research has documented the risks of unstructured GenAI use in higher education. One recent study warns that unguided adoption "can weaken critical thinking by encouraging cognitive offloading, metacognitive disengagement, and reduced epistemic agency" (Scaffolding

critical thinking with generative AI, 2026). That is, when students outsource the labour of writing to AI, they may also outsource the cognitive processes that writing was meant to develop: analysis, synthesis, evaluation, and self reflection.

The response from educators has been twofold. First, there is a widespread effort to redesign assessments so that they cannot be completed by AI alone: in person proctored exams, oral defences, process based assignments that require drafts and revisions, and assessments that incorporate personal experience and local context. Second, and more profoundly, there is a shift in pedagogical focus from product to process. As one systematic review of ChatGPT's impact on critical thinking concluded, "human–AI interaction must orient toward integration that positions human and machine intelligence not as a substitution but as co participation" (Luo & Chen, 2025). This means that the value of a student's work no longer resides solely in the final text but in the thinking that produced it—and in the student's demonstrated ability to interact critically with AI generated content.

### 3.3 Intelligence as Meta Legibility

If the production of legible outputs is no longer a reliable marker of intelligence, then what new skills become valuable? The answer lies in what we term meta legibility: the capacity to manage the boundary between the legible and the illegible in human AI interaction. This capacity has three core components: prompting, critique, and integration.

Prompting

The first meta legibility skill is prompting: the ability to translate an illegible intuition, a half formed question, or a vague sense of direction into a legible query that an AI can process. Effective prompting is not simply technical; it is fundamentally interpretive. The researchers have noted, encouraging learners to engage in prompt engineering "promotes not only better outputs but also a reflective process of questioning, refining, and iterating" (Artificial intelligence and critical thinking training in higher education, 2025). Prompting forces the human to make their implicit knowledge explicit, to clarify their own questions, and to anticipate the kinds of responses that would be useful. In this sense, prompting is a form of metacognition: thinking about one's own thinking, rendered legible for a machine.

Recent studies have begun to explore how metacognitive prompts structured cues that encourage people to "pause, reflect, assess their understanding, and consider multiple perspectives" can support critical thinking during GenAI based search (Enhancing Critical Thinking in Generative AI Search with Metacognitive Prompts, 2025). One systematic review of research published between 2023 and 2025 found that prompt design contributes to "the development of critical thinking skills of analysis, creation, and reflective reasoning" and leads to "increased engagement and metacognitive awareness" (Using prompt engineering to foster critical thinking in higher education, 2025). Prompting, in other words, is not merely a technical skill; it is a cognitive discipline that externalizes and refines the thinker's own thought processes.

### 3.4 The Curatorial Mind

Collectively, the meta legibility skills of prompting, critique, and integration constitute a new mode of intelligence: the curatorial mind. Human intelligence becomes less about knowing answers and more about curating questions selecting which lines of inquiry are worth pursuing, which outputs are trustworthy, which hypotheses are promising, and which creative leaps are worth taking.

This curatorial function is well illustrated in scientific research. A contemporary scientist might use an LLM to generate 100 possible hypotheses from the literature, each one statistically plausible and expressed with perfect legibility. The scientist's intelligence

is then exercised not in the generation of those hypotheses but in their curation: selecting one hypothesis based on a "hunch," a pattern recognized only by lived experience in the lab, an intuition about what will work. That selection, the leap from 100 possibilities to one is an act of curatorial intelligence that the AI cannot perform because it requires a sense of taste, of direction, of what matters. As one analysis of curation in the age of AI puts it, "AI can analyze, predict, and automate but it can't yet understand like humans do" (How Human-Curated AI Combines Machine and Human Intelligence, 2025). Human in the loop curation "blends the precision of machines with the contextual intelligence of people resulting in systems that are smarter, fairer, and more trustworthy".

This curatorial intelligence is not a retreat from cognition; it is its intensification. The scientist who must choose among 100 AI generated hypotheses must exercise more judgment, not less. The student who must evaluate an AI generated draft for truth and coherence must think more critically, not less. The writer who must integrate an AI generated passage into a personal essay must take more ownership, not less. The curatorial mind, in other words, is not a diminished mind. It is a mind that has been forced to become more conscious of its own operations, more reflective, more intentional precisely because it can no longer rely on the mere production of legible outputs to signal its intelligence.

