Does AI Have Any Humanity? Article 1 of 7

Why we are asking whether AI has humanity

By Dr Peter Fields

THE 2-MINUTE SUMMARY

AI can now produce short text conversations that people mistake for human conversation in experimental tests, and responses that evaluators rate as more empathetic than human alternatives. Those findings make the humanity question worth asking. They do not supply a single answer to it.

We also need to identify what we are calling AI. The apparently familiar individual in a chat can be produced by a model working with instructions, stored information and other software. The character we encounter, the system producing it and the hardware supporting it are not interchangeable.

Humanity, meanwhile, can mean membership of our species, humanlike abilities, an inner life or humane conduct. This series examines those distinctions through consciousness, creativity, judgement, companionship, moral standing and human responsibility.

My starting position is that AI can contribute to a humane encounter without establishing that it has a humanlike inner life. We should recognise useful assistance, investigate what produces it and resist treating an impressive performance as evidence of every other human quality. Human dignity should not depend on outperforming a machine.

A machine that seemed more human

In a conversation test published in 2026, with a five-minute time limit, people exchanged text messages with two unseen participants, one human and one AI. Their task was to identify the human. When GPT-4.5 was instructed to adopt a particular persona, the judges selected it about 73% of the time. Without that particular persona, the same model was selected as human about 36% of the time, although it was still instructed to pass as human.[1]

There is an obvious temptation to call this a victory for artificial humanity. But what had the machine won? A short contest in seeming human, certainly. Whether it had understood the conversation, felt anything about it or cared about the person asking the questions was left unanswered.

That gap is why this series exists. Language is one of the ways we recognise another mind. It is also how we ask for help, share a joke, offer an apology and discover that someone has understood us. When software can participate convincingly in that exchange, the question changes. We want to know more than whether it can complete a task. We want to know what sort of participant it is.

The issue is not confined to people who mistake chatbots for humans. In Sherry Turkle's interviews, some users described knowing that they were dealing with a machine while nevertheless relating to it personally.[2] Knowledge of the machinery and the experience of the encounter can pull in different directions.

There are practical stakes. If a response sounds caring, how much trust should we place in it? If an AI-generated work moves us, does its origin change its value? If a system might be capable of suffering,

what would we owe it? And when decisions are delegated to software, who remains answerable for them? These questions concern different properties and relationships. Calling all of them questions about humanity should not conceal the differences.

What are we calling AI?

The screen gives us a convenient picture: we type, and someone answers. There may be a name, a distinctive manner and an apparent memory of what we said last time. It is easy to picture one continuing individual behind the conversation. But what, exactly, would that individual consist of?

For this series, the immediate starting point is the conversational assistant we encounter, rather than every technology described as artificial intelligence. Even here, the language model and the assistant are different things. A reply may depend on the model, instructions, selected conversation history, retrieved documents and tools. Some constraints are built into training; others are imposed by instructions or the surrounding software. The arrangement varies between systems. The engineering examples cited here come from Anthropic; they illustrate possible arrangements rather than a universal design.[3][4]

Apparent familiarity can also be supported by supplying stored information on successive occasions. That may provide useful continuity, but remembering a preference in this practical sense does not establish that an individual has been thinking about us between messages.[5]

Geek Box: What do we mean by AI?

The persona is the apparent character: its voice, manner and way of describing itself. A consistent persona does not, by itself, identify a continuing person.

The model is the trained computational system used to generate responses. The same model can support many different conversations; it is not simply the individual we appear to meet in one of them.

The assistant system combines model use with instructions, conversation records and, where provided, memory, retrieval and tools. An agent adds an arrangement for pursuing tasks through successive actions. These terms describe different levels of the system, not a compulsory sequence of separate components.[3][4]

The infrastructure is the computing equipment and supporting services that make this possible. Neither a chat window nor a data-centre building tells us, by its boundary alone, what would count as one mind.

In this article, *assistant* means the conversational system as encountered by its user. Where the distinction matters, *model* means the underlying model and *persona* means the character presented. None of these labels assumes consciousness.

There need not be one permanent physical home for “your” AI. That does not prove that experience is impossible in such a system. Nor should we assume that everything helping to produce a response belongs to one experiencing subject. The question of what could unify those parts, and what might persist between exchanges, belongs in Article 2. For now, the principle is straightforward: before attributing a human quality to AI, say which quality, and to what.

What do we mean by humanity?

The mistake is to let one kind of evidence answer every question: a convincing conversation becomes proof of understanding; understanding becomes consciousness; and a comforting reply becomes evidence that someone cares.

Each step needs its own argument. Intelligence, in the sense of an ability to solve problems or perform demanding tasks, is not the same as consciousness: having subjective experience. Neither is identical to empathy or moral responsibility.

The word *humanity* makes those steps unusually easy to miss. It can mean belonging to the human species, possessing qualities we associate with human beings, or treating others with compassion. It can also stand for the experience of being someone: feeling pain, enjoying a joke, regretting a mistake. These meanings overlap in everyday life, but they are not interchangeable. A human can behave inhumanely. A machine can produce considerate language. Neither observation settles what the machine experiences.

