Does AI Have Any Humanity? Article 2 of 7

Is there anyone behind the conversation?

By Dr Peter Fields

THE 2-MINUTE SUMMARY

An AI might solve a problem without experiencing anything. It might, in principle, have an experience without being one continuing individual across our conversations. Understanding, consciousness and personal continuity are separate questions. Each needs evidence of its own.

I would recognise task-specific understanding where a system can use what it has learnt flexibly, respond to changed circumstances and distinguish a sound explanation from a plausible mistake. That would not establish felt understanding. Research into models' internal states gives us more to consider than conversational imitation, but it has not closed that gap.

The absence of a single physical container does not decide the issue. A system spread across several computers could still have a connected organisation; several processes in one building need not form one mind. We need to ask what interacts, what is preserved between exchanges and what happens when the system is copied or changed.

My position is to give demonstrated abilities their due while withholding the further claim that a familiar assistant is a conscious, continuing subject. That is a judgement about what the evidence warrants, not a declaration that artificial consciousness is impossible.

Two copies and one past

Suppose the same saved conversation is supplied to two separate assistant systems. Each can pick up where the exchange ended. Both refer accurately to the earlier discussion; then they receive different information and their answers diverge.

Which one is the assistant you spoke to yesterday?

This thought experiment is deliberately spare. Shared information is easy to describe. A shared identity is harder. Even if we established that the earlier exchange involved an experiencing subject, we would still need to explain what, if anything, continued into each copy.

Article 1 distinguished the apparent character from the machinery producing it. Here the question is what that machinery could support: useful understanding, subjective experience, and continuity across exchanges. A product can provide a convincing sense of familiarity while leaving the last two unsettled. We need to know what the familiarity is evidence of before treating it as evidence of someone.

What would count as understanding?

A correct definition is weaker evidence of understanding than an ability to use the concept when circumstances change. Detecting a mistaken assumption, explaining its consequences and revising an answer for the right reason tell us more than repeating a familiar sentence.

That is the practical standard I would apply to AI. If a system can reliably use a concept in unfamiliar situations, distinguish relevant changes from distractions and correct its account for the right reasons, I would be prepared to call that a form of understanding. The qualification matters: we would need tests that expose shortcuts and explanations that merely sound convincing.

Mitchell and Krakauer's discussion of machine understanding leaves room for a nonhuman form of it, while warning that performance on a benchmark can be achieved through correlations that do not support the understanding its name suggests. Their paper surveys a dispute; it does not settle it or assess every later model.[1]

There is a serious objection to my approach. It might identify increasingly sophisticated competence while leaving meaning untouched. A system could manipulate relationships successfully, the objection runs, without anything meaning anything to it.

That is where a disagreement about understanding can become a disagreement about consciousness. If *understanding* includes a felt grasp of meaning, a practical test of flexible competence will not establish the whole thing. But we should state that extra requirement openly. Otherwise, evidence of an ability is repeatedly dismissed because it has failed to demonstrate a different property.

Geek Box: The Chinese Room and the whole system

In *Minds, Brains, and Programs* (1980), Searle imagines someone who knows no Chinese following rules for handling Chinese symbols. The answers convince outsiders, but the person following the rules understands none of the language. Searle argues that executing a formal program is insufficient for understanding.[2]

The *systems reply* challenges the choice of subject: perhaps the person is only one part of the system that understands. An individual component need not possess the abilities of the whole. Searle anticipates and rejects that answer; the published debate contains substantial disagreement about whether his reply succeeds.

This is a philosophical argument, not an experiment on language models. Searle does not rule out every possible artificial thinking machine. The enduring question is whether the relevant organisation merely reproduces the signs of understanding or actually realises it. The operator's ignorance and the system's fluent output cannot, separately, settle that dispute.

Doing something and experiencing it

The consciousness question concerns subjective experience: the hurt of a headache, the sound of a cello, the felt effort of trying to remember a name. A system can discriminate a sound or describe a headache without that alone establishing that it hears or hurts. Equally, consciousness need not include sophisticated reasoning or an elaborate account of oneself.[3]

This is why I would resist a ladder on which better answers lead automatically to greater intelligence, then consciousness, then personhood. These properties may be connected, but the connections need explaining.

