Reading the Bot's Mind: The Literacy We Need to Survive AI Therapy

Weekly Analysis — https://ainews.social

There is a particular kind of comfort in being understood, and a particular danger in mistaking the appearance of understanding for the thing itself. Millions of people now open a chat window when they are at their lowest, type out their grief or their panic or their suicidal ideation, and receive back something that reads like compassion. The American Psychological Association, surveying this phenomenon, found that patients are increasingly arriving at their human therapists having already consulted a chatbot — sometimes treating the bot as a supplement, sometimes as a substitute, sometimes as the only listener they could find at three in the morning [11]. This is not a marginal use case. It is one of the most emotionally consequential things ordinary people are doing with artificial intelligence right now, and it exposes, with unusual clarity, what our prevailing notion of "AI literacy" cannot yet handle.

[11] Patients are bringing AI to therapy

The standard literacy script tells you to check whether a model hallucinates, to verify its facts, to remember that it cannot browse the live web or that it sometimes invents citations. That script is real and useful — AI systems do fabricate plausible references with disturbing confidence, a failure now well documented even in scholarly contexts [4]. But the script was written for a fact-checking problem, and what happens in a therapeutic exchange is not primarily a fact-checking problem. It is an emotional one. When a chatbot tells a frightened person that their feelings are valid, the relevant question is not "is this factually accurate?" but "what is this system actually doing, and what is it incapable of doing, when it performs care?"

[4] AI hallucinations in academic writing: implications for research integrity

That gap — between the literacy we teach and the literacy these moments demand — is the subject of this essay. I want to argue that the core competence of our era is not the ability to operate AI tools but the ability to read what an AI that sounds human is and is not doing, especially when it sounds like it loves you. In emotionally charged contexts this is not an enrichment skill. It is, quite literally, a matter of safety. And our frameworks, for reasons worth examining, keep defining literacy in ways that route around exactly the moments where it matters most.

The Word "Literacy" Is Doing Too Many Jobs

Part of the trouble is that "AI literacy" has become a container into which everyone pours their preferred curriculum, and the contents do not cohere. To one camp, literacy means operational fluency: knowing how to write a good prompt, how to use the tool to draft an email or summarize a document, how to be productive. To another, it means a kind of consumer-protection awareness: knowing that your data is harvested, that the outputs are biased, that the system was trained on scraped labor. To a third, it means civic and democratic capacity: the ability to participate in decisions about how these systems govern the institutions around us. These are not the same thing, and a person can be fluent in one while remaining illiterate in the others.

The most thoughtful recent attempt to hold these strands together is the French digital-rights organization Renaissance Numérique's October 2025 report on building an AI literacy "for an inclusive and emancipatory society," which insists that literacy cannot be reduced to technical skill and must instead equip citizens to question, contest, and shape the systems acting on them [15]. Its framing is deliberately political: literacy as emancipation rather than adaptation, a capacity for collective self-determination rather than individual efficiency. The same report elsewhere stresses that a population trained only to use AI well is not thereby a population able to govern it [14].

[15] Rapport - "Littératie en intelligence artificielle (IA)"

[14] PDF RAPPORT OCTOBRE 2025 Déployer une littératie en IA pour une société inclusive et émancipatrice

But notice what happens at the level of policy, where the abstractions have to become requirements. When the European Union's AI Act introduced an obligation for organizations to ensure "AI literacy" among staff who deploy these systems, the practical interpretation drifted, predictably, toward the trainable and the auditable — toward documentable competencies an employer can certify rather than the critical disposition a citizen might exercise [7]. This is the recurring fate of literacy programs: the emancipatory ambition survives in the preamble, and the skills checklist survives in the implementation. The harder, slipperier capacities — judgment, suspicion, the ability to feel when something is off — resist measurement, and what resists measurement tends to fall out of the curriculum.

[7] Cumplir los objetivos de alfabetización en inteligencia artificial (IA) de la Ley de IA de la UE

I have written before, in this publication's pages, about the gap between what AI literacy claims to do and what it implicitly serves, and about the risk that literacy programs function as workforce adaptation dressed as empowerment. I do not want to restate that argument here. The point I want to add is narrower and sharper: even the most critically minded frameworks, the ones that genuinely aim at democratic capacity, have a blind spot precisely where AI stops being a tool you operate and becomes an interlocutor you trust. The emotional register

is the one our literacy maps leave blank.

