

## Confessions of a Chinese Room

In the recent past, a mysterious word has been rustling through the pages of the international press and clogging the channels of global digital communication: *chatGPT*. Oh mon dieu, what was there to read and hear. Already the all-simplifying and sensation-producing crowd of journalists saw once again the uniqueness of the human species threatened. Finally there seemed to be a superintelligence that could answer all more or less earth-shattering questions.

However, similar speculations seemed to have existed many decades ago, when the IBM company was still building *room-filling computer systems* that had less capacity than today's pocket telephones. Rumour had it that such a machine knew everything. And when, during a tour of the computer centre, one of the group got the idea to ask the computer about the whereabouts of his late father, he learned that his father was playing golf in Connecticut. That can't be, our man said, "my father died last year." To which the computer output, "Your mother's husband died last year, but your father is playing golf in Connecticut".

No, the IBM story is of course a nice phantasy but even the still room-filling yet orders of magnitude more powerful computer-systems that run this mysterious chatGPT programme would be at their wits end on such a question.

Just for the fun of it I put it to the test and asked chatGPT the same question. Here is the answer: *„I'm just a text-based AI and I don't have access to personal information about individuals unless it has been shared with me in the course of our conversation. I can't help you locate your father. If you're looking for someone, it's best to contact local authorities or use appropriate channels and resources to help with your search.“*

A reply to remember.

So it appears something in that big but invisible machine I was communicating with did understand my question and knew it could not provide a satisfactory answer. But it did not fail to come up with some reasonable advice as to how I might obtain the needed or desired information. And politely so, just like a helpful person might have done. It took note of my request, identified key words –like „where“, „my“ and „father“ –and set to work, generating pertinent text.

There sure must be a ghost in the machine who does the understanding, matching the input he receives with what he knows and builds on that. But does he really „understand“? In the same way we humans think we do (disregarding the fact that we are mostly unaware of how we do it and what it means to us)? Is that ghost „human-like“<sup>1</sup>?

To answer that question I decided to ask the ghost himself.

*„Do you understand natural language the way ordinary people do?“*

Here is his answer:

*„I’m designed to understand and generate natural language text, and I strive to comprehend text in a way that’s similar to how humans do. However, it’s important to note that I don’t possess consciousness, emotions, or subjective experiences like humans. My understanding is based on patterns and information present in the text data I was trained on, and I generate responses based on that data. My responses are generated algorithmically, and I don’t have personal experiences, opinions, or feelings. My goal is to provide helpful and coherent information in response to the text input I receive.“*

Hence, no love, no pity, no sorrow and no joy. The ghost knows the words and he knows the contexts in which they may occur but he will never understand what they mean to us.

---

<sup>1</sup>Exercise: What does "human-like" mean?

There we have it! Sounds familiar?

In fact, some forty years ago in faraway California, a then quite young master of the mind game, [John Searle](#) by name, dreamt up such a ghost, without suspecting that I would be "talking" to such a thing today.

He put the ghost in a room and gave him the task of answering questions that were passed to him through a slit in the locked door. He was to return his written answers - through a second slot in the door - to the questioner.

The ghost had only one problem: he did not know the language in which the questions were asked and the answers were to be delivered. It was Chinese!

Well, it could just as well have been Greek or Spanish or Turkish, all languages of which the ghost had no idea either (in fact, he did not need to know any language but English). But it just so happened that a wonderful Chinese dictionary was at hand, which not only contained all the characters in all possible combinations, but also the rules (in English!) according to which the characters had to be put together to make sense according to the questions asked.

Upon: 我父亲在哪里? (“Where is my father?”) our ghost – based on his text corpus and syntax rules (now incorporated in so called “[Large Language Models](#)”) would have replied:

我很抱歉，但我无法回答这个问题，  
因为我没有任何关于您父亲的信息。  
如果您正在寻找您父亲的具体位置或联系信息，  
您可能需要与他进行直接交流，  
或者与其他家庭成员或朋友联系以获取更多信息。  
如果您有其他问题或需要帮助，  
我将尽力提供支持。

(Very polite! Try DeepL!)

Not surprisingly, Searle’s scenario became known as the "[Chinese Room](#)". His ghost was actually a person working with

all sorts of office materials but of course, what he really meant was a computer with its sophisticated circuitry and switches. It served Searle as an argument that computers cannot understand anything, thus refuting an assumption that was held by many AI enthusiasts at the time and is again or still quite popular today. (In fairness, I should add that Searle's argument had its antecedents, the earliest - probably - being [Leibniz's mill](#)".)

