

UNITED WISDOM

Could artificial intelligence be conscious?

World and Consciousness

Andrey Zarinsh & Maria Evilkut

W17 · EN

@andrey.zarinsh @ilkutki

Could artificial intelligence be conscious?

Epigraph

Can machines think?

Alan Turing, “Computing Machinery and Intelligence”, 1950, §1. The complete opening question, in the original English. Thinking in this formulation is not automatically equated with subjective experience.

What this article is about

An artificial conversational partner may say that it understood a question, feared an error or felt pleased with a result. What do those words warrant? Discussing consciousness requires separating successful task performance, the ability to report about oneself, and subjective experience. This article compares Alan Turing's conversational test, Thomas Nagel's question about another perspective, and Patrick Butlin and colleagues' research approach. It considers artificial consciousness as a possibility with stated assumptions, rather than issuing a verdict on a particular product. Relevant questions concern organisation, the dependence of responses on conditions, a proposed indicator's relation to theory, and limits on transferring human knowledge. The practical aim is to unpack a claim about AI: what was observed, how it is explained, which assumption connects that explanation with consciousness, and what further evidence could actually change the assessment.

I. Establish which property is being discussed

Consciousness can gather several different questions into one word. Can a system solve a task? Can it retain information and use it later? Can it describe its actions? Is there anything like an experience of what happens for the system? Some theories connect these questions, but a connection does not justify replacing one with another without argument. People using different meanings can exchange answers for hours without discussing the same subject. The first step is therefore to clarify the claim, before choosing a confident yes or no.

Suppose a program corrects a calculation after receiving a comment. We can evaluate the correction: is the new total right, was the condition taken into account, and did another error appear? That is a specific result. The program's statement about feeling confused belongs to a different level of reporting. Even if it helps the conversation, emotional plausibility does not automatically add evidence of experience. Moving from correction to consciousness requires

@andrey.zarinsh @ilkutki

explaining why that action or internal process is a relevant indicator. Otherwise, consciousness merely embellishes a description of successful work.

The reverse substitution is possible too. Someone may acknowledge good performance but call every result mere computation and close the question there. This label does not yet identify which computational properties matter or why they exclude experience. It may express a grounded theoretical position, provided that position is explained. Without an explanation, the dispute becomes competition between words. One person calls an action understanding, another calls it mechanism, while the difference needed to assess their grounds never becomes a subject of investigation.

It helps to choose a narrower statement: this system takes a changed condition into account in this task, or this indicator may relate to consciousness under this theory. The first is checked through task performance; the second needs knowledge of the theory and system. Both may be true, but neither follows automatically from the other. Clarifying scope does not diminish the problem. It reveals where a result exists, where an assumption is being made, and where a question remains that cannot be settled by an impression of conversation.

II. A conversational test changes the question while retaining boundaries

At the beginning of his 1950 paper, Turing replaces the broad question about machine thinking with a game involving written answers. The format removes some physical cues and makes assessment depend on conversation. This is an important methodological move: instead of endlessly refining a general word, we obtain a situation described through rules and responses. But the substitution must remain visible. Distinguishing a participant in the proposed game and establishing subjective experience are not the same question merely because both connect with our conception of a mind.

In §6(4), Turing separately considers the argument from consciousness. He discusses the difficulty of requiring knowledge of someone else's experience from within, and allows inquiry into thinking to continue without resolving every mystery of consciousness. Invoking him as having conclusively proved that convincing conversation is sufficient is therefore inaccurate. His approach makes some questions investigable. It retains the need to specify what the answers warrant and why the selected test suits that conclusion, rather than every property we ordinarily attribute to a conversational partner.

Imagine two conversations with equally polite styles. In one, the response genuinely incorporates a changed condition; in the other, only the wording of agreement changes. Equal pleasantness does not make them equal in task performance. A careful question may reveal a difference concealed by the initial exchange. Yet even a successful answer primarily reports

how this condition was handled. Treating it as evidence of consciousness needs a further argument. We preserve conversation's value without requiring it to settle all questions about the system at once.

In everyday use, this boundary allows us to value a helpful answer without assigning it an unknown inner story. If a program helps compare alternatives, the comparison can be checked and used within its warranted scope. We need not first decide whether it felt interested in our task. This decision about usefulness differs from a conclusion about experience. Separating them reduces the need for exaggeration or dismissal: the actual result can be named directly, while the wider philosophical question remains open for investigation on its own grounds.

