

Many have put the right diagnosis on why debates on the consciousness of LLMs, on their intelligence or the individuality of their agents, are leading nowhere: we do not yet know what these properties really mean *for a LLM*. We do have plenty of definitions, but most seem to be contextual and not able to capture what happens with this new artifact called an LLM. My point is that this is a bigger problem of how we use natural and scientific language, and its limits in the face of a new phenomenon, and not strictly one of definitions. Two decades of effort put into this attempt to better talk about AI failed for a more familiar reason than one may suspect at first, one that Bacon analyzed at the dawn of early modern science in *Novum Organum*: we are forced to use words from natural language that describe phenomena in a subjective way. Bacon saw that a new language is needed, one in which the words are anchored in objective experiments, and not in subjective perception. He failed to apply his project to “heat”, mainly because you cannot build a new language word by word. But his lesson stood, taken up as the constitution of the Royal Society and built into the practice out of which scientific language grew: you cannot fix scientific terms by argument. It happened again when QM emerged and natural language spilled back in because nothing else was available; the atom was “a planetary system”, and later Einstein and Schrödinger found it intolerable that “a particle has no position”. The situation today in AI is eerily similar.

Let us start with a clear example. Is a cat conscious? No doubt, no matter what definition or set of criteria we employ. But why is it then that most current thinking on machine consciousness assumes that the right kind of language processing would itself amount to consciousness? Many conscious beings may not ever talk, and yet, we consider coherent talk as a mark of consciousness. Is it? Why?

The problem traces back to Block’s split: phenomenal and access consciousness. “What is like to be something?” vs. “a state available for reasoning, for speech, and for rational control of action”. The distinction was drawn carefully, by argument alone *and not by phenomenological evidence*. It is *access consciousness* that is invoked in LLMs’ case, because it appears that there is *something* that is accessed to produce coherent speech. This has its own problems, because as many have observed, this is our intentional stance at work: we do not really understand why they are so coherent and effective, and we basically project on them a reason so that they would make sense to us.

There is a good paper to be written on this, but I won’t get into it. I’ll just observe a strange fact: almost nobody says “LLMs have access consciousness” instead of “LLMs are conscious”. There are reasons for that, the term is not made for the phenomenon we are dealing with, so that it is stretched and bent, up to the point of becoming a new problem rather than a solution.

One thing people involved with this do to stabilize “consciousness” is to apply the Socratic method: they argue at length about its meaning, refining it through examples and counterexamples, failing to see that whatever gets out of this is still the vocabulary they tried to fix. It is a circularity that one who has not

read *Novum Organum* has great difficulty becoming aware of; it is precisely like redefining “heat” through the subjective experiences natural language has attached to it: it burns, it rises, it bites. Before Bacon this yielded heat as a real quality inhering in substances, something a body possesses more or less of, far from what physics understands by “heat” today, energy transferred between systems in thermal contact. In *Novum Organum* Bacon basically failed to redefine “heat”, but his “first vintage” (something like a first iteration) sounds remarkably similar to our use today: “Heat is a motion, expansive, restrained, and acting in its strife upon the smaller particles of bodies.”

For a Baconian then, there is not much surprise in why LLMs end up as some sort of conscious creatures; it is to be expected. But more importantly, this indicates that his “Tables of Discovery” may point a way out of the circularity.