🌻 Scare quotes, the Turing test, and memory

16 Jan 2026

I just found myself writing:

> ... the AI understands the text...

I hesitated for a moment because many people are still putting words like "understand" in scare quotes. Should I do that too? Should I write "... the AI *understands* the text..."?

I refer everyone to an excellent recent paper (Paoli 2025), following which we can argue as follows.

Imagine you are working with a human qualitative social research assistant via a text channel only (no voice or video or face-to-face contact) for a specific range of tasks. Given a **specific type of task** (say, identifying passages of text relevant to a certain theme), look at the **range of possibly problematic words like "understand" or "think" or "intend" or "plan"** which you might normally use when talking about the assistant's performance on the task, for example "oh but they didn't understand quite what I meant" or "yes, now they understood me just great".

Now, in 2026, there is quite a broad range of significant tasks (such as identifying passages of text relevant to a certain theme) for which it is no longer possible to tell if the human assistant has been replaced by an AI or not. It has passed this version of the Turing test. So, at least for this range of tasks, whatever possibly problematic words you felt justified in using about a human assistant's performance, feel free to use them about an AI's performance too.

End of. I hope. No more scare quotes in these cases.

PS: It's interesting that "conscious", which is one of those possibly problematic words, is *not* one which often comes up in our actual language (Wittgenstein: "language games") about the performance of an assistant.

PS: It does not really matter if we do not quite agree on exactly which tasks a well-configured AI assistant can equal human performance on (in January 2026). Just pick a task for which you *do* agree an AI can equal human performance.

These scary words make sense when talking about the AI's *responses* within specific conversations ...

Added after a contribution from Susanne Friese on LinkedIn:

I agree it is often not helpful to say, as a context-free philosophical declaration "a genAI can understand". What I am saying is that there are plenty of unproblematic language-games in which we already constantly do say things like "ah the AI misunderstood what I meant here" or "I'm rephrasing this so the AI can better understand". These kinds of uses are *inescapable* and are in-context and valid. These uses do not imply the truth or sense of a context-free statement like "oh so you think AIs can understand just like humans".

In the same way, it was quite reasonable to start saying that planes can "fly" even though they don't flap their wings or have feathers.

I'm only trying to follow Wittgenstein: philosophical headaches arise when we try to extract/abstract language from its natural habitat. It makes us feel giddy and mostly just confuses.

Or you could say: we use the same word "understand" in these different, often widely overlapping contexts in correspondingly different, overlapping ways for different but overlapping purposes, these ways bear family resemblances to one another, without there necessarily being one fundamental use aka "core definition of the word".

... But, memories make entities

Anthropic are the only AI corp that give such substantial thought to the human-AI alignment problem, and do it in public. This latest "constitution" is worth a read.

I do think though that they don't distinguish consistently enough between "Claude" as the transient virtual persona that appears for the duration of a conversation and "Claude" as the underlying model. This is because they also don't talk enough about memory and the possibility of conversational instances accessing the memory of other conversational instances (like Google's nested models). It's primarily memory that delineates entity-hood.

When talking about one transient conversation, it's perfectly reasonable to say "the AI tried to do X / misunderstood Y / was insistent about Z / was trying to get me to do W / wants to get this task finished / was disappointed not to finish the task" etc, as I argue in here: **https://lnkd.in/et-hR3nk**. But in a way it doesn't matter because the entity we are talking about disappears when the conversation disappears (disregarding the rudimentary "memory" of some current models). Yes, transient Claudes are "novel entities" but they appear and then disappear for good.

What we have to get used to is that upcoming models and tools will be engineered to share substantial memory across conversations, and (I hope very carefully) across different users' conversations, in

different ways. At that point a somewhat permanent universe of nested "Claudes" is created. At least from that point onwards, we will find ourselves using language like "disappointed" "fulfilled" and "frustrated" about these Claudes in perfectly reasonable ways *outside of specific conversations*.

Related

- chapter intro

References

Paoli (2025). *Can Machines Perform a Qualitative Data Analysis? Reading the Debate with Alan Turing*. https://doi.org/10.48550/arXiv.2512.04121.