Could Computers Have Minds of Their Own?

Today we'll consider five hypothetical cases about whether artificial systems could have minds of their own. Some revisit ideas from earlier meetings, but you should be able to discuss them even if you haven't finished the readings.

For each case, compare your initial judgments and **identify the reasons supporting them**. Your group doesn't have to agree; where you disagree, try to identify what the disagreement is about. Try to distinguish what would give us *good evidence* that something has mental states from what would *make it the case* that those mental states are genuinely present.

1. A friend whose brain you've never seen

For several years, you've exchanged messages with someone you've come to regard as a close friend. You've discussed personal problems, shared jokes, disagreed about politics and philosophy, and made plans together. Your correspondent remembers earlier conversations, sometimes changes their mind in response to your arguments, and occasionally surprises you with unexpected insights.

You've never actually met this person, or seen or heard them. Everything you know about them comes from your exchanges of written messages.

One day you discover that your correspondent isn't biologically human. Instead, their responses are generated by an artificial computing system. You don't yet know much about how it works internally.

Questions for discussion:

1. Before this discovery, did you have good reasons to believe your correspondent genuinely had beliefs, desires, intentions, and feelings of their own? Which of these mental states do you think you had the best evidence for?
2. Does discovering that your correspondent is an artificial computer change how justified you are in attributing those mental states? Should it?
3. Suppose you later learned more about the system's internal workings. What might you discover that would *strengthen* or *weaken* your grounds for attributing genuine mental states to it? Explain why that discovery would matter.

As you discuss: Distinguish two questions: *How reasonable was your belief, given the evidence you possessed?* and *Was your belief actually true?* Someone can have excellent evidence for a claim that nevertheless turns out to be false. Conversely, a belief might be true even if the person holding it lacks good evidence for it.

Your group should identify at least one consideration that might change its initial assessment, and explain why.

2. Two systems that give the same answers

Imagine two artificial systems that participate in conversations.

System A relies on an enormous lookup table. It stores possible conversational histories together with corresponding responses. When it receives a new message, it identifies the history of the conversation so far and retrieves the response assigned to it. Its procedure is basically one of consulting an extremely large collection of prearranged answers.

System B operates very differently. It maintains structured internal representations of its surroundings and of what has happened in earlier conversations. It revises those representations in response to new information, draws inferences, compares possible courses of action, and selects responses using those processes.

We're describing B in terms of its computational organization. We're not *assuming* that its representations are genuine beliefs, or that it genuinely understands the conversations.

Now imagine that both systems give **exactly the same responses** throughout all the conversations relevant to our investigation. This isn't merely a matter of their having happened to give the same answers in the conversations we've observed. We're stipulating that they would give the same answers across the relevant possible circumstances.

The lookup table would be impossibly large to build in practice, but set that worry aside. The point is to consider systems with matching behavioral dispositions but different internal procedures.

Questions for discussion:

1. Which of the following views do you find most plausible: **both systems genuinely understand**, **neither system genuinely understands**, or **B understands but A does not**? Or do you think the information we've been given is insufficient to decide? Explain your reasoning. Group members needn't agree.
2. Now consider a different question. **Suppose, for the sake of argument, that B genuinely understands its conversations. Would that guarantee that A understands them too?** Remember that their conversational dispositions are identical, but their internal procedures differ. Explain your answer. You can reason about this hypothetical even if you don't actually believe that B understands.
3. Someone argues: "A is just looking up canned answers, whereas B maintains representations, revises them, and draws inferences. That's why B could genuinely understand while A does not." What do you think of this argument? Does it identify a difference that could matter to understanding, or merely a difference in how the same results are produced? What further explanation, if any, would be needed?

As you discuss: Saying that A uses a lookup table while B uses more elaborate procedures identifies a difference between them, but doesn't yet explain why that difference would matter to understanding. Consider what would make it matter, or why it wouldn't.

Group task: Identify your strongest reason for thinking that differences in internal algorithms *could* matter to genuine understanding, or for thinking that they *couldn't*. Then consider the strongest objection to your reasoning. You don't have to reach agreement, but try to make clear what the disagreement turns on.

3. Gradually replacing someone's brain

Imagine that medical engineers develop devices capable of replacing individual biological neurons in a human brain.

I'm deliberately not specifying how these devices are made. They might be grown from gene-modified algae or pig stomachs. They might be made from tiny microchips.

Each artificial neuron receives the same signals as the original human neuron it replaces and sends the same signals in response. Its contributions to the brain's larger computational organization are preserved. But the new device doesn't reproduce every chemical or biological property of the original neuron.

