What lies behind AGI: ethical concerns related to LLMs

Giada Pistilli

What lies behind AGI: ethical concerns related to LLMs

Giada Pistilli[1]

Abstract

This paper opens the philosophical debate around the notion of Artificial General Intelligence (AGI) and its application in Large Language Models (LLMs). Through the lens of moral philosophy, the paper raises questions about these AI systems' capabilities and goals, the treatment of humans behind them, and the risk of perpetuating a monoculture through language.

Key Words: Artificial Intelligence – Natural Language Processing – AI Ethics – Value Theory

Introduction

The confusion around the term "Artificial General Intelligence" (AGI), often trapped and disputed between the marketing and research fields, deserves to be defined. In 1980, American philosopher John Searle published an article in which he argued against what was then called "strong AI". Following the legacy of Alan Turing, the question Searle posed was: "is a machine capable of thinking?" (Searle, 1980). To briefly summarize the experiment, the philosopher illustrated a thought experiment known today as "the Chinese room" to attempt to answer his question. The thought experiment consists of imagining a room in which an artificial intelligence has at its disposal a set of documents (knowledge base) with Chinese sentences in it. A native Chinese speaker enters the room and begins to converse with this AI; the latter can answer, considering it can easily find which sentence corresponds to the questions asked. The American philosopher's argument is simple: although AI can provide answers in Chinese, it has no background knowledge of the language. In other words, the syntax is not a sufficient condition for the determination of semantics.

Although the term "strong AI" is replaced by "AGI" nowadays, the two terms do not mean the same thing. More importantly, there is still a lot of confusion among pioneers and AI practitioners. Machine Learning engineer Shane Legg describes AGI as "AI systems that aim to be quite general, for example, as general as human intelligence" (Legg, 2021). This definition seems to be a philosophical position rather than an engineering argument. In this article, I will not discuss human intelligence, a topic arousing debates for centuries in many social sciences (e.g., epistemology, philosophy of mind, cognitive psychology, anthropology, etc.), but rather AGI capabilities. Therefore, the interpretation I will use to the term "Artificial General Intelligence" points to AI systems as increasingly specialized in precise tasks, specifically in processing natural language[2]. The idea is then to multiply (scale) exponentially the capabilities of a given AI system. I will not discuss the possibility of theoretical physics to realize this idea, but rather its philosophical implications and, specifically, its moral implications.

Natural Language Processing (NLP)

[1] Sorbonne Université & CNRS – giada.pistilli@paris-sorbonne.fr

[2] Language is defined as "natural" when it belongs specifically to humans (e.g. Chinese, Spanish, German), as opposed to "artificial" language of machines (e.g. different code languages).

Before discussing the AGI moral implications, it is essential to situate our arguments and clarify a few technical details. Without going too deep in technical details, I strongly argue that social scientists must know the technical specifics of the AI system we are studying, highlighting the possibilities it offers and the existing potential risks for its users. It is indeed impossible to talk about artificial intelligence as if it were a single monolithic block; it is crucial to keep in mind and compare each technical specifics in order to evaluate them in the social context.

Natural Language Processing (NLP) is a field of research, a subfield of AI, that focuses on the interactions between human and machine language. Initially based on a symbolic recognition system (called symbolic AI), learning in NLP today refers more to statistical probability methods in Neural NLP. NLP systems based on Machine Learning algorithms are increasingly popular, and one type of learning is making waves: the Transformer model. It is an attention-based learning model, a technique that attempts to mimic human cognitive attention. As Wittgenstein would say, a word only makes sense in its context (Wittgenstein, 1953). Similarly, the Transformers models, in the pre-training phase, make connections between words. The principle is to use a very large dataset and focus the attention of the model on a small but important part of it, depending on the context (Vaswani, 2017).

Generative Pre-trained Transformer 3 (GPT-3)

To illustrate our point, we will take the GPT-3 model as a case study. My goal is not to sling mud at OpenAI or any other type of product that exists out there; if I talk about GPT-3 is because I have got access to their API last year, and I have been studying it since. My arguments wish to be a philosophical conceptual basis for thinking about ethical issues related to Large Language Models (LLMs) and asking questions for the future.