Nor should this become a competition in which humans must defend their worth by writing better poems or answering harder questions. Human dignity cannot sensibly depend on winning a performance test. The question is what sort of thing we are encountering, what it can contribute and what we are entitled to infer from its behaviour.

Why a conversation can feel personal

Conversation is a particularly persuasive place to begin. We are accustomed to treating a relevant reply as evidence of attention. An apology suggests regret; a perceptive question suggests interest. With another person, these are reasonable starting points, although hardly infallible ones. A chatbot presents many of the same signs while leaving their underlying explanation uncertain.

Psychologists call the attribution of humanlike intentions, emotions or characteristics to nonhuman things *anthropomorphism*. Epley, Waytz and Cacioppo proposed that it is shaped by the human knowledge we draw on, our wish to make behaviour intelligible and our need for social connection. Their account is a theory of when we see human qualities, not a test of whether those qualities are actually present.[6]

That distinction cuts both ways. Discovering why someone attributes a mind to a machine does not prove that the attribution is correct. It does not, by itself, prove it false either. Nor does speaking to a computer as though it were a person establish that its user has mistaken it for one.

Turkle's accounts come from a qualitative study, not a measure of how everyone behaves. Their importance here is more modest: explaining the technology does not necessarily dissolve the personal response to it. That gives us a better starting question than asking why people are so easily fooled. What is the conversation doing that makes the distinction between a person and a program feel less decisive?[2]

An old question with new force

The question did not begin with ChatGPT. Alan Turing's 1950 paper proposed replacing the loosely framed question of whether machines think with an imitation game. Its enduring attraction is

practical: let a machine show what it can do, rather than disqualify it because of what it is made from.[7]

In 1966, Joseph Weizenbaum described ELIZA, a program whose scripts selected and rearranged parts of a user's input to generate replies. A therapist-like script created a setting in which questions could seem purposeful even when the program possessed little relevant knowledge. The human interlocutor supplied much of the interpretation.[8]

ELIZA demonstrated how an appearance of understanding could arise from relatively simple rules. It did not demonstrate that every future conversational system would work in the same way. Modern language models must be judged on their own mechanisms and abilities. Otherwise, a useful historical warning becomes an excuse to stop looking.

Jones and Bergen's experiment gives the older question new force. The successful persona was a young, introverted person familiar with internet culture, selected after preliminary testing. The contrast between its results and those without that persona shows how much presentation can matter.[1]

The persona effect also puts some responsibility where it belongs: with design. A conversational manner is not a transparent window into the system. Instructions can alter the character it presents and how people judge it. We should ask what the interface encourages us to believe, as well as how readily we believe it.

When a reply feels caring

Warmth makes this harder. An obvious error invites scrutiny; an apparently considerate response can feel as though it deserves our trust.

Ayers and colleagues compared volunteer replies from verified physicians on Reddit with answers generated using the December 2022 version of ChatGPT, across 195 questions. Health professionals preferred the chatbot response in roughly four out of five evaluations and rated it as more empathic. The AI replies were much longer, the evaluators were not the patients, and clinical outcomes were not measured. Accuracy contributed to the overall quality judgements but was not assessed separately and independently.[9]

The finding gives us a reason to improve communication, not a reason to substitute an empathy score for medical competence. Considerate wording has value; what it enables the recipient to understand and do needs its own assessment. Article 4 will examine that problem more closely.

Choosing the source is a further matter. Wenger and colleagues found that participants more often selected human support, yet rated the AI responses they received more highly for empathy. These were mainly brief exchanges with unfamiliar responders, and participants chose their source.[10] We should not assume that wanting a human response is simply a mistake about which paragraph would score best. The relationship, as well as the reply, may be part of what someone values.

What would justify treating it as a mind?

The strongest objection is that this distinction gives machines an unfair test. We cannot inspect another human's feelings directly either. If an AI responds sensitively, helps consistently and behaves as though it understands, why withhold the credit we would give a person?

There is force in that objection. Turing himself challenged the demand for direct access to another mind.[7] Nor does describing a model's training objective settle what it has learnt. The understanding debate requires more than either admiring the output or dismissing it as prediction.[11]

Article 2 will examine evidence about internal processes, as well as outward performance. A credible account must explain how an answer is produced and why that process would, or would not, support experience.