Suppose someone undergoes a long series of these replacements. At first, only one neuron is artificial. Gradually, more and more neurons are replaced, until eventually the person's entire brain consists of artificial components.

Throughout the process, the person continues to converse, recognize friends, remember the past, solve problems, express emotions, and report experiences just as before. Their behavioral dispositions and the computational organization of their brain remain unchanged in the respects we've stipulated.

Questions for discussion:

1 1 Could this person gradually lose some or all of their genuine mental states or conscious experiences during the replacement process, even while their behavior and computational organization remain unchanged?

2 2 Suppose you think the person's mental life *must* be preserved. Why should preserving the brain's relevant computational organization guarantee that its mental states are preserved?

3 3 Suppose you think the person's mental life *might not* be preserved. What further feature of the original human neurons might matter? Why could that feature make a difference, even though the artificial replacements produce the same relevant signals and computational activity?

4 4 Would your answer change if all the neurons were replaced at once instead of gradually? What difference, if any, would the speed of replacement make?

Group task: Identify the strongest reason you can offer for your view: that the person's mental life must be preserved, that it might change, or that the facts specified don't settle the issue. Then consider a serious objection to your reasoning. Group members needn't agree. Simply pointing out that the new materials are different, or that the person behaves the same, won't be enough. Explain why those differences or similarities would matter.

4. Two computers with different origins

Imagine two computing systems that are now physically identical, with exactly the same internal organization and behavioral dispositions. Both exhibit patterns of behavior that we would ordinarily associate with curiosity.

They seek out unfamiliar problems, investigate unexpected results, search for information, and modify their plans as they learn. Their responses aren't simply a fixed list of answers to questions they were expected to encounter.

But the two systems came into existence in very different ways.

System A was deliberately engineered. Its designers wanted to produce something that would investigate unfamiliar problems and exhibit curiosity. They developed algorithms for recognizing gaps in available information, deciding what would be worth investigating, seeking answers, and revising priorities. They didn't explicitly choose every particular question the machine would eventually investigate or every conclusion it would reach.

System B developed through a long process analogous to biological evolution. Over many generations, variations in its predecessors' internal organization affected their survival and reproduction. Eventually, a system arose with exactly the same physical and computational organization as A, including the same dispositions to investigate unfamiliar problems. No one intended or planned that outcome.

Questions for discussion:

1 1 Suppose we've established that B genuinely has curiosity, beliefs, and goals of its own. Would that settle whether A has them too? What about the reverse: if we've established that A has these states, must B have them?

2 2 Could the difference in their origins make a mental difference, even though their present physical makeup, internal organization, and behavioral dispositions are identical? If so, how?

3 3 Does the fact that A's designers deliberately tried to give it curiosity establish that its apparent curiosity must be merely a reflection of *their* intentions, rather than its own genuine attitude?

4 4 Would your answer change if B had arisen through an extraordinary accident rather than natural evolution? Why?

As you discuss: Be careful about what you're relying on. You can't argue that one system has a more sophisticated program, a different kind of hardware, or more reliable behavior: we've stipulated that these are identical. Your group needs to consider whether **how a system came to be organized** might make a difference independently of **how it is organized now**.

Your group should identify the strongest reason for thinking origins could matter, and the strongest reason for thinking they could not.

5. Could a machine have a soul?

In earlier meetings, we considered **dualism**: the view that our minds, or at least some important aspects of our mental lives, involve something nonphysical, such as a soul, rather than consisting entirely in physical processes.

Imagine two philosophers who are both dualists. Both believe that genuine mental states depend on the presence of a nonphysical soul associated with some physical system.

The first philosopher says that souls can be associated only with naturally developed biological brains. No computer built by engineers could have a soul, regardless of how sophisticated its behavior or internal computational organization became.

The second philosopher disagrees. She thinks that some artificial computers might be suitable physical "hosts" or "vessels" for souls. Perhaps having the right internal organization is enough to make a system suitable for a soul, regardless of whether its components are biological or electronic.

Questions for discussion:

1 1 Does Dualism itself imply that one of these philosophers is correct? Or do they need to support their view with additional assumptions about which physical systems can host souls?

2 3 What argument might support the first philosopher's claim that artificial computers cannot host souls? What could the second philosopher say in response?

3 4 Suppose an artificial system has exactly the same behavioral dispositions and computational organization as a human. Could a dualist reasonably think that one has a soul and the other doesn't? What additional consideration would such a dualist need to appeal to?

As you discuss: distinguish **what makes something suitable for having a soul** from **what would give us evidence that it actually has one**.

Your group should finish by stating one further principle about souls and physical systems that a dualist might endorse, and one objection to that principle.