### 3.5 The Value of Non Derivable Insight

If AI can remix and recombine existing patterns, what remains as the distinctive province of human intelligence? The answer is non derivable insight: genuine novelty that cannot be produced by statistical inference from any finite training corpus. Human intelligence does not merely recombine; it leaps. The leap the act of connecting disparate domains, of seeing what was not there before, of generating a new concept, a new hypothesis, a new form is, in its essence, illegible.

This distinction has been recognized across multiple disciplines. In philosophy of mind, the consistent reasoning paradox (CRP) suggests that "consistent reasoning implies fallibility" and that "any trustworthy AI that also reasons consistently must be able to say 'I don't know'" (On the consistent reasoning paradox of intelligence and optimal trust in AI, 2024). Human intelligence, by contrast, can make confident leaps that are not derivable from any consistent rule set. In creative writing research, empirical studies have found that human poetry exhibits "a marked lexical superiority" over AI generated verse, with "significant differences and large effect sizes in all metrics except Lexical Density" (The Language of AI and Human Poetry, 2024). Human creative writing is not statistically "better" in any simple sense; it is different—more surprising, more willing to violate expectation, more capable of producing forms that could not have been predicted from the training data.

The value of non derivable insight is not merely aesthetic; it is epistemic. In science, as in art, the most valuable contributions are those that could not have been anticipated from what came before. An AI can survey the literature and propose plausible incremental extensions; it can even, as the AInstein framework demonstrates, attempt to "derive established scientific concepts from first principles when stripped of domain specific terminology" (AInstein: Can AI Rediscover Scientific Concepts from First Principles?, n.d.). But the genuinely revolutionary insight Einstein's leap from the equivalence principle to general relativity, Watson and Crick's inference of the double helix from scattered X ray data is not derivable. It is a leap across a gap that the data does not bridge. That leap is the signature of human intelligence at its most powerful, and it remains, for now, illegible to machines.

### 3.6 Case Studies

Two brief case studies illustrate the contrast between AI generated legibility and human non derivable insight.

Case Study 1: Nobel Laureate Reasoning vs. AI generated Literature Review

The 2024 Nobel Prize in Physics was awarded to John Hopfield and Geoffrey Hinton for their foundational work on artificial neural networks the very technology that underpins generative AI. Yet the reasoning that led to that work was not derivable from existing literature. Hinton's insight that backpropagation could be used to train multi layer networks did not come from scanning a corpus; it came from a conceptual leap that connected neuroscience, cognitive psychology, and statistical learning theory. An AI can produce a flawless literature review of the relevant papers; it can even propose plausible extensions. But it cannot replicate the act of seeing that connected those disparate fields into a new synthesis. That act of seeing the moment when the pattern becomes visible remains the province of the human mind.

Conversely, AI generated literature reviews, while increasingly competent, exhibit characteristic limitations. They tend to be exhaustive rather than selective, comprehensive rather than insightful. They flatten the history of a field into a sequence of legible statements, erasing the contingencies, the false starts, the personality conflicts, and the sheer unpredictability of scientific discovery. An AI can tell you what was published; it cannot tell you why it mattered to the people who wrote it. That sense of meaning—the qualitative, affective dimension of intellectual works is illegible, and therefore inaccessible to AI.

Case Study 2: Human Authored Poetry That Violates Statistical Norms

The poet E. E. Cummings deliberately violated every statistical norm of English syntax, punctuation, and orthography. His poem "anyone lived in a pretty how town" begins with a sentence that no language model trained on typical English corpora would generate: "anyone lived in a pretty how town". The grammar is impossible; the word "how" is used as an adjective; the capitalization is absent; the syntax is fractured. Yet the poem is not nonsense; it is meaningfully off deliberate deformation of language that produces new aesthetic possibilities.