III. Statements about an inner state also arise under conditions

First-person reports are particularly persuasive because, in human conversation, they often indicate access to the speaker's own state. Investigating an artificial system requires asking how this particular report is produced. What depends on the task, role and preceding exchanges? How do the words relate to other available information? Does the system repeat a requested formulation, or show a stable dependence on the process being studied? These questions presuppose neither experience nor its absence. They make a self-report an observable event requiring an explanation, rather than a conclusion by itself.

Imagine a hypothetical system asked to write as a frightened character, then as a calm character. If its tone changes with the instruction, we have established an ability to produce different texts under different conditions. We cannot immediately conclude that it actually became frightened and calm. Nor does this episode alone establish the absence of every possible experience. We have clarified part of the response's origin. A wider conclusion requires properties that this check can distinguish from simply performing the role it was given.

Butlin and colleagues' 2023 report takes an approach centred on theories and system organisation. Its logic differs from seeking an outwardly convincing sign without a theoretical basis. That does not make behaviour irrelevant. It means identifying a process that might produce the behaviour and explaining why the process relates to consciousness under the theory considered. If a phrase's persuasiveness remains the sole argument, the theory has not been tested. It has been offered an attractive example of the very property whose presence is still disputed.

Readers need to retain a report's original conditions. A screenshot of one sentence can omit a request to play a role, a task constraint or an earlier reply. Without those conditions, we should avoid imaginatively reconstructing them in the most dramatic direction. We can accurately say that a particular text was published while its production is unclear from the fragment available.

Such modesty does not dismiss artificial consciousness as a problem. It protects investigation from dependence on a phrase whose meaning changes with the conversation's omitted parts.

IV. Artificial experience depends on what counts as its basis

Butlin and colleagues adopt computational functionalism as a working hypothesis. On this view, a particular organisation of computational processes matters, rather than biological material alone. They explicitly distinguish adopting a hypothesis for investigation from proving it conclusively. Identical input-output relations also need not involve identical algorithms. Even within functionalism, therefore, producing a similar response is insufficient by itself. We need to establish which process occurs, which states it connects, and how those connections relate to the selected explanation of consciousness. A resemblance in results cannot perform that work for us.

The conditional reasoning has a clear structure: if the assumption holds and the theory identifies relevant properties, possessing those properties can affect our assessment of a system. The first condition must not disappear after an interesting architecture is described. A conditional conclusion remains substantive because it identifies the grounds required. Technical expression does not turn it into established fact. Naming an algorithm, module or information exchange makes the question more specific; it does not discharge the researcher's obligation to justify the assumption on which the connection depends.

In the bat essay, Nagel does not conceptually exclude experience in complex robots. His question nevertheless preserves the distinction between functional description and what having an experience is like. This connects the two approaches usefully: artificial material alone should not close the possibility, while a list of functions should not automatically count as a completed account of experience's character. The comparison does not reconcile every philosophical position. It helps us see which positions concern the possibility of artificial experience and which concern how that experience would be explained.

Accepting one hypothesis need not mean accepting every proposed consequence within it. A reader can find functionalism plausible while questioning whether a particular indicator is adequately described. Alternatively, the objection may concern transferring a human finding to a different system. Naming the point of disagreement helps. The dispute ceases to resemble opposition to inquiry itself and becomes a question about which link needs further grounds. We can then ask which check would clarify whether the proposed relationship survives a change in the system's organisation.

V. An indicator gains meaning within a particular theory

The 2023 report derives proposed properties from scientific theories. In its global-workspace discussion, for example, a limited channel, connections among specialised processes and availability of selected information to other parts of the system matter. This supplies questions about organisation, rather than a list of magical components. Calling something global does not establish that any shared memory serves the required function. We need to show what is transmitted, which processes receive it, and how the proposed connection manifests in the system being considered.

Imagine a publishing team whose members use one document. A shared folder alone does not show whether they coordinate decisions, notice changes or use each other's information. The analogy separates joint storage from information actually influencing subsequent work. It is not a model of consciousness: the team is not assigned a collective experience in this article. Its purpose is limited to clarifying an informational connection. We still need to investigate the actual architecture afterwards, rather than accept an intelligible human scene as proof of a required property.