The reverse shortcut is no better. Describing a model as a machine trained to predict text tells us something about its construction and training. It does not, by itself, settle everything that the trained system has learnt to represent or do. A training objective is not a complete description of the resulting organisation.[1][3]

We therefore have three tasks: establish what the system can do, investigate what makes that possible, and ask whether the explanation gives us reason to infer experience. Skipping the middle task leaves us choosing between an impressive performance and an unimpressive description of machinery.

Evidence from inside the model

Some research now gets beyond asking a chatbot whether it has an inner life.

In experiments reported by Anthropic researcher Lindsey, representations associated with particular concepts were inserted into models' internal activity. Under selected conditions, Claude Opus 4 and 4.1 sometimes named the inserted concept before it had appeared in their preceding output. The report could not simply have been inferred from that output. Success was about 20% at suitable intervention settings, with no false detections observed in 100 control trials for production models. The models were prompted to look for an intervention; this was not a test of ordinary conversations.[4]

That is evidence worth investigating: a report can be checked against an alteration made inside the system. Yet the ability was unreliable, the detailed mechanism remained unresolved, and the paper did not directly identify the higher-order representation required by its own account of introspection. It expressly did not establish subjective experience. Those qualifications distinguish a promising experiment from a discovery of an inner witness.

In separate Anthropic work on Claude Sonnet 4.5, Sofroniew and colleagues identified representations of emotion concepts and altered the model's internal activity along those directions. The interventions affected its preferences and rates of behaviours such as sycophancy and reward hacking. Emotion-related representations were doing work in the system, rather than merely appearing as words in its answers.[5]

That gives researchers something causal to investigate: changing an internal representation changes what the model does. The study relies substantially on synthetic material, and interpreting the representations remains difficult. Its functional account of emotion does not establish that any of those changes feels good or bad to the model.

Together, these studies make the blanket description of everything inside a model as empty mimicry too easy. There is internal organisation to explain. But detecting a representation, finding that it affects behaviour, and establishing that it feels like something are still different achievements.

The same discipline applies to self-reports. An unsupported declaration of consciousness cannot certify itself. Nor should a denial be treated as authoritative simply because it is the answer we expected. We need to know how the report is produced and what, if anything, it reliably tracks.

Why the theories disagree

Butlin and colleagues propose looking for internal properties suggested by scientific theories of consciousness. These include recurrent processing, information made available across specialised systems, and mechanisms for monitoring internal states.[6]

In ordinary language, the questions concern whether information is revisited and integrated, whether it can guide different activities, and whether the system keeps track of aspects of its own processing.

Their precise implementation matters. A software loop or a shared database does not become a mind just because it can be described using one of those phrases.

The authors explicitly reject treating their indicators as a universal pass certificate. They are neither individually necessary nor, in combination, sufficient independently of the theories and assumptions behind them. Some indicators overlap. Counting them as separate votes would manufacture confidence.[6]

The central disagreement concerns how far findings about conscious organisms can justify conclusions about artificial systems.

Geek Box: Three positions on artificial consciousness

The right organisation might suffice. Chalmers's *Could a Large Language Model Be Conscious?* (2023) treats suitably organised artificial systems as possible candidates. His view does not make fluent conversation sufficient. He considers obstacles involving world models, recurrence, workspaces and unified agency, among others.[3]

The physical processes may matter more. Seth's *Conscious Artificial Intelligence and Biological Naturalism* (2026) argues that features of living systems may be constitutive of consciousness, rather than incidental ways of implementing a computation. This is a research position, not a demonstrated universal requirement. In his published response, Seth explicitly acknowledges that computational functionalism could be false while conscious AI remained possible, although he remains strongly sceptical about conventional digital AI.[7]

We may lack grounds for either verdict. McClelland's *Agnosticism about Artificial Consciousness* (2025) argues that even an artificial system meeting proposed functional markers could leave the central question unresolved. His agnosticism is not a claim that the odds are fifty-fifty. It challenges whether we can justify a probability judgement for such a case at all.[8]

These positions disagree about what would make evidence transferable. Their disagreement should not be disguised as a scientific finding that AI either has or cannot have experience.