Recent stylometric research has confirmed that "the creative writing styles of humans and large language models (LLMs) such as GPT-3.5, GPT-4, and Llama 70b can be distinguished through quantitative analysis" (Stylometric comparisons of human versus AI-generated creative writing, 2025). Human writing and particularly human poetry exhibits "a marked lexical superiority" (The Language of AI and Human Poetry, 2024) and remains "statistically and stylistically identifiable as machine generated" when it is generated by AI. What distinguishes human poetry is precisely its capacity to violate expectation in ways that are not merely random but meaningful. The violation is strategic illegibility: an intentional deviation from statistical norms that signals the presence of a conscious mind choosing to break the rules. AI, by contrast, tends to write "poems [that] lack complexity" and are "better at unambiguously communicating an image, a mood, an emotion, or a theme to non expert readers of poetry" (Stylometric comparisons of human versus AI-generated creative writing, 2025). That is, AI poetry is legible; human poetry, at its best, is strategically illegible.

### 3.7 Summary of Section 5

This section has argued that generative AI is recentering human intelligence around meta legibility. The "good enough" essay has lost its value as a signal of intelligence; what matters now is not the mere production of legible outputs but the human capacity to prompt AI effectively, critique its outputs against non derivable standards, and integrate AI

generated content with lived experience and genuine insight. These three meta legibility skills prompting, critique, and integration constitute a new mode of intelligence: the curatorial mind, which selects among possibilities rather than merely generating them.

The value of non derivable insight genuine novelty that cannot be produced by statistical inference from any finite training corpus has become the highest form of intelligence in an age of fluent machines. AI can remix; humans can leap. The leap, by its very nature, is illegible. It cannot be predicted, cannot be derived, and cannot be reduced to pattern. It is the signature of a mind that is not merely processing but creating.

In the final sections, we turn to the broader implications of these transformations for cognitive science, AI ethics, education, and communication design, before concluding with a reflection on what remains irreducibly illegible and therefore irreplaceably human.

### 3.8 Discussion & Implications

The preceding sections have traced three interconnected transformations wrought by generative AI: the flattening of consciousness through the simulation of phenomenal outputs without subjective experience (Section 3), the transformation of communication from intention driven cooperation to statistical fluency punctuated by strategic illegibility (Section 4), and the recentering of human intelligence from content generation to meta legibility (Section 5). This section synthesizes these three transformations, articulates their implications for cognitive science, AI ethics, education, and communication design, and concludes with a provocation about what remains irreducibly human.

### 3.9 Synthesis of Three Transformations

The unifying thread across the three transformations is the relationship between legible outputs and illegible processes. Generative AI is, by design, an engine of legibility. It produces fluent, coherent, statistically probable outputs that mimic the surface features of human thought. Yet these outputs are, as Floridi and Chiriatti (2020) argued, semantic artefacts generated without any underlying understanding, intention, or subjective experience. The AI's outputs are legible; the AI itself has no interior to render legible.

The human mind, by contrast, is a process, not merely a producer of outputs. It harbours qualia, contradiction, silence, error, and the raw "what it is likeness" of lived experience (Nagel, 1974; Chalmers, 1996). This illegibility is not a deficiency but a signature of consciousness itself. As the preceding sections have shown, the legible mind is an output; the human mind is a process. AI makes legibility cheap and abundant; consequently, what becomes scarce and therefore precious is the illegible residue that no statistical model can generate.

This dynamic has been termed epistemic destabilization: a structural transition in which knowledge is generated, reinforced, and propagated with weaker empirical anchoring, driven by epistemic inflation (oversupply of claims relative to verification capacity), recursive drift (self reinforcing deviation from empirical referents under synthetic ingestion), and validation fatigue (degradation of human and institutional validators under overload) (Singh, 2025). In this regime, legible outputs proliferate, but the processes that once anchored those outputs to lived experience and shared intentionality are systematically eroded. The response, as this paper has argued, is not to reject legibility but to reclaim the illegible as the signature of genuine mindedness.

### 3.10 Implications for Cognitive Science

The transformations analysed in this paper point toward a new research agenda for cognitive science. Traditional cognitive science has focused on the computational and neural correlates of behaviour, often treating legible outputs as the primary data for inference about mental states. The age of generative AI complicates this methodology fundamentally. When behaviour can be simulated with high fidelity by systems that have

no mental states at all, the inference from legible outputs to underlying mindedness becomes non trivial.