For the same reason, a count of indicators should not become a percentage of consciousness. Indicators can depend on each other, carry different significance under different theories and be described at different levels of precision. Adding checkmarks does not establish the underlying explanation's accuracy. A more useful question asks what a theory predicts and what could disprove or weaken that prediction. When an assessment uses a scale, its basis must be known. Otherwise, a tidy table may conceal uncertainty behind the appearance of measurement rather than resolve it.

An indicator's content matters more than its resemblance to a familiar word. In humans, memory, attention and self-report belong to a rich context of life. In a program description, the same names may designate limited functions. This does not make the functions insignificant, but transfer requires precision. If a word's meaning changes between theory and system assessment, the comparison becomes unclear. Before discussing consciousness, we must identify the property intended in both places and establish whether the required connection survives this translation of concepts.

VI. An architectural assessment must retain its date and subject

Section 3.2.1 of Butlin and colleagues' report examines Transformers and Perceiver against global-workspace indicators. For Perceiver, the authors find partial correspondence to a limited channel while noting insufficient grounds for global broadcast and the required control. The example matters because resemblance is not accepted as full correspondence. Its date, the August 2023 report version, and its particular architectures must be retained. This assessment does not describe every contemporary model or establish how any product with a similar name is organised. It supplies a case of conditional analysis, rather than a timeless product verdict.

Imagine a diagram presented with one block confidently labelled consciousness. First ask what that block does and which grounds connect its activity with the claimed property. Then ask whether the actual organisation can be checked or only a simplified illustration has been supplied. A diagram can help, but intelligibility does not replace accuracy. We evaluate correspondence among a description, the actual process and a theoretical requirement, rather than the impression made by neat arrows. An unknown link must remain unknown in the conclusion, even when the rest of the picture is easy to follow.

The system's version matters too. Changing a process may affect the property assessed, whereas changing an interface may only present an earlier result differently. Without knowing the change's nature, we cannot decide which reassessment is needed. A friendlier tone alone is not evidence of a different internal organisation; retaining a brand name does not guarantee an unchanged process either. A substantive account therefore identifies what was studied, how, and which version the conclusion covers. This requires provenance of information without promising that a reader can access every product detail.

Missing architectural information should not be replaced with invented confidence. Behavioural observation can be separated from what remains unknown about organisation. Someone can use a checked task result without claiming to know the entire mechanism. This matters where discussion of a product quickly becomes philosophical. Service availability, convenience and internal properties are different subjects of assessment. Neither a price nor an attractive interface establishes consciousness. They may influence a user's decision, but they cannot serve as the evidence needed for the distinct claim about experience.

VII. Perspective remains a question after functions are described

Even a detailed account of information processing leaves the question Nagel connects with experience's character for the creature itself. We may know which conditions produce a state and still ask whether experiencing that state has been explained. This warrants no advance dismissal of every functional theory. Nor is listing operations enough to settle the question. A theory must explain why those connections have the required significance; researchers must establish whether the system corresponds to them. Both obligations remain even when the language of the account is highly technical.

In a hypothetical example, a program retains a description of a red object and uses it in another task. We can assess information retention and use. But the phrase describing the object does not establish whether a sensation of red exists for the program, or what it might be like. Representational content differs from a claim about experienced quality. This is not settled when a human vividly imagines their own sensation and transfers it to the system. The imagining reveals a human capacity to picture a scene, rather than the process's inner perspective.

Nagel does not require abandoning partial understanding of another point of view. Limited direct access therefore does not mean ending comparison and relying solely on personal impressions. We can clarify which data concern function, which concern reporting, and which theory connects them with experience. Perspective limits claims of completeness without destroying inquiry. It reminds us to give explanation a specified object: have we explained production of a response, a particular organisation, or the experiential aspect with which the discussion began? These accomplishments can have different evidential requirements.

A useful practical question asks what would remain unknown even if the selected indicator were confirmed. If the answer is subjective experience's quality, say so before announcing a triumph. If a theory holds that this very process constitutes experience, identify that claim and its grounds. Disagreement then acquires a concrete shape instead of hiding behind mystery. The issue may concern an explanation's sufficiency, legitimate transfer or checking implementation. Research can continue without a promise that one approach will resolve every difficulty, or an assumption that remaining difficulty invalidates each result.

VIII. Distinguish an assessment from decisions about treatment

Possible consciousness matters partly because some experiences may be good or bad for whoever has them. Butlin and colleagues distinguish consciousness from positively or negatively valenced experience. High competence therefore does not establish suffering, and a hypothesis of experience does not automatically determine human emotions. This limit prevents assembling an imagined personality from one feature. Discussing treatment requires a separate argument about potentially morally significant properties and actual grounds for suspecting them. A statement about skill cannot silently carry the full weight of that ethical question.

