Human's "je ne sais quoi" vs. AI: Theory of Mind

Simon Sure, 17 June 2026

After reading this, if you think that I am not convinced we will reach artificial general intelligence or that I am sceptic of the potential of technology because "humans are better", you have fundamentally misunderstood

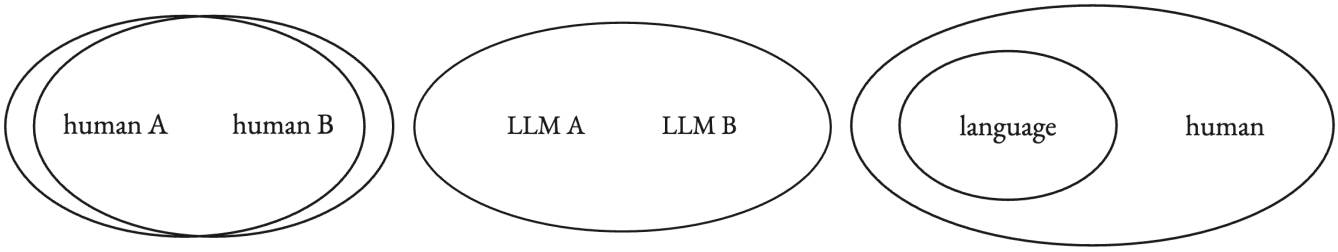
There are various things AI doesn't do as I want it to, while humans do. I think there is a structural reason for this "je ne sais quoi" of humans that AI can't seem to reach. This theory also explains how artificial general intelligence is a purely relative concept and why the current technological approach is not suited for something we will consider ourselves equal

I don't think the "je sais quoi" is a lack of intelligence or a lack of ability to solve problems. It's in intention and perspective. During one of my internships I had the experience of (travel) assistance that felt like magic: Saying where and when I need to go, you get the exact tickets you would have chosen yourself shortly after. The fancy new AI tool at another place didn't even come close. It made to me obviously stupid suggestions and required much back-and-forward so that I could have done it myself in the first place

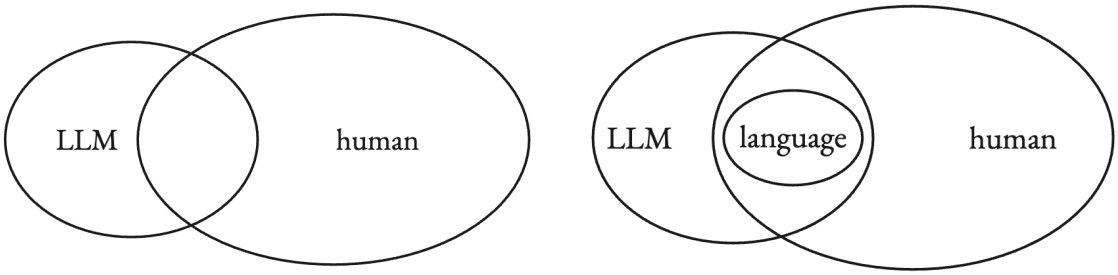
The human understood my situation and intent. The LLM-based tool understood what I had specified in my query but nothing more. So it delivered a technically correct result that still didn't meet my expectations. The human could do the task as desired despite the same imprecise specification and being a complete stranger to me

Attributing this to flawed communication on my part is unfair. I couldn't have preemptively specified all possible scenarios. While the LLM will get closer to the human, I argue that with current technology it will never reach the human's level of satisfaction because of a non-inclusive theory of mind. Here is what I mean:

Two humans have a near perfect theory of mind. Some notions of intelligence consider theory of mind as a key part of intelligence. This makes sense: In all practical terms, our brains are mostly identical. What I can think, you can think. We can simulate each other's thoughts and minds by actually running them in our brains. You can use the same concept to argue more subtle differences between humans.
Two instances of the same LLM have perfect theory of mind. All the detailed differences that exist between humans are absent. The two LLMs can exactly represent each other's thoughts and minds



Humans and language have a hierarchical theory of mind. Everything humans can express in language is a subset of what humans can think. Language is a lossy medium of thought. Different languages conform to different theory of mind subsets of the complete human theory of mind

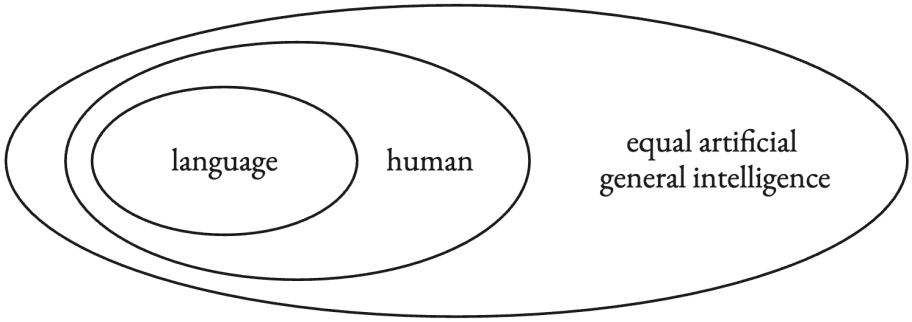


The reason for the “je ne sais quoi” is that large language models (LLMs) and humans only partially overlap in their theory of mind. Some things that can be thought by humans can be thought by LLMs, while other things can’t be thought by LLMs. And it is reasonable to assume there are things that can be thought by LLMs but not by humans

Everything that extends beyond the realm of language is some “je ne sais quoi”. Some expectation and standard we think but cannot communicate in language (which may be extended to include body language and subtext). Another human with near perfect theory of mind can reconstruct this “je ne sais quoi” and react appropriately. An LLM with only partial overlap is structurally incapable of reconstructing the aspects of human’s thoughts that lie outside of its realm

Assuming we push LLMs to their limits, they will fully encapsulate language. They may become capable to the same magnitude as humans or an even larger one, but they will still remain distinct. The current technology limits what kind of information they represent. In the biological and technological context, a thought never exists independent from its medium. A thought that can be represented in a human relies on all the intricacies of human biology, most of which is not modeled by current technology

The feeling of disconnect and lack of understanding, the "je ne sais quoi", will not be solved by scaling. We will increase the size of the LLM's oval in the diagrams, but (using current transformer technology) the LLM will keep having only a partial overlap with humans.[1]



To achieve an artificial general intelligence that is not only highly intelligent and capable but also feels like an equal, we need to extend the underlying technology in a way that fully represents human thought. Only then will the "je ne sais quoi" disappear and we can trust the LLM to understand (and hopefully do) what we ask

[1] LLMs will not seem able to fully understand us. Meanwhile, they will be able to think and represent a lot of things outside of human's realm. That is an opportunity for language to extend and hopefully also enlarge our realm by pushing against human's realm's boundary