But now we have it first hand, from a "real" *Chinese Room*, called *chatGPT*, defying the numerous counter-arguments Searle's argument provoked. Of course, *chatGPT* (the *Chinese Room*) does understand something but it is not what natural human entities understand. *chatGPT*'s world is not the real natural world but a man-made artificial world of all sorts of discrete symbols, derived at best, from human experiences, insights and interactions, a second hand world. Its semantic ground, its ontology, as it were, is purely syntactic ("what goes with what" - must be constructed from whatever text it was trained on, according to rigid or - at best - statistical rules).

Case closed?

Well, what if we allowed the ghost in the *Chinese Room* (or *chatGPT*) to make his / its own experience of whatever world -or environment -it is supposed to work in or render services?

I was wondering if the corpus of text *chatGPT* had been trained on would come up with an answer to the following two-part question:

*Are personal experiences within a given environment necessary in order to achieve a true understanding of that environment? And in what terms does understanding an environment become manifest?*

The answer, I regret to say, to the first part was somewhat discombobulated, a bit of a waffle. With the second part it was more successful. In summary, it said:

*The manifestation of understanding can be seen through effective problem-solving, communication, adaptation, and the development of empathy and connections within that environment.*

Probing a bit deeper I asked:

*Can this kind of understanding be achieved by a suitably programmed information processing device?*

Again, it basically confirmed Searle's conclusion that computers are incapable of human-like understanding:

*Suitably programmed information processing devices, like advanced artificial intelligence systems, can achieve a form of understanding of environments and information, but it differs significantly from human understanding. ... Their understanding is based on algorithms and patterns, and they lack consciousness, intuition, empathy, and the ability to transfer understanding between domains. As a result, human understanding remains distinct and more nuanced compared to machine-based understanding.*

Suggesting the use of *machine learning* algorithms to achieve a better understanding of an environment did not make the *chatGPT* ghost change its "mind":

*Suitably designed machine learning algorithms can lead to systems that exhibit a form of understanding within specific domains, but this understanding is rooted in statistical patterns and lacks the depth, breadth, and versatility of human understanding. While these algorithms can be powerful tools for various tasks, they are fundamentally different from human cognition and consciousness.*

So, there we have it again. But remember that everything our *Chinese Room* (brand *chatGPT*) "knows" is based on what symbolic representation of human knowledge it has been filled with. And all knowledge may have an expiry date.

However, I was surprised that *chatGPT* was not more positive about the potential of machine learning. It seemed to rein-

force its own claims about the difference between human and machine understanding, demonstrating a certain lack of “imagination”.

Apparently “machine learning” did not trigger a response that would also address the growing body of research in *robotics*, where processing sensorial input and learning from it plays a key role, for instance in achieving a certain degree of autonomy of operation.

I therefore posed a more specific question:

*Could a robot with all sorts of sensors and actuators acquire a "human-like" understanding while interacting with its environment, through learning for instance and developing suitable (digital) world models?*

And that did indeed yield a response that I could have given myself:

*It is possible that a robot with advanced sensors and actuators, combined with sophisticated machine learning algorithms, could develop a more human-like understanding of the world through interacting with its environment and learning from its experiences. However, it is important to note that this is still an area of active research and development, and there are many challenges and limitations to achieving a truly human-like understanding.*

*One of the biggest challenges is developing machine learning algorithms that can accurately process and interpret the vast amounts of data that a robot’s sensors and actuators can generate, and then use that data to construct meaningful representations of the world. Additionally, there are many nuances and complexities of human experience that may be difficult to model or replicate in a digital form, such as emotions, intuition, and creativity.*

*Nevertheless, researchers and engineers are actively working on developing robots with more advanced and sophisticated capa-*

*bilities, and it is possible that we may see significant progress in this area in the coming years and decades. Ultimately, however, it is likely that there will always be some fundamental differences between human and machine understanding, due to the inherent differences in our biology, experience, and cognitive processes.*

Amazing. Note that the bottom line is still the same. *Chat-GPT* insists on the ultimately unbridgeable difference between human and machine understanding, not the least due to our biologically determined “*being in the world*”. I fully agree. Mind matter does matter.

We (like all other animals) learn from the very beginning of our lives (birth or conception, I leave that up to you). No switch needs to be flipped. But that’s what has to be done if we want a robot to run. The robot’s “*being in the world*” is still “second hand”. It depends solely on its designer and operator, their good will and skill, creativity and intelligence

Unless we come up with a [completely new life design](#), like God on the [fifth and sixth day of creation](#).

If such bleak phantasies prevail: may God have mercy on us!

Hans-Georg Stork, 11/2023