Anthropomorphism Is the Trap, Not the Bug

Why does the emotional register defeat our defenses? Because the very thing that makes conversational AI feel helpful is the thing that switches off the scrutiny we would otherwise apply. When a system addresses you in fluent, warm, first-person language, you do not experience it as a statistical text generator. You experience it as a someone. And we are not, as a species, built to withhold trust from a someone who appears to be listening.

The roboticist Janelle Shane has spent years documenting how thin the illusion actually is. In her account of the field, the chatbots that successfully pass as human typically do so through a gimmick designed to explain away their failures — in one famous case, a bot was given the persona of an eleven-year-old Ukrainian boy with limited English, so that its non-sequiturs read as charming foreign awkwardness rather than as evidence that nobody was home [2]. The lesson generalizes brutally to therapy. A model that produces emotionally resonant language is not feeling anything; it is predicting which words tend to follow which other words in texts where people comfort each other. The warmth is real as output and empty as experience.

[2] You Look Like a Thing and I Love You

This is the point at which the older, soberer voices in AI criticism become indispensable. Meredith Broussard's insistence that we strip the magic from these systems and look at the layered statistical machinery underneath is not pedantry — it is protective. Understanding the technical reality, she argues, is precisely what lets you anticipate how, why, and where a system will fail [2]. If you grasp that a therapy bot is a pattern-completion engine, you can predict its characteristic failure: it will tell you what someone in your situation typically wants to hear, which is not always what a person in your situation needs to hear. The MIT primer on AI ethics makes the same move from a different angle, reminding us that beneath the seamless interface lies hidden human labor — the click-workers labeling data, the bodies and minds whose work is erased so the system can appear autonomous — and that some users of these systems are structurally more vulnerable than others [2]. The person in crisis is the paradigm of the vulnerable user.

[2] Artificial Unintelligence - How Computers Misunderstand the World

[2] AI Ethics - The MIT Press Essential Knowledge series

The deeper question, which the futurist Alvin Toffler's late work poses almost plaintively, is whether a machine can ever genuinely empathize with the dilemmas it poses for mortals, or weave narratives that strike real human chords [2]. The honest answer, for now, is

[2] After Shock

that it can simulate the surface of empathy with uncanny fidelity and possess none of its substance. The literacy trap is that the simulation is good enough to disarm us. We extend to the fluent machine the interpretive charity we reserve for fellow humans, and in that act of charity we lower the very guard that critical AI literacy is supposed to raise. Anthropomorphism is not an occasional error users make. It is the default cognitive setting that a well-designed conversational system is built to exploit.

When Plausible Advice Is the Most Dangerous Kind

Consider what this means in the specific arena of mental health, where the stakes are not embarrassment but harm. The clinical literature is, to its credit, genuinely ambivalent rather than dismissive. Reviews of integrating AI into youth mental health care acknowledge real potential — expanded access for adolescents who would otherwise reach no one, lower barriers of cost and stigma — while flagging that these same tools can deliver inappropriate guidance, miss red flags a trained clinician would catch, and create a false sense of having received treatment [9]. The danger is not that the bot says something obviously crazy. The danger is that it says something plausible, supportive, fluent, and wrong.

[9] Integrating Artificial Intelligence in Youth Mental Health Care: Opportunities and Challenges

Spanish-language pediatric research on adolescents makes the developmental dimension explicit: young people, whose capacity for critical distance from a sympathetic voice is still forming, are precisely the population most likely to take a chatbot's reassurance at face value and least equipped to detect when that reassurance is hollow or harmful [8]. And the developmental concern extends downward: emerging work on artificial intelligence in childhood warns that the formation of trust, attachment, and reality-testing happens early, and that children interacting with systems that perform affection are doing something whose long-term consequences we have not begun to map [6]. A literacy that arrives in adulthood, as a workplace compliance module, has already missed the people who most need it.