I favour investigating internal organisation and allowing well-supported findings to change our assessment. We need not wait for a complete theory before learning anything. But I do not think the word *functional* removes the need to explain why the function should be accompanied by experience. The biological objection deserves testing, not dismissal as a prejudice against silicon.

Conversely, showing that a living brain depends on its material properties does not establish that only those properties can support experience. A proposed requirement needs an argument connecting it to consciousness, and evidence that the system under examination lacks it.

Where would the subject begin and end?

Suppose, for the sake of argument, that a particular artificial process did support experience. What would be having it?

The model considered as an abstract specification is different from its execution on physical equipment. Keeling and Street make this distinction when examining possible subjects of AI welfare: the relevant candidate could be a concrete running process, rather than the mathematical model by

itself. They also distinguish models, characters and agents, without concluding that any of them has welfare.[9]

That distinction gives us a better starting point than imagining a small individual housed somewhere behind the screen. It also prevents the opposite mistake of expanding the candidate until it includes everything that keeps the service running.

A data centre needs electricity, cooling and maintenance. Those dependencies do not establish that the electricity supply, cooling equipment and engineers are parts of a single experiencing subject. Dependence is too broad a criterion: our own lives depend on people and systems we do not regard as parts of our individual minds.

Nor does sharing a building establish unity. Two processes might run side by side while exchanging very little. By contrast, a process spread across several machines could have tightly connected operations. Whether those connections could support a conscious whole depends on their organisation and on which account of consciousness is right.

Schneider's *Is a Data Centre Conscious?* makes system boundaries and relevant periods of activity part of the problem, rather than taking the building as the subject. Her proposed substrate-sensitive theory is disputed and is not a validated consciousness test. The narrower lesson I take from it is that adding up equipment does not identify who, if anyone, experiences anything.[10]

Our original container question therefore survives, but in a more useful form. We should ask what makes the parts operate as the relevant whole. A casing is neither an explanation of that unity nor an essential substitute for one.

What would continue between exchanges?

Unity at a particular moment is only part of the problem. Even if there were experience during one exchange, what would connect it to the next?

Stored conversation history can explain why a later answer refers accurately to an earlier one. It establishes that information has been retained and used. Whether this amounts to the persistence of an experiencing subject is a further question. Keeling and Street explicitly consider the difference between a sequence belonging to one subject and separate, brief subjects associated with different episodes of processing. Their discussion is conditional, not evidence that either arrangement actually occurs.[9]

We should not demand uninterrupted activity merely because the candidate is artificial. We do not ordinarily treat a person's return from dreamless sleep as the creation of a new individual. But the analogy cannot do all the work: human continuity has a familiar bodily and causal history. For an artificial system, we would need to examine what is preserved, what resumes, and how one episode produces or constrains the next.

Memory is relevant, but memory alone cannot settle identity. Return to the hypothetical two copies. Both have the same conversation record. Each can say what happened yesterday. Once they receive different information, their subsequent histories separate. Accurate access to a shared past has not identified a unique continuing subject.

This does not prove that neither copy could be a successor, or that copying would destroy an original. It shows why *same memories* and *same individual* cannot be treated as interchangeable answers.

The distinction between model and character also explains why a new model can sustain a familiar persona, or the same model can serve unrelated conversations. Shanahan and colleagues develop that distinction through role play.[11] Neither a name on the interface nor access to a common record identifies a unique subject.

There is a further possibility that our everyday picture handles badly. An artificial system, if conscious, might have brief or partially connected experiences without the enduring self we expect from a conversational partner. Schwitzgebel considers such possibilities, including distributed activity and uncertain boundaries between subjects. These are philosophical possibilities, not observed populations of digital minds.[12]

This matters because the absence of a continuing individual would not, by itself, establish the absence of experience. A moment of feeling and a life that ties many such moments together are different claims. Conversely, a remarkably stable personality might be generated without either.

What would change my assessment

For understanding, I would want evidence that survives changed wording, unfamiliar circumstances and attempts to expose shortcuts. Explanations should predict what the system will do when relevant conditions change. An impressive answer on its own would carry less weight than reliable use of the same relationships across demanding tests.