GPT-3, Generative Pre-trained Transformer 3, is an autoregressive language model that uses deep learning to produce human-like text (Broackman et al, 2020). OpenAI's API can be applied to virtually any task that involves understanding or generating natural language. On their API webpage, there is a spectrum of models with different power levels suitable for various tasks. Examples of GPT-3 models are: chat (it simulates an AI assistant to converse with), Q&A (where you can ask questions on any topic and get answers), Summarize for a 2nd grader (makes a summary in simple words of a provided text), classification (you write lists and ask for categories to be associated with them), and much more.

The problem of Artificial "General-purpose" Intelligence (AGI)

As seen above, there are several definitions of what an AGI is. Another interesting definition for our analysis is the one proposed by Goertzel and Pennacin in their 2007 book *Artificial General Intelligence*:

> Artificial General Intelligence (AGI) refers to AI research in which 'intelligence' is understood as a general-purpose capability, not restricted to any narrow collection of problems or domains and including the ability to broadly generalize to fundamentally new areas. (Goertzel & Pennachin, 2007)

The various definitions of AGI often recall a cross-cutting capability of the language model, defined as "general-purpose". If we are taking GPT-3 as a case study, is because OpenAI defines its API like following: "unlike most AI systems which are designed for one use-case, the API today provides a general-purpose "text in, text out" interface, allowing users to try it

on virtually any English language task" (Broackman et al, 2020). The simplicity of using this type of AI system is that users can exploit them with almost no computer skills. Users simply have to write their request in natural language in the "prompt". GPT-3 will respond with content generation that attempts to match the answer ("text-out") to the question ("text-in").

Some ethical concerns about "general-purpose" Large Language Models

Developing an AI system, specifically, a Large Language Model with "general-purpose", or better said, without a precise goal but rather a broad spectrum of capabilities, raises various ethical concerns on different levels. I will not explore all ethical concerns, but rather focus on three in particular.

1. The first ethical problem we face is related to the innumerable capabilities of the AI model. In moral philosophy, which deals with defining, suggesting, and evaluating the choices and actions that put individuals in a situation of well-being, it isn't easy to assess an artifact morally with an assortment of different objectives. Moreover, the capacities of a Large Language Model like GPT-3 are often defined but can multiply with its use. Given the breadth of possible uses in natural language, the model's capabilities can be infinite if not defined a priori and framed by its developers. If the goal of an AGI is to no longer recognize itself in a list of skills but rather to have an infinity of them, the situation becomes highly complex to keep under control. It won't be easy to assess and make value judgments about something whose full range of capabilities is still unknown. Also, it will be challenging to control possible malicious uses, to name a few: spam, fake product reviews, fake news, fake social network accounts, homework cheating (e.g. generated essays), etc. In other words, I argue that in order to make a moral judgment about a technological artifact, it is essential to know and define its goals. In the absence of these conditions, ethics will hardly find its usefulness. Calculating the risks, consequences, context, and model use would be very challenging, or even impossible, if its capabilities are infinite.
   Moreover, without going into the psychology and characteristics of human intelligence, there is confusion among AGI pioneers between the latter and Human-Level AI (Goertzel, 2014). Nils Nisson described the AGI as a machine capable of autonomous learning; the question emerging here is: without *a priori* fixed limits, how can control be exercised over its possible and various uses? (Nilsson, 2010) What safeguards are in place to prevent abuse and misuse? Furthermore, what are the limits set on the machine learning of this AGI? Given these technologies' state of the art, the current state of moral analysis around these systems often seems to dwell on the technical limits of the machine or human intelligence. Quid about the boundaries of the latter's capabilities?

2. Secondly, as already pointed out by Goetze and Abramson in their paper "Bigger isn't better" (Goetze & Abramson, 2021), by sociologist Antonio Casilli's studies of "click workers" (Casilli, 2019) and researcher Kate Crawford (Crawford, 2021) there is an ethical concern related to social justice. Crowdwork, often used to train such large models, does not guarantee the quality of the dataset and perpetuates wages inequalities.

   > Crowdworkers are generally extremely poorly paid for their time; ineligible for benefits, overtime pay, and legal or union protections; vulnerable to exploitation by work requesters [...]. Moreover, many crowdworkers end up trapped in this situation due to a lack of jobs in

their geographic area for people with their qualifications, compounded with other effects of poverty. (Goetze & Abramson, 2021)

For example, the famous ImageNet dataset was labeled by an equally renowned crowdwork: Amazon's Mechanical Turk, which offers tailored services to adjust and improve AI systems' data and knowledge bases while training them to enable automation (Crawford, 2021). The way these Large Language Models are trained is a bit obscure and raises issues of social justice and relevance when annotating data that will need to feed a globally targeted AI model. This set of issues raised seems to refer to the logic of what some contemporary philosophers call the "technoeconomy" (Sadin, 2018). According to this logic, the economy would find itself driving technical and technological developments, seeking to minimize their costs to produce maximum benefits.