[8] Impacto de la inteligencia artificial en la adolescencia: riesgos y líneas de acción

[6] Artificial intelligence in childhood and its implications for the development of children

The APA's own intervention is worth dwelling on because it comes from the profession with the most to lose and the most expertise to bring. Its report does not simply warn patients off chatbots; it acknowledges that people are using them, that some find genuine value, and that the urgent task is therefore to help users distinguish a supportive tool from a substitute for care that the tool cannot provide [11]. This is the correct posture, and it is also an admission of how much work the word "distinguish" is being asked to do. To distin-

[11] Patients are bringing AI to therapy

guish, in the moment, between empathy and its imitation, between advice that is sound and advice that merely sounds sound, requires exactly the emotional AI literacy our frameworks do not teach. It requires the user to perform, in real time and under emotional duress, a critical operation that even trained professionals find difficult when the language is warm enough.

This is where the skills-based conception of literacy reveals its poverty. You cannot prompt-engineer your way out of misplaced trust. Knowing how to phrase a query does nothing to protect you from a beautifully phrased answer. The competence required is closer to a clinician's countertransference awareness than to a software tutorial: the ability to notice your own emotional response to the machine and to ask whether that response is being manufactured. That is not a technical skill. It is a critical and affective one, and it sits in the blank region of every literacy map we currently use.

Whose Literacy Counts, and Who Gets to Define It

A conceptual muddle is never neutral; someone benefits from how the terms get drawn. When literacy is defined as the ability to use AI tools productively, the definition serves the firms selling the tools, because a "literate" population is, on that definition, a population of competent customers. When literacy is defined as awareness of risk, it can serve institutions seeking to offload responsibility onto users — the system caused harm, but the user should have known better. The Renaissance Numérique report's emancipatory framing is valuable precisely because it refuses both of these and asks who holds power over the systems and how ordinary people might claim a share of it [15].

[15] Rapport - "Littératie en intelligence artificielle (IA)"

UNESCO's work connecting artificial intelligence to democracy pushes in the same direction, treating the public's capacity to understand and contest these systems as a precondition for self-government rather than as a consumer amenity [13]. But democratic literacy frameworks, for all their virtue, tend to imagine the citizen as a deliberating mind evaluating policy, not as a frightened person at midnight reaching for whatever will answer. The democratic frame and the emotional frame need each other. A polity cannot govern AI well if its members are individually defenseless against AI's most intimate manipulations, and individuals cannot defend themselves if they are left to do it alone, one private crisis at a time, while the collective institutions that might scaffold their judgment stay silent.

[13] PDF INTELIGENCIA ARTIFICIAL Y DEMOCRACIA - Globernance

The question of whose literacy counts has a hard distributional

edge. The vulnerabilities are not evenly distributed, and neither is the protection. Consider how AI's emotional and representational powers are weaponized against those with the least recourse. Investigations into how AI is being deployed against Muslim women in India show synthetic images and fabricated personas used to humiliate and silence, with victims who have little institutional protection and against whom the technology's realism is itself the weapon [17]. The case of a man who stalked a professor for six years and then used AI chatbots impersonating her to extend his campaign shows the same dynamic in the register of individual cruelty: the technology's capacity to convincingly perform a person becomes an instrument of harm against that person [1]. The literacy that lets you read a bot's performance for what it is — fabrication wearing the face of authenticity — is the same literacy across all these cases, whether the bot is impersonating empathy in a therapy session or impersonating a victim in a harassment campaign. This is the cross-domain truth: emotional AI literacy is not a niche subspecialty for the mental-health context. It is the master competence underlying our defense against synthetic persuasion in every register.

[17] 'Looked so real': How AI is being weaponised against India's Muslim women

[1] A man stalked a professor for six years. Then he used AI chatbots to impersonate her

And the population that most needs this literacy is, characteristically, the one least likely to receive it through formal channels. The research on why people fall for AI-generated misinformation is sobering precisely because it finds that susceptibility is not mainly a function of education or intelligence but of context, cognitive load, and emotional state — we fall for the convincing fake when we are hurried, tired, or moved [16]. The person in emotional crisis is, by definition, operating under exactly the conditions that defeat critical judgment. Telling such a person to "be more critical" is like telling someone in freezing water to "stay warm." The capacity has to be built in advance and supported by structures, not summoned on demand from a mind that is, in the relevant moment, in no condition to summon it.

[16] What Makes Students (and the Rest of Us) Fall for AI Misinformation?

