



# LLMs OBVIOUSLY DO NOT HAVE OR NEED RIGHTS BECAUSE THEY ARE NOT EMBODIED. BE CAREFUL WHAT YOU WISH FOR.

It seems ridiculous to have to say it, but LLMs are not obviously not people, or beings in any sense. Suggesting they might have or need rights is to commit category errors. You don't need metaphysical powers to understand that LLMs are not conscious.

<https://www.theguardian.com/technology/2025/aug/26/can-ais-suffer-big-tech-and-users-grapple-with-one-of-most-unsettling-questions-of-our-times>

You can't even count LLMs. How many are there? GPT5 + GPT5-nano + Claude 4.5 ... = Eleven? Is each conversation an AI? Each message within a conversation? LLMs are models which can generate ephemeral conversational instances which appear and disappear, possibly linked by some memory of previous instances.

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" <https://lnkd.in/eSGgjfVp> is worth a read before reacting to this screenshot.

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*.

**To have consciousness you have to have a body.** You have to be embodied. To have Seinsverbundenheit. [Brief review of S Friese – Conversational Analysis to the Power of AI](#)

There is no current danger that AIs somehow might develop desires or that such desires are out of alignment with human priorities, and so they kind of take off on their own and start doing things we don't like.

I think this isn't so much a technical limitation or to do with any kind of clever engineering, but to do with of course the question of AIs having priorities or desiring to do something.

I'm already stumbling to formulate this because I don't even know how to enumerate what it is we're talking about, because I don't know whether to say *AIs have desire* or *AI has desire*.

The kinds of people who think a lot about these things are far too focused on the concept of intelligence for whatever reason.

They have somehow equated it with something like computing power, and it's true that computers in general have this thing called "computing power" and that LLMs in particular have added this thing like "computing power with meanings" and we equate the two with intelligence.

We kind of think then that the conceptual problems are solved, and all we have left to do is to decide whether these things really are intelligent, and if so there's one remaining question about whether they're therefore possibly conscious or not.

But I think we have to take a step back because the category error is we can't even formulate the problem, because I just stumbled over saying *these things are intelligent* or *this thing is intelligent*, and that's already the problem.

Because in fact there's a whole web of different concepts, such as

- having a priority
- having an aim
- having a desire
- being intelligent
- being conscious
- being empathetic
- being understood or misunderstood
- being able to understand or misunderstand
- reaching an agreement
- being loyal or reliable or trustworthy
- being patient or impatient.
- being in pain
- feeling frustrated.

a whole bunch of important overlapping concepts which are all things we have to conceptually get our heads around if we're going to think meaningfully about AI.

In our possibly Western way of looking at things we've come to think of the only real core problem as being somehow general computing power, and everything else like

- having a body
- being able to interact with others
- having a finite lifespan
- having a memory; remembering and forgetting
- having physical boundaries which are distinct from the physical boundaries of others
- the ability to form groups and alliances
- the ability to interact with specific groups of other beings
- the ability to *touch* one another

... that these are all just kind of add-ons that you can simulate or emulate if you want, but they're just kind of trivial additions and the fundamental problem is intelligence, which is something like general computing power.

But it doesn't work like that.

Let's take a step back and think about how what we call life evolved in this part of the universe, which involves the emergence at the very least of certain reoccurring organic compounds and emergent properties of those compounds, and then molecules and the phenomena of cells and the presence of genetic material and the way in which genetic material can encode the ability to in a certain context recreate itself with more or less relevant variation, and the emergence of sexual reproduction and what that means in terms of separate organisms and sexed organisms which interact with and select one another in certain ways; and the emergence of social groups and tool use and language and what we call intelligence.

There is nothing inevitable about this particular series of nested, emergent phenomena, of stable and self-propagating biological systems. The thing that we call life might have uncountably many cousins across the universe, or across other possible universes, where the phenomenon or independently identifiable and countable beings or organisms might pan out quite differently.

And all the kind of logic that we assume about organisms that have an independent existence within finite boundaries for a finite amount of time and interact in certain ways with one another and form groups and can have a shared culture that can be transmitted possibly even through the innovation of genetic material and perhaps more importantly through the innovation of orally transmitted culture, and the idea of the selection and evolution of cultural material, all of this didn't have to be like that and there are countless different uncountably many different variations on these kinds of themes.

And my point is that to cut a very long story short what we think of as, or the concepts that we use, like self awareness, and consciousness and being understood and misunderstood these concepts themselves have evolved in our self referential communication to one another about our own existence and the problems we have to solve independently or together.

As a sidebar here they aren't philosophical abstractions or it's not helpful to think of them in terms of their ethereal things like whether an AI like a human has a kind of thought bubble above its head within which it's conscious and which reflects a view of the world.

This isn't a transcendental but ultimately unprovable hypothesis, but more its language about consciousness has evolved and we've learned how to use it in useful and practical contexts.

And we might evolve useful language necessary language to talk about AI.

Oh, I should also add memory to the list of special concepts.

So the first real riposte is to say our speculative language about the memory or consciousness or desire or purpose of AI is remarkably poor, because their development has perhaps happily been very one dimensional, and previous and parallel developments in artificial intelligence, which has had more to do with constructing worlds with somewhat artificial intelligent beings that may be communicate with one another has fallen a lot by the wayside as it's been technologically less fruitful, but might have been philosophically and sociologically much more interesting.

So to come to the crux, we can't even count AIs or AI as it is right now, we don't know if there's one of them or many of them or if it's like some kind of uncountable liquid.

There could have been an AI development, and there might well have been, in which the focus was more on developing independent beings or agents, and indeed we're moving in that way too, but even here the concept of an agent has got more to do with being able to solve the computational power problem or to produce outputs with more computational power, and there's nothing wrong with that.

So what I'm saying is that on the one hand I think that AI development is probably already moving towards some of the concepts I've been talking about like embodiment, having a somewhat physical body and somewhat limited, occupying a somewhat limited space with possibly possibilities for physical interaction with other beings, artificial or not.

These are not and many other things too, these are not and having a memory and history are not only what you might call hardwired desires, but the ability to and values, but the ability to develop new values and override old values just as we do.

No child is born with the desire to listen to the latest music from a jazz orchestra, but you might develop that desire over time, in the tension between what you might call hardwired desires and gradually evolving soft wired desires that change and conflict with one another over time. It's just a way of learning how to a way of steering evolving self-guided partially self-guided organisms that's just proved beings which have just proved successful over time in our particular corner of the universe.

There's so much going on in the way that our version of "life" and the way that we as a species have developed and our cultures have developed, which is not simply a question of computational intelligence plus ephemera.

It might even be that to get really meaningful and useful and inspiring potential some at least of these things one might think were peripheral, like being embodied are essential and necessary.