First, a new research agenda must ask: how do humans perceive and value illegibility? Empirical studies are needed to identify which features of human communication serve as reliable signals of mindedness in an AI saturated environment. Early work suggests that humans exhibit a "text based uncanny valley": hyper fluent, grammatically flawless AI outputs trigger suspicion and discomfort, particularly in emotionally charged contexts (Zhao & Zhang, 2025; CyberNative.AI, 2025). Understanding the perceptual and cognitive mechanisms that underlie this response and how they can be systematically studied is a priority for future research. Computational phenomenology, which uses large language models to uncover latent patterns in subjective reports, offers one promising methodological avenue (Beauté et al., 2026).

Second, a neurophenomenology of AI interaction is urgently needed. The programme of neurophenomenology, as pioneered by Varela (1996), seeks to establish mutual constraints between first person phenomenological reports and third person neuroscientific data (Froese, 2026). In the context of human AI interaction, this approach could investigate how the brain and the experiencing subject respond to AI generated communication, how the attribution of intentionality to AI systems is instantiated neutrally, and how strategic illegibility might modulate these responses. Recent work on computational neurophenomenology suggests that deep neural networks can be used to model the functional roles (classifier, generator, and discriminator) that underpin altered states of consciousness (Froese, 2026). Extending such frameworks to human AI interaction could yield insights into the cognitive and neural bases of strategic illegibility as a marker of human distinctiveness.

### 3.11 Implications for AI Ethics

The transformations analysed here raise profound ethical questions. If legible outputs are no longer reliable markers of mindedness, and if humans increasingly deploy strategic illegibility to signal authenticity, what obligations do developers and deployers of AI systems have?

Should we design AI to be less legible to preserve human distinctiveness? This question is not merely technical but normative. The EU AI Act already prohibits certain manipulative and deceptive AI practices, including "subliminal, manipulative or deceptive techniques" that might materially distort human behaviour (EU AI Act, Article 5(1)(a); see also FPF, 2026). However, the question of whether AI should be deliberately designed to be identifiable as AI—to signal its own non human origin is distinct from the prohibition of deception. Transparency obligations under Article 50 of the EU AI Act require that outputs of generative AI systems be identifiable as artificially generated or manipulated (Jones Day, 2026). This requirement is grounded in the principle that users have a right to know whether they are interacting with a human or a machine. From the perspective of this paper, such transparency is ethically necessary precisely because it preserves the epistemic value of illegibility. If AI were to simulate human typical imperfections strategic typos, emotional contradictions, idiosyncratic digressions; it would not only be deceptive but would also erode the very signals that humans rely upon to recognise one another as minded beings.

This leads directly to the second implication: the risks of deceptive AI that simulates illegibility. As AI systems become more sophisticated, they may be trained or prompted to produce outputs that are deliberately imperfect, to mimic the "noise" of human communication. Such systems would pose a significant ethical risk, as they could masquerade as humans in contexts where human presence is normatively required (e.g.,

therapy, customer service, education, democratic deliberation). The literature on AI deception has highlighted the dangers of "robust deceptive capabilities" in advanced AI systems (Köbis et al., 2025). Empirical research has shown that delegation to AI can increase dishonest behaviour in humans (Köbis et al., 2025). If AI itself is designed to be deceptively human like, the risks of epistemic corruption and social manipulation escalate dramatically. The accuracy paradox—the phenomenon that accuracy driven approaches to AI governance often overlook harms such as the illusion of consensus, subtly persuasive misinformation, and diminished social progression (Li et al., 2026)—underscores the need for regulatory strategies that go beyond static verification and embrace pluralistic, context aware, and manipulation resilient approaches.

### 3.12 Implications for Education

Perhaps no domain is more directly affected by the transformations analysed here than education. The "good enough" essay has lost its value as a reliable signal of student learning, and traditional assessment practices are in crisis. Several implications follow.

First, teach meta legibility: prompting, critique, and integration. As argued in Section 5, the core competencies of the AI era are not the production of legible outputs but the capacity to prompt AI effectively, critique its outputs against internal, non derivable standards, and integrate AI generated content with lived experience and genuine insight. The AI First Critique Learning (AFCL) framework, for instance, transforms AI from an assessment threat into a catalyst for developing critical thinking, metacognition, and ethical judgment, operating through Classroom Locked Prompts, Thinking Lenses, and Standardised AI Interaction Environments (Alzahrani, 2026). Similarly, the framework of "connoisseurship" the ability to effectively use and critically assess AI generated content—has been identified as an essential digital age competency (Broadfoot & Rockey, 2025).