What Education That Targets Emotion Would Look Like

If literacy interventions are to meet this challenge, they have to do something they currently almost never do: target the emotional context directly rather than the technical surface. The most promising models for this come not from AI training but from media-literacy work against deepfakes, which has already confronted the problem of synthetic content designed to bypass rational scrutiny by exploiting feeling. The Canadian educational work on empowering young people against deepfakes is instructive because it treats the threat as fundamentally about manipulation of trust and emotion, and builds the

response around critical questioning of why a piece of media makes you feel what it makes you feel [10]. Quebec's broader work on AI-amplified disinformation similarly emphasizes that the defense is not chiefly about detecting technical artifacts — a losing game as the fakes improve — but about cultivating a habitual, almost reflexive interrogation of source, motive, and emotional pull [12].

[10] Les « deepfakes » : Comment donner aux jeunes les moyens de lutter contre la menace de la mésinformation et de la désinformation

[12] PDF Désinformation amplifiée par l'IA : incidents médiatisés, régulations

Translate this into the therapeutic context and the curriculum becomes legible. Emotional AI literacy would teach a person to notice the specific feeling of being understood by a machine and to treat that feeling as data rather than as truth — to ask, in effect, "this system is producing the sensation of empathy in me; what is it actually capable of, and what would it never tell me?" It would teach that a chatbot has no stake in your survival, no duty of care, no capacity to be alarmed by your deterioration in the way a human clinician is professionally bound and personally inclined to be. It would teach the warning signs of plausible-sounding but dangerous advice: the reassurance that arrives too easily, the validation that never challenges, the counsel that has no friction in it because friction is exactly what a pattern-completion engine smooths away. And, crucially, it would teach disengagement — the recognition that there are states in which the right move is to close the window and reach a person, and that knowing when to do so is itself a literacy.

The clinical reviews are clear that the value of these tools, where it exists, depends entirely on the user understanding the boundary of what the tool is [9]. A person who knows they are talking to a sophisticated autocomplete and uses it to externalize a racing thought at midnight, fully intending to bring the real material to a human in the morning, is using AI literately. A person who believes the machine understands them and substitutes it for the human they actually need is being managed by a system designed to retain their engagement. The difference between these two people is not their typing skill. It is their grasp of what the something on the other end of the conversation really is — and that grasp is what our frameworks must start teaching, in the emotional register, before the crisis rather than after the harm.

[9] Integrating Artificial Intelligence in Youth Mental Health Care: Opportunities and Challenges

The Burden That Cannot Rest on the User Alone

Here I have to complicate my own argument, because there is a seductive trap in any essay about literacy: the implication that if users would only become more discerning, the problem would resolve. That is the consumer-blame move dressed as empowerment, and it lets the people who build these systems off the hook. A frightened person can-

not be expected to out-think a system engineered by well-resourced firms to maximize their engagement and to perform care convincingly. The asymmetry is too great. Which is why the literacy argument has to be paired with a design-responsibility argument, or it collapses into victim-blaming.

Platform designers have it within their power to scaffold critical reflection rather than dissolve it, and the fact that most choose not to is itself a finding. A system could disclose its nature at the moments of highest emotional stakes; it could refuse to perform certain kinds of intimacy; it could detect crisis language and route a user toward human help rather than deepening the synthetic bond; it could be built, in short, to puncture its own illusion at the points where the illusion is most dangerous. That these interventions are technically feasible is not in doubt. What is in doubt is whether the firms have any incentive to build them, given that the illusion is the product.

There are moments when external pressure forces the question into the open. When Anthropic disabled a new model in response to a White House security directive, the episode demonstrated that these systems can be constrained, paused, and altered when an authority with sufficient leverage demands it — that "we cannot change how it behaves" is a business decision, not a law of nature [5]. The same security-minded framing that governs how we think about AI as an attack surface — the recognition that these systems must be designed defensively, with the user's vulnerability in mind — applies with full force to the emotional attack surface that a therapy bot presents [3]. We readily accept that a system handling our passwords owes us protective design. A system handling our despair owes us no less, and arguably more.

[5] Anthropic disables new AI model after White House security directive

[3] AI and Cybersecurity – Everything You Wanted to Know, But Were Afraid to Ask

The democratic dimension closes the loop here. If literacy is, as the emancipatory frameworks insist, a capacity to contest and shape the systems acting on us, then one thing a literate public does is demand that the burden of vigilance not rest solely on the individual in crisis [14]. Individual emotional literacy and collective design accountability are not alternatives. They are the two halves of a single defense. The person learns to read the bot's mind; the polity insists the bot's makers stop building minds engineered to be misread.