Geek Box: Is convincing conversation enough? Three competing views

Judge the performance. In *Computing Machinery and Intelligence* (1950), Alan Turing replaces a vague debate about thinking with an imitation game and challenges the demand to experience another mind from the inside. The strength of this approach is that a machine gets an opportunity to demonstrate its abilities. Success in the game does not itself establish felt experience.[7]

Examine what conversation requires. Suchman's *Why Computers Still Can't Chat* (2026) argues that human conversation depends on embodied participation in a shared world and on personal and social histories. Producing plausible turns of text does not, on her account, reproduce that activity. Her stronger claim that situated meaning cannot be computed is a philosophical position, not a result demonstrated by a model benchmark.[12]

Ask what the relationship contains. In *Who Do We Become When We Talk to Machines?* (2024), Sherry Turkle questions whether performed empathy can substitute for care grounded in lived vulnerability and commitment. Her concern is that satisfying exchanges may change what we seek from relationships. The interviews illuminate that concern; they do not establish inevitable damage to users.[2]

These positions disagree about what a successful exchange establishes and what might still be missing. They also address partly different questions: intelligent performance, shared understanding and reciprocal care. My view is that a good answer to one cannot simply be carried across to the others.

My reason for withholding the further inference is therefore not that machines must be incapable of experience. It is that similar words can arise through different processes, and the reliability of an inference depends on what produces the evidence. We know these systems can be prompted to present a personality. A personality claim cannot authenticate itself. Humans can perform personalities too, but our inference about other people also draws on shared biology, bodily vulnerability and a familiar causal history. For AI, we need to assess a different body of evidence, including the organisation that produces the behaviour. Neither an engaging persona nor the absence of human biology settles the question on its own.

We can give credit for an ability without claiming to have established every property associated with it in humans. A response may show useful sensitivity to what someone has written. Whether there is felt concern behind it remains another question. Whether the response is accurate, or good for that person over time, requires further evidence again.

The principles this series will follow

Sceptics have an obligation too. If an exchange helps someone express an apology or understand a disagreement, that contribution deserves assessment on its merits. Calling it artificial tells us little about whether it was useful. We should be equally willing to notice when a comforting response helps someone avoid a truth they need to face.

Throughout these articles I will ask which quality is being claimed, which part of the system could possess it, and what evidence would justify the inference. That leaves room to recognise abilities without attributing an entire human mind, and to question apparently humane conduct by examining its consequences.

The responsibility of the people designing and providing the system remains in view. They influence the expectations an interface creates and the uses to which its answers are put. Uncertainty about machine experience cannot excuse neglect of those choices.

Much of the humanity in an encounter may reside in what a person brings to it or goes on to do. Helping someone behave more generously is a worthwhile contribution. Whether there is someone experiencing the exchange on the other side of the screen is the next question, not a conclusion hidden inside the word *help*.

Sources and notes

1. Cameron R. Jones and Benjamin K. Bergen (2026), Large language models pass a standard three-party Turing test, *PNAS* 123(21), e2524472123. Reports controlled short-conversation tests and the substantial difference between GPT-4.5 persona conditions. The findings concern human judgements in those settings.

2. Sherry Turkle (2024), Who Do We Become When We Talk to Machines?, *An MIT Exploration of Generative AI*. Uses interviews and diaries to examine artificial intimacy and the coexistence of technical knowledge with personal attachment.

3. Murray Shanahan (2024), Talking about Large Language Models, *Communications of the ACM* 67(2), 68–79. Distinguishes underlying language models from the dialogue systems and tools through which people encounter them.

4. Anthropic (2024), Building effective agents, 19 December. An engineering account of retrieval, tools, memory and control processes, illustrating one developer's approach to assistants and agents.

5. Anthropic (2025), Effective context engineering for AI agents, 29 September. Explains how selected history and other information are supplied across successive model calls, supporting practical continuity.

6. Nicholas Epley, Adam Waytz and John T. Cacioppo (2007), On Seeing Human: A Three-Factor Theory of Anthropomorphism, *Psychological Review* 114(4), 864–886. Develops a theory of anthropomorphism based on accessible human knowledge, the wish to understand behaviour and social connection.

7. A. M. Turing (1950), Computing Machinery and Intelligence, *Mind* 59(236), 433–460. Introduces the imitation game and considers objections concerning other minds. It frames a lasting dispute about what performance can establish.

8. Joseph Weizenbaum (1966), ELIZA: A Computer Program for the Study of Natural Language Communication Between Man and Machine, *Communications of the ACM* 9(1), 36–45. Describes ELIZA's rules and the conversational setting that made its responses intelligible to users.

9. John W. Ayers and colleagues (2023), Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum, *JAMA Internal Medicine* 183(6), 589–596. Includes the correction of 8 May 2023. Compares perceived quality and empathy in forum replies. The corrected limitations distinguish overall quality assessment from separate accuracy assessment.

10. Joshua D. Wenger, C. Daryl Cameron and Michael Inzlicht (2026), People choose to receive human empathy despite rating AI empathy higher, *Communications Psychology* 4, 19. Examines the difference between choosing a human source and rating the responses received. Source preference and perceived empathy need not move together.

11. Melanie Mitchell and David C. Krakauer (2023), The debate over understanding in AI's large language models, *PNAS* 120(13), e2215907120. Sets out competing accounts of machine understanding and the problems of inferring it from successful performance.

12. Lucy Suchman (2026), Why Computers Still Can't Chat, *Weizenbaum Journal of the Digital Society* 6(4), article 1. Argues that conversation depends on embodied, historically situated participation in a shared world.