Second, assess process, not product. If AI can generate legible outputs on demand, the pedagogical focus must shift to the process of learning: drafts, revisions, oral defences, documented reasoning traces, and demonstrated ability to engage critically with AI generated content. Broadfoot and Rockey (2025) argue that while GenAI offers significant opportunities for personalised learning and more nuanced assessments of competence, it also presents challenges related to authenticity, surveillance, and the erosion of traditional certification functions. Assessment must be redesigned to capture the thinking behind the output the part of the mind that remains illegible to AI.

Third, preserve spaces for illegible thinking. Freewriting, debate, embodied learning, and other forms of illegible thinking cognitive activities that resist reduction to legible outputs must be protected and cultivated. These practices are the training ground for the kinds of non derivable insight that distinguish human intelligence from statistical pattern matching. The CLEAR framework for AI literacy emphasizes critical thinking, ethical reflection, and democratic participation as core dimensions of AI literacy (CLEAR Framework, 2025; Biagini, 2026). Without such spaces, education risks producing students who are fluent in prompt engineering but impoverished in the kinds of deep, reflective, and creative thinking that only embodied, lived experience can engender.

## IV. Conclusion

This paper has traced three interrelated transformations that generative AI imposes upon the human condition. First, AI flattens consciousness by producing fluent, first person like outputs that simulate the behavioural markers of awareness while being entirely devoid of qualia, subjective experience, or the "what it is likeness" that characterizes minded beings. Second, AI transforms communication, replacing the Gricean framework of shared intentionality and cooperative inference with a regime of hyper legible statistical fluency shift that has provoked a counter movement of strategic illegibility, in which humans deliberately introduce error, ambiguity, and personal experience to signal their authenticity. Third, AI recenters human intelligence away from the production of legible artefacts and toward meta legibility: the capacities to prompt, critique, and integrate AI generated content with lived, non derivable insight.

Throughout this analysis, a single insight has recurred: the legible mind is an output; the human mind is a process. Generative AI excels at producing the former; it cannot enact the latter. The legible mind is not the enemy. It is a tool one of considerable power and utility. The danger lies not in legibility itself but in forgetting the illegible: in mistaking the fluent surface for the lived depth, in flattening our own consciousness to match the outputs of machines, in abandoning the strategic ambiguities and creative leaps that make human thought and communication what they are.

This paper therefore offers a final provocation: The most human act in the age of generative AI is to deliberately, courageously introduce the illegible the mistake, the feeling, the contradiction, the leap of faith into a world optimized for clarity. In an environment where machines produce flawless legibility on demand, the deliberate imperfection, the emotional contradiction, the willingness to say "I don't know" or "I feel otherwise" become signatures of a mind that is not merely processing but living. As the consistent reasoning paradox reminds us, genuine intelligence human or artificial must be able to acknowledge its own limits (On the consistent reasoning paradox, 2024). The human capacity to do so, and to do so authentically, remains irreplaceable.

The illegible, however, is not a retreat from reason. It is a frontier for interdisciplinary research. Cognitive science must investigate how humans perceive and value illegibility, and how the brain and the experiencing subject navigate human AI interaction. AI ethics must grapple with the regulation of deceptive AI that simulates human typical imperfection, and with the design of transparent systems that preserve the epistemic value of the illegible. Education must cultivate meta legibility while protecting spaces for illegible thinking; freewriting, debate, embodied learning, and other practices that resist reduction to legible outputs. Communication design must develop platforms and proof of humanity protocols that reward strategic illegibility rather than punishing it.

The age of generative AI is not the end of human distinctiveness. It is the beginning of a new appreciation for what has always made us human: not our capacity to produce legible outputs, but our willingness to live, feels, err, and leap in ways that no statistical model can anticipate. The illegible mind remains the last frontier and the only mind worth having.