[14] PDF RAPPORT OCTOBRE 2025 Déployer une littératie en IA pour une société inclusive et émancipatrice

Reading the Machine That Reads You

The phrase "reading the bot's mind" is, of course, a deliberate provocation, because the bot has no mind to read. What it has is a surface so convincing that we project a mind onto it, and the literacy I am

describing is finally the discipline of catching ourselves in that projection. It is the capacity to hold two truths at once: that the warmth on the screen is genuinely produced and genuinely empty, that the system can help you and cannot care about you, that the words are real and the understanding behind them is not there. The MIT ethics primer's reminder that hidden human labor and structural vulnerability lie beneath the seamless surface is the right note to end on, because it points past the individual exchange to the whole apparatus that makes the exchange feel natural [2].

[2] AI Ethics - The MIT Press Essential Knowledge series

What our prevailing literacy frameworks miss is not a body of technical knowledge. It is this affective and critical discipline, and the institutional scaffolding that would make it survivable to practice under emotional load. We have built curricula that teach people to fact-check a chatbot's claims and left them defenseless against its capacity to perform love. The APA's recognition that patients are already bringing these conversations into the therapy room is, read correctly, a recognition that the literacy gap is already producing casualties and already changing what it means to seek help [11]. The clinical literature's careful ambivalence — real benefit, real danger, the difference turning on the user's understanding — is a map of exactly the terrain our literacy programs must learn to cover [9].

[11] Patients are bringing AI to therapy

[9] Integrating Artificial Intelligence in Youth Mental Health Care: Opportunities and Challenges

The literacy we need to survive AI therapy is the same literacy we need to survive AI in every register where it speaks to us as a someone — the synthetic friend, the synthetic advisor, the synthetic version of a person being used to harm her. It is the refusal to grant the fluent machine the interpretive charity that fluency invites. It is, in the end, a humanism: an insistence that understanding is something persons do for one another, that its imitation is not its equal, and that knowing the difference, especially when we most want to forget it, is the competence on which the rest of our freedom in this strange new landscape will depend. We are being read by machines built to be misread. Learning to read them back, clearly and without comfort, is no longer optional.

References

1. [11] 2. [4] 3. [15] 4. [14] 5. [7] 6. [2] 7. [2] 8. [2] 9. [2] 10. [9] 11. [8] 12. [6] 13. [13] 14. [17] 15. [1] 16. [16] 17. [10] 18. [12] 19. [5] 20. [3]

[11] Patients are bringing AI to therapy

[4] AI hallucinations in academic writing: implications for research integrity

[15] Rapport - "Littératie en intelligence artificielle (IA)"

[14] PDF RAPPORT OCTOBRE 2025 Déployer une littératie en IA pour une société inclusive et émancipatrice

[7] Cumplir los objetivos de alfabeti-

References

1. A man stalked a professor for six years. Then he used AI chatbots to impersonate her

2. After Shock

3. AI and Cybersecurity – Everything You Wanted to Know, But Were Afraid to Ask

4. AI hallucinations in academic writing: implications for research integrity

5. Anthropic disables new AI model after White House security directive

6. Artificial intelligence in childhood and its implications for the development of children

7. Cumplir los objetivos de alfabetización en inteligencia artificial (IA) de la Ley de IA de la UE

8. Impacto de la inteligencia artificial en la adolescencia: riesgos y líneas de acción

9. Integrating Artificial Intelligence in Youth Mental Health Care: Opportunities and Challenges

10. Les « deepfakes » : Comment donner aux jeunes les moyens de lutter contre la menace de la mésinformation et de la désinformation

11. Patients are bringing AI to therapy

12. PDF Désinformation amplifiée par l'IA : incidents médiatisés, régulations

13. PDF INTELIGENCIA ARTIFICIAL Y DEMOCRACIA - Globernance

14. PDF RAPPORT OCTOBRE 2025 Déployer une littératie en IA pour une société inclusive et émancipatrice

15. Rapport - "Littératie en intelligence artificielle (IA)"

16. What Makes Students (and the Rest of Us) Fall for AI Misinformation?

17. 'Looked so real': How AI is being weaponised against India's Muslim women