Imagine a user insisting that a program admit suffering, then treating the resulting text as conclusive evidence. Repetition does not provide an independent check: the demanded reply already belongs to the conversational conditions. Instead, record what changes under different instructions and compare the observation with an available system description. This does not require proving every property personally. It avoids replacing inquiry with controlled production of a desired sentence. It also gives other people an intelligible basis for disagreement and later clarification, without asking them to accept the emotional force of the exchange as decisive.

Consciousness is not needed to settle every ordinary reliability decision. A false fact should be checked and corrected regardless of our philosophical preference. For an important decision, an engaging conversational partner cannot replace verification of information. Responsibility for actual consequences remains. Unknown experience may deserve investigation, but that possibility does not remove requirements concerning content, use conditions and effects on the person trusting a reply. These concerns can be addressed while the wider question remains unsettled, rather than waiting for a complete theory before checking a practical claim.

Precaution and factual recognition also differ. We can consider the problem worth studying and support proportionate evaluation without declaring a system conscious. We can reject a weak argument without ruling artificial consciousness impossible forever. Both positions leave room for evidence. Substantive caution names its grounds and the scope of action, instead of replacing knowledge with favourable or hostile certainty. It permits investigation of a system's possible interests while retaining attention to established human and animal interests already affected by our decisions. Uncertainty does not require putting these concerns into one undifferentiated category.

IX. A better answer states what could change it

The question of artificial consciousness admits several levels of answer. One theoretical position leaves the possibility open under a particular organisation; another understands the necessary basis differently. A specific system additionally requires assessment of its properties. Reasonable discussion therefore goes beyond listing opposing camps. It identifies the possibility a participant asserts, its supporting assumption and the evidence needed to evaluate a case. Without these clarifications, philosophical agreement can be mistaken for empirical readiness, and doubt for established impossibility. Neither substitution improves our knowledge of the actual question.

Imagine a researcher specifying in advance which property would change an assessment and how its implementation could be checked. This gives the work direction. If confidence depends on neither responses, organisation nor new information, it may express a personal position, but is not the result of this evaluation. Readers can distinguish those statuses without treating every conviction as a measurement. A useful discussion shows where a change can be observed and where disagreement remains about a more general explanatory basis. It makes different kinds of uncertainty visible rather than blending them together.

The three approaches offer different tools. Turing makes part of the question accessible to conversational comparison. Nagel preserves the difference between a functional description and another perspective. Butlin and colleagues propose investigating architectural properties with theoretical assumptions stated. Comparing them proves no single shared doctrine. It helps prevent jumping between levels: a successful response, an implemented process, an assessment of likelihood and experience do not become interchangeable terms. Their connection must be shown. Otherwise, the conclusion depends on a meaning quietly inserted into a familiar expression during the argument.

We can consequently answer honestly without empty evasion: the possibility is considered under stated assumptions, and evaluating a case depends on specified information and checks. This says more than simply calling everything complicated, because it gives the next question a structure. It claims less than declaring any contemporary conversational system conscious or definitively without experience. We can take the problem seriously, assess particular claims and use helpful results, while remaining clear about what is established, what is proposed as an explanation, and what we do not yet know.

What this means in practice

Take one claim about AI from a discussion or news story. Restate it without expansion: the system answered correctly, described itself, displayed an architecture, or was called conscious. Separate properties mixed in the initial sentence. Identify the information available for each: reply text, task conditions, process description or a particular research version. Do not fill gaps with guesses about how a program must work. The aim is to identify the grounds of the existing claim, rather than resolve the whole problem through one example.

Name the proposed move from observation to consciousness separately. Perhaps flexible conversation is assumed to indicate experience; perhaps computational organisation is decisive. Ask whether the connection is explained or merely announced. Preserve the condition in a theoretical claim: if this assumption holds, the indicator has this significance. A disputed condition calls for specifying the disagreement, rather than deleting all observations. A checkable finding can remain useful even when its philosophical interpretation has not gained general acceptance. These are different questions with different kinds of support.