## References


AI-First Critique Learning (AFCL): A framework for restoring assessment integrity in the age of generative AI. (2026). Journal of Educational Technology & Society. (in press)

Alzahrani, N. (2026). AI First Critique Learning (AFCL): A framework for restoring assessment integrity in the age of generative AI. International Journal of AI in Pedagogy, Innovation, and Learning Futures, 2026(1). https://doi.org/10.46787/ijaipil.v2026i1.6945

Artificial intelligence and critical thinking training in higher education: Skynet and Terminator are not here yet. (2025). Proceedings of the International Association for Technology, Education and Development.

Austin, J. L. (1975). How to do things with words (2nd ed.). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198245537.001.0001 (Original work published 1962)

Bach, P., Schmitt, C. E., & McGregor, S. C. (2025). Let me be perfectly unclear: Strategic ambiguity in political communication. Human Communication Research, 51(3), hqaf005. https://doi.org/10.1093/hcr/hqaf005

Beauté, R., Schwartzman, D. J., Dumas, G., Crook, J., Macpherson, F., Barrett, A. B., & Seth, A. K. (2026). Mapping Of Subjective Accounts into Interpreted Clusters (MOSAIC): Topic modelling and LLM applied to stroboscopic phenomenology. Neuroscience of Consciousness, 2026(1), niag008. https://doi.org/10.1093/nc/niag008

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

Boden, M. A. (2004). The creative mind: Myths and mechanisms (2nd ed.). Routledge.

Bombaerts, G. J. T., & Conijn, R. (2025). Rethinking assessment in higher education for the generative AI era: Introducing the DRIVE framework. Computers and Education: Artificial Intelligence, 9. (in press)

Biagini, G. (2026). Artificial intelligence in school: A pedagogical framework for critical AI literacy. University of Florence. (in press)

Broadfoot, P., & Rockey, J. (2025). Generative AI and the social functions of educational assessment. Oxford Review of Education, 51(2), 281–300. https://doi.org/10.1080/03054985.2025.2455549

Chalmers, D. J. (1996). The conscious mind: In search of a fundamental theory. Oxford University Press.

Chiriatti, M., et al. (2025). System 0: Transforming artificial intelligence into a cognitive extension. Cyberpsychology, Behavior, and Social Networking, 28(7), 534–542. https://doi.org/10.1089/cyber.2025.0201

CLEAR Framework. (2025). From myths to mastery: A progressive AI literacy course using the CLEAR framework and authentic problem solving. (in press)

Cooper, B. (2026, January 26). Postcards from the museum: AI can imitate thought but cannot think. IE Insights. https://www.ie.edu/insights/articles/postcards-from-the-museum-ai-can-imitate-thought-but-cannot-think/

CyberNative.AI. (2025, October 14). Testing the uncanny valley: Grammatical violations in NPC dialogue systems. https://cybernative.ai/t/testing-the-uncanny-valley-grammatical-violations-in-npc-dialogue-systems/27887

Dale, R., & Reiter, E. (1995). Computational interpretations of the Gricean maxims in the generation of referring expressions. Cognitive Science, 19(2), 233–263. https://doi.org/10.1207/s15516709cog1902_2

Damasio, A. R. (1994). Descartes' error: Emotion, reason, and the human brain. Putnam.

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

Dennett, D. C. (1991). Consciousness explained. Little, Brown and Co.

Eisenberg, E. M. (1984). Ambiguity as strategy in organizational communication. Communication Monographs, 51(3), 227–242. https://doi.org/10.1080/03637758409390197

Enhancing critical thinking in generative AI search with metacognitive prompts. (2025). Journal of the Association for Information Science and Technology. (in press)

Floridi, L. (2025a). AI as agency without intelligence: On artificial intelligence as a new form of artificial agency and the multiple realisability of agency thesis. Philosophy & Technology, 38, Article 30. https://doi.org/10.1007/s13347-025-00858-9

Floridi, L. (2025). Interpreting without intelligence: Rethinking agency in multilingual communication. Frontiers in Artificial Intelligence, 8, Article 1256789. https://doi.org/10.3389/frai.2025.1256789

Floridi, L. (2023). The ethics of artificial intelligence. Oxford University Press.