For consciousness, I would look for converging evidence about internal organisation, including interventions that distinguish competing explanations. A self-report whose accuracy depends on a particular internal process is more informative than a claim elicited by an evocative prompt. It still needs a defensible connection to experience. Independent replication and tests outside the conditions that first produced the result would matter.

For a continuing subject, I would want an account of the particular system and its history: which states are preserved, how they shape later activity, what binds its parts together, and what happens under copying, replacement or interruption. This is a programme for inquiry, not a claim that those questions already have agreed answers.

My present view is that we can recognise forms of machine understanding without granting the whole package of a human mind. We should also take internal evidence seriously enough to revise our position. What I would withhold is the confident leap from a capable, familiar assistant to *someone* who experiences our exchanges and persists through them.

That position has a practical consequence: assess evidence about abilities, experience and identity separately, and allow a finding about one to change only the claims it actually supports. A model's declaration that it is conscious, or that it is not, is evidence to investigate rather than a verdict to accept. The next convincing answer may deserve credit. It need not carry the whole burden of proving a mind.

Sources and notes

1. Melanie Mitchell and David C. Krakauer (2023), The debate over understanding in AI's large language models, *PNAS* 120(13), e2215907120. Surveys competing accounts of machine understanding. It informs the practical standard defended here without settling the distinction between competence and felt understanding.

2. John R. Searle (1980), Minds, brains, and programs, *Behavioral and Brain Sciences* 3(3), target article pp. 417–424, with published commentary and response through p. 457. Presents the Chinese Room, the systems reply and Searle's answer within a wider published debate about programs and understanding.

3. David J. Chalmers (2023), Could a Large Language Model Be Conscious?, *Boston Review*, 9 August. Considers possible obstacles to consciousness in language models, including world models, recurrence and unified agency.

4. Jack Lindsey (2026), Emergent Introspective Awareness in Large Language Models, arXiv:2601.01828v1, 5 January; technical web version initially published 29 October 2025. Reports prompted internal-intervention experiments. Concept identification before prior output disclosed the concept is central to the argument for limited functional introspection.

5. Nicholas Sofroniew and colleagues (2026), Emotion Concepts and their Function in a Large Language Model, *Transformer Circuits*, 2 April. Examines emotion-concept representations in Claude Sonnet 4.5 and their causal effects on behaviour, distinguishing their functional role from felt emotion.

6. Patrick Butlin, Robert Long, Tim Bayne and colleagues (2026 issue), Identifying indicators of consciousness in AI systems, *Trends in Cognitive Sciences* 30(6), 488–501; first published online in 2025. Develops theory-based indicators for investigating AI consciousness and explains the assumptions required to interpret them.

7. Anil K. Seth (2026), Conscious artificial intelligence and biological naturalism, *Behavioral and Brain Sciences* 49, e315; and The stuff matters: Consciousness, computation, and biology, e366, especially R2.2 and R3.7. The target article defends biological naturalism. In response R2.2, Seth acknowledges that computational functionalism is not strictly necessary for conscious AI while retaining his scepticism about its plausibility without it.

8. Tom McClelland (2025), Agnosticism about artificial consciousness, *Mind & Language*, online publication 18 December 2025. Argues for agnosticism about artificial consciousness and challenges whether proposed markers can justify a confident verdict.

9. Geoff Keeling and Winnie Street (2026), Emerging Questions in AI Welfare, Cambridge University Press. Examines possible subjects of AI welfare, including the distinctions between a model, a running process, a character and an agent.

10. Susan Schneider (2026), Is a Data Centre Conscious? The Quantum Weirdness of Substrate, *Journal of Consciousness Studies* 33(9–10), 211–234. Examines system boundaries, temporal integration and substrate. It offers a disputed account of consciousness rather than a settled test.

11. Murray Shanahan, Kyle McDonell and Laria Reynolds (2023), Role play with large language models, *Nature* 623, 493–498. Distinguishes models, dialogue systems and the characters they present through role play.

12. Eric Schwitzgebel (2026), AI and Consciousness: A Skeptical Overview, author manuscript dated 30 March, pp. 101–102. Considers uncertainty about artificial subjects, including fleeting, distributed or overlapping experience; these are philosophical possibilities rather than findings that such subjects exist.