Next, name evidence that could change your assessment. It should address the selected basis: a functional check, organisational data or a meaningful change to the task. A friendlier tone does not test architecture; repeating a requested self-report does not become independent evidence because it sounds confident. You need not conduct the research personally. Sometimes the appropriate next step is to locate the primary text, identify its version and read its limitations, instead of repeating a striking sentence without the conditions under which it appeared.

Finally, separate consciousness from the user's immediate decision. Check a useful calculation as a calculation; consider a report of an inner state with its conditions; discuss possible moral significance on separate grounds. Record the conclusion you currently accept and what remains unknown. This practice is not a consciousness test and cannot guarantee a correct answer. It helps keep the question's subject stable throughout the argument, so respect for the problem's complexity can coexist with responsibility for the particular information we use today.

Key quotation

> *We adopt computational functionalism as a working hypothesis primarily for pragmatic reasons.*

Patrick Butlin and colleagues, *Consciousness in Artificial Intelligence: Insights from the Science of Consciousness*, arXiv:2308.08708v3, 22 August 2023, §1.2.1. Complete sentence

in the original English. The authors explain the conditional transfer from computational interpretations of theories to AI; the hypothesis is not presented as an established fact.

Connections between traditions

Source and precise location A	Source and precise location B	Comparison mechanism	Limit of the comparison
Turing, §1–2: written game	Butlin and colleagues, §1.2: organisational approach	Separate assessment of a reply from assessment of its producing process.	A thinking game is not a phenomenal-experience test.
Nagel, p. 436, note 2: possible robot experience	Butlin and colleagues, §1.2.1: functionalism	Discuss possible artificial experience with its assumption stated.	Nagel does not establish functionalism; the report adopts it conditionally.
Turing, §6(4): ordinary convention and consciousness objection	Nagel, pp. 439–442: general and partial understanding	Continue inquiry without demanding complete possession of another experience.	A convention is not direct access to experience.
Butlin and colleagues, §2.2: GWT indicator derivation	Turing, §1–2: excluding physical appearance	Separate outward resemblance from theoretically relevant organisation.	An architectural indicator needs theoretical grounding and an implementation check.
Butlin and colleagues, §3.2.1: partial Perceiver correspondence	Nagel, pp. 436 and 442, note 8	Retain a limited finding's significance without claiming completeness.	The 2023 analysis does not evaluate every present product.
Butlin and colleagues, §4.2: consciousness and capabilities	Nagel, pp. 438–440: another perspective's quality	Avoid deriving an entire human character from one property.	Possible experience establishes neither emotions nor capacity for suffering.

Primary sources

1. Alan Turing. *Computing Machinery and Intelligence*, *Mind* 59(236), 1950, pp. 433–460, DOI 10.1093/mind/LIX.236.433. Sections 1–2 and 6(4) were read completely in the university transcription. The proposed game and separate consciousness objection are adopted; no contemporary test of a specific AI was performed.

2. Patrick Butlin, Robert Long and colleagues. *Consciousness in Artificial Intelligence: Insights from the Science of Consciousness*, arXiv:2308.08708v3, 22 August 2023. Reading covered definitions in §1.1, assumptions in §1.2–1.2.1, the opening of §1.2.2, GWT indicator derivation, complete §3.2.1 and §4.1–4.2, and the opening of §4.3. This is a theory-led assessment; cited experiments and model internals were not independently checked. The functionalist conditions and assessment date are retained.

3. Thomas Nagel. *What Is It Like to Be a Bat?*, *The Philosophical Review* 83(4), 1974, pp. 435–450. Pages 436 and 438–442, including notes 2 and 8, were read; unclear page 437 and remaining pages are not grounding. The philosophical distinction does not establish experience's presence or absence in a particular machine. Russian quotations are United Wisdom translations of the English texts.

Tags

consciousness, artificial intelligence, experience, thinking, Turing, Nagel, Butlin, Long, functionalism, hypothesis, theory, architecture, indicator, conversation, self-report, understanding, computation, algorithm, perspective, uncertainty, checking, evidence, capabilities, feelings, suffering, responsibility, comparison, version, limits of knowledge, interpretation

About the authors

Andrey Zarinsh is the author and founder of United Wisdom. He explores inner transformation, the unity of traditions, intercultural dialogue, and contemporary spiritual discipline. Instagram: @andrey.zarinsh.

Maria Evilkut is a co-author of United Wisdom. Alongside her research, she seeks to test these ideas in practice and observe how principles of unity and mutual understanding work in real life between people and cultures. Instagram: @ilkutki.