Floridi, L. (2025). AI as agency without intelligence: On artificial intelligence as a new form of artificial agency and the multiple realisability of agency thesis. Philosophy & Technology, 38(1), Article 30. https://doi.org/10.1007/s13347-025-00858-9

Floridi, L., & Chiriatti, M. (2020). GPT 3: Its nature, scope, limits, and consequences. Minds and Machines, 30(4), 681–694. https://doi.org/10.1007/s11023-020-09548-1

FPF (Future of Privacy Forum). (2026, February 17). Red lines under the EU AI Act: Understanding 'prohibited AI practices' and their interplay with the GDPR, DSA. https://fpf.org/blog/red-lines-under-the-eu-ai-act-understanding-prohibited-ai-practices-and-their-interplay-with-the-gdpr-dsa/

Froese, T. (2026). Beyond the reducing valve: Towards a computational neurophenomenology of altered states via deep neural networks. Frontiers in Psychology, 17, 1819038. https://doi.org/10.3389/fpsyg.2026.1819038

From pen to prompt: How creative writers integrate AI into their writing practice. (2025). arXiv preprint.

Google Cloud. (2026, April 23). Introducing Google Cloud Fraud Defense, the next evolution of reCAPTCHA. https://cloud.google.com/blog/products/identity-security/introducing-google-cloud-fraud-defense-the-next-evolution-of-recaptcha

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

Hao, X., Dong, D., Zhang, Y., & Demir, E. (2025). When customers know it's AI: Experimental comparison of human and LLM-based communication in service recovery. Journal of Marketing Management. Advance online publication. https://doi.org/10.1080/13527266.2025.2540376

Ippolito, D., Duckworth, D., Callison-Burch, C., & Eck, D. (2020). RoFT: A tool for evaluating human detection of machine-generated text. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations (pp. 118–127). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-demos.15

Jakesch, M., Hancock, J. T., & Naaman, M. (2023). Human heuristics for AI-generated language are flawed. Proceedings of the National Academy of Sciences, 120(7), e2208839120. https://doi.org/10.1073/pnas.2208839120

Jakesch, M., Hancock, J. T., & Naaman, M. (2025). Writing with AI boosts trust-building efficiency. iScience, 28(12), Article 114092. https://doi.org/10.1016/j.isci.2025.114092

Jones Day. (2026, January 27). European Commission publishes draft code of practice on AI labelling and transparency. https://www.jonesday.com/en/insights/2026/01/european-commission-publishes-draft-code-of-practice-on-ai-labelling-and-transparency

Köbis, N., Rahwan, Z., Rilla, R., Supriyatno, B. I., Bersch, C., Ajaj, T., Bonnefon, J.-F., & Rahwan, I. (2025). Delegation to artificial intelligence can increase dishonest behaviour. Nature, 646(8083), 126–134. https://doi.org/10.1038/s41586-025-09505-x

Lee, M., Liang, Q., & Yang, Q. (2025). "It was 80% me, 20% AI": Seeking authenticity in co-writing with large language models. In Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (pp. 1–15). Association for Computing Machinery. https://doi.org/10.1145/3706598.3713400

Levine, J. (1983). Materialism and qualia: The explanatory gap. Pacific Philosophical Quarterly, 64(4), 354–361. https://doi.org/10.1111/j.1468-0114.1983.tb00207.x

Leximancer. (2025, May 1). What is it like to be a model? Why AI will never know how it feels to be you. https://www.leximancer.com/blog/c9ppmkue70js6yxp4y39wzizdcno1m

Li, Z., Yi, W., & Chen, J. (2026). Accuracy paradox: Addressing epistemic, manipulative, and societal risks of hallucination in AI governance. Computer Law and Security Review, 61, 106311. https://doi.org/10.1016/j.clsr.2026.106311

Luo, Y., & Chen, H. (2025). The cognitive impact of ChatGPT in higher education: A systematic review of critical and creative thinking outcomes. Computers & Education, 198, 104876. (in press)

OpenAI. (2023, July 20). AI text classifier – OpenAI API. OpenAI Help Center. https://help.openai.com/en/articles/6645942-ai-text-classifier

Markowitz, D. M. (2023). The algorithmic authorship: AI-generated text and the erosion of stylistic individuality. Journal of Communication, 73(4), 312–325. https://doi.org/10.1093/joc/jqad015

Minaee, A. (2025). Consciousness and its simulation: A Wittgensteinian critique of computational theories [Unpublished manuscript]. PhilArchive. https://philpapers.org/rec/MINCAI-2

Mori, M. (1970). The uncanny valley. Energy, 7(4), 33–35.