3. This last argument allows us to make a transition to our third ethical problem: language. Speaking of Natural Language Processing and Large Language Models, it is inevitable to talk about it. I argue that the language-related problem in Large Language Models is of two different natures. The first is the difficulty in controlling the text generation ("text-out") produced by the model. As an example, GPT-3 has a content filter to warn the user when confronted with content that is unsafe (text containing profane, discriminatory, or hateful language) or sensitive (the text could be talking about a sensitive topic, something political, religious, or talking about a protected class such as race or nationality). Unfortunately, this content filter is inaccurate and unsatisfactory, as the content generated by GPT-3 is often toxic. Its creators are aware of this and are studying what might be possible solutions. The "Process for Adapting Language Models to Society (PALMS) with Values-Targeted Datasets" paper (Solaiman & Dennison, 2021) introduced by the OpenAI research team explores possible solutions to the problem of toxic content generated. Their goal is to modify the behavior of the language model through small samples in which there are value-targeted models. Although OpenAI has succeeded in decreasing the toxic content generated by GPT-3, studies such as PALMS highlight other ethical concerns related to the values that may be contained in the language model and, consequently, transmitted to the users who employ it.
   Regardless, it still will be difficult to tame this titan under these AGI conditions of "general-purpose". In this case, the limits are not only ethical but also technical. Being the text generation a probabilistic calculation of which word will follow within the same sentence, GPT-3 will always be in the condition to give different answers from each other, according to the examples inserted in its prompt. Therefore, if the text-in already presents toxicity, finding it in the text-out will be easy. Differently, if in the prompt there are no toxic contents, there will always be the probability that GPT-3 answers with a text-out containing toxic elements. Once again, the ethical problem here is related to the vastness of the language model and the desire to open it up to a multitude of capabilities.
   The second nature of the language-related ethical problem when it comes to Large Language Models is the absence of diversity. Diversity is understood not just as a representation of gender and ethnicity but also as an actual language (Spanish, Portuguese, Danish, etc.). In fact, according to OpenAI, 93% of the training data was in English. The next most represented language was French (1.8%), followed by German (1.5%), Spanish (0.8%), Italian (0.6%), and so on (Brown et al, 2020). Researchers have already begun to explore the multilingual capabilities of GPT-3,

noting for example how it works poorly in minority languages such as Catalan (Armengol-Estapé et al, 2021). Since the absence of a piece of data is as important as its presence, the very scarce presence of languages other than English leads us to some rather negative considerations, given the multilingual and universal nature that an AI model like AGI is intended to take. The overwhelming and cumbersome omnipresence of the English language is a serious problem that needs to be addressed as soon as possible if we want to make AI accessible to everyone. Because GPT-3 is a system that uses natural language to function and provide answers, orienting it exclusively to English and the values that revolve around American culture will not do justice to the pluralism of values in which we live in our diverse societies. The risk of implicitly promoting a monoculture fostered by large American industries is indisputable. The danger here is twofold: on the one hand, the propagation of the monoculture may be permeated by the implicit or explicit values of the industries developing these AI systems. On the other hand, this same monoculture can be promoted and shared, implicitly or explicitly, through the value systems belonging to the culture dominating these new technological developments. One striking example is related to the recent testimony of the Facebook whistleblower. During her testimony, Frances Haugen pointed out that the lack of moderation tools in languages other than English allowed users of the online platform to freely share content in violation of Facebook's internal policy (Hao, 2021).

Potential solutions to be explored

First, a challenging but fundamental question must be asked: what then is the ultimate goal of these Large Language Models? What is the purpose of AGI? Since in ethics the "I do it because I can" paradigm can't stand, we should be able to define "the" purpose clearly and not settle for the vague "general-purpose". In its absence, as said above, it will be difficult to find a justification and, consequently, evaluate it morally. Considering the advances and the current state of the art of machine learning technologies, automating it more and more can only be desirable after well-framed safeguards have been put in place. If this can still be part of building an AGI, developing *ex-ante* well-structured capabilities limits would be necessary.