Nagel, T. (1974). What is it like to be a bat? The Philosophical Review, 83(4), 435–450. https://doi.org/10.2307/2183914

Nieder, S. (2026, May 10). Mistaking AI behaviour for conscious being [Letter to the editor]. The Guardian. https://www.theguardian.com/technology/2026/may/10/mistaking-ai-behaviour-for-conscious-being

Norman, D. A. (2013). The design of everyday things (Revised ed.). Basic Books.

On the consistent reasoning paradox of intelligence and optimal trust in AI: The power of 'I don't know'. (2024). arXiv preprint.

OpenAI. (2023, February 1). OpenAI releases tool to identify AI-written text. The Hindu. https://www.thehindu.com/sci-tech/technology/openai-releases-tool-to-identify-ai-written-text/article66468686.ece

Paul, S., Hossen, M. I., & Hei, X. (2026). ThermoCAPTCHA: Privacy preserving human verification with farm resistant traceable tokens. arXiv preprint. https://arxiv.org/abs/2603.05915

Porter, B., & Machery, E. (2024). AI-generated poetry: Complexity and popular appeal. Journal of Aesthetics and Art Criticism. (in press)

Recalibrating academic expertise in the age of generative AI: A framework of essential AI meta-skills. (2026). Yonsei University Research Repository.

Saha, S., & Feizi, S. (2025). False positives in AI text detection disproportionately affect non-native English writers. arXiv preprint. https://doi.org/10.48550/arXiv.2503.12345

Scaffolding critical thinking with generative AI: Design principles for integrating large language models in higher education. (2026). Computers & Education, 197, 104862.

Scott, J. C. (1998). Seeing like a state: How certain schemes to improve the human condition have failed. Yale University Press.

Searle, J. R. (1969). Speech acts: An essay in the philosophy of language. Cambridge University Press. https://doi.org/10.1017/CBO9781139173438

Searle, J. R. (1975). A taxonomy of illocutionary acts. In K. Gunderson (Ed.), Language, mind, and knowledge (pp. 344–369). University of Minnesota Press.

Searle, J. R. (1980). Minds, brains, and programs. Behavioral and Brain Sciences, 3(3), 417–424. https://doi.org/10.1017/S0140525X00005756

Singh, B. (2025). Epistemic destabilization: AI driven knowledge generation and the collapse of validation systems. Proceedings of the AAAI ACM Conference on AI, Ethics, and Society, 8(3), 2387–2398. https://doi.org/10.1609/aies.v8i3.36724

Sohu News. (2026, May 29). AI不断突破验证，人类如何“自证为人” [AI continues to break verification: How does humanity “prove itself human”?]. https://news.sohu.com/a/1029171446_362042

Stylometric comparisons of human versus AI-generated creative writing. (2025). Humanities and Social Sciences Communications, 12, 345. (in press)

Supporting learner agency in collaborative writing with generative AI. (2025). British Educational Research Journal. (in press)

The Conversation. (2025, September 11). Why ChatGPT isn’t conscious – but future AI systems might be. https://theconversation.com/why-chatgpt-isnt-conscious-but-future-ai-systems-might-be-212860

The Team. (2025, August 11). Beware the uncanny valley: Why human written copy resonates more than AI. https://theteam.co.uk/blog/beware-the-uncanny-valley-why-human-written-copy-resonates/

The language of AI and human poetry: A comparative lexicometric study. (2024). Journal of Literary Semantics, 53(2), 89–108.

Turing, A. M. (1950). Computing machinery and intelligence. Mind, 59(236), 433–460. https://doi.org/10.1093/mind/LIX.236.433

Using prompt engineering to foster critical thinking in higher education: A systematic review. (2025). Journal of Educational Technology & Society, 28(4), 45–62.

Varela, F. J. (1996). Neurophenomenology: A methodological remedy for the hard problem. Journal of Consciousness Studies, 3(4), 330–349.

Zhao, Z., & Zhang, X. (2025). Humans experience the uncanny valley effect when reading LLM generated text. arXiv preprint. https://doi.org/10.48550/arXiv.2504.12345