Secondly, we need to start shedding light on these dark processes behind the AI industry regarding our social justice issue. The "black box" is not only found within the algorithms, but also on the exploitative processes that often bind the poorest part of the world to make us believe that these processes are automated - but they are not. The demand for human labor to produce the datasets needed to run these Large Language Models grows exponentially. As a result, national and international institutions need to start asking questions quickly, in order to bring answers and a clear legislative framework for these new "data labeler-proletarians".

Finally, the issue related to language is, in my opinion, one of the thorniest to deal with. Aside from the concealed hypocrisy found among the AGI pioneers, who sell their products as being "universal", the problem here is structural. Today we're talking about Large Language Models, but I'd like to point out that the entire Internet ecosystem is governed by the English language and an American monoculture that permeates every corner of it. Today we are facing a difficulty that we can turn into a possibility: we can fix this kind of problem in language models and try to integrate the feedback from its users as much as possible. The process will undoubtedly be longer, but it could be the beginning of a fruitful collaboration. In addition, it might help to change the paradigm of AGI and make it rather "narrow AI":

oriented toward specific capabilities and circumscribed to its context. In this way, each context could appropriate its model and make it its own, thus ensuring a plurality of values relevant to its social context.

Conclusion

In conclusion, we have seen how technical problems often go hand in hand with ethical issues. Given the interdisciplinary nature of the scientific domain of artificial intelligence, these ethical problems cannot be solved without the help of engineers. And when I talk about philosophers and engineers working together, it also means that engineers shouldn't make themselves out to be ethicists. Indeed, yet another major problem that plagues the field of AI today is a significant confusion that exists in distinguishing between the expertise of engineers and that of philosophers. To put it another way, the public scene is occupied by a majority of so-called ethicists but are in reality computer scientists. The interest that is increasingly being brought to moral philosophy is fundamental and is a pressing issue. Still, as we have seen today, the problems that arise from these technologies cannot be solved by technical work alone. The role of the social sciences is to come to the rescue and help the discipline. Because if science describes reality, ethics suggests how this reality should be tomorrow.

References

Armengol-Estapé, J., Bonet, O.D., & Melero, M., "On the Multilingual Capabilities of Very Large-Scale English Language Models", in *arXiv*, 2021. abs/2108.13349

Brown, T.B. et al, "Language Models are Few-Shot Learners", in *arXiv*, 2020. arXiv:2005.14165

Casilli, A., *En attentant les robots*, Editions Seuil, Paris, 2019.

Crawford, K., *Atlas of AI*, Yale University Press, Yale, 2021.

Goertzel, B., Pennachin, C., *Artificial General Intelligence*, Springer-Verlag Berlin Heidelberg, Berlin, 2007.

Goertzel, B. "Artificial General Intelligence: Concept, State of the Art, and Future Prospects", in *Journal of Artificial General Intelligence* 5(1), pp. 1-46, 2014.

Goetze, T. S. & Abramson, D., "Bigger Isn't Better: The Ethical and Scientific Vices of Extra-Large Datasets in Language Models" in *WebSci '21 Proceedings of the 13th Annual ACM Web Science Conference (Companion Volume)*, 2021.

Hao, K., "The Facebook whistleblower says its algorithms are dangerous. Here's why", in *MIT Technology Review*, October 5th 2021: https://www.technologyreview.com/2021/10/05/1036519/facebook-whistleblower-frances-haugen-algorithms/

Nilsson, N. J., *Quest for Artificial Intelligence: A History of Ideas and Achievements*, Cambridge University Press, Cambridge, 2010.

Sadin, E., “Le technolibéralisme nous conduit à un ‘avenir régressif’”, in *Hermès, La Revue*, vol. 80, no. 1, 2018, pp. 255-258.

Searle, J., “Minds, brains, and programs. Behavioral and Brain Sciences”, 3(3), 1980, pp. 417-424. doi:10.1017/S0140525X00005756

Solaiman, I, Dennison, C., “Process for Adapting Language Models to Society (PALMS) with Values-Targeted Datasets”, in *OpenAI Research*, 2021.

Vaswani, A. et al., “Attention is All you Need”, in *ArXiv* abs/1706.03762, 2017.

Wittgenstein L., *Philosophische Untersuchungen* (1953), *Philosophical Investigations* [PI], translated by G. E. M. Anscombe, 1953.