Safety issues

We had argued in Cezanne-ai paper on five directions linked with the impact of the AGI component in our open-framework:

1. We used soft AGI solutions with limited impact.
2. A conversation without 2 parties involved is not a conversation, and thus the AGI component is mandatory in Conversational AI.
3. The existing models are riskier than our framework as they cannot control the biases from the datasets used by pre-trained models or the extraction/generation of the output from the web/3rd party’s data environment.
4. AGI is not used on core-inputs.
5. We are using AGI for intuition over the states and policies related to the conversational flow and not as an instrument for answers generation.

Even so, we find it is our responsibility to put in place safety issues procedures related to the implication that an AGI model could have on the users.

**For that reason, everyone that is using Cezanne-ai open-framework to build conversational AI bots must commit in the beta-testing and in roll-out to determine possible reactions of the users that could have negative repercussions on different levels (phycological, social…).**

*Conversational AGI component in the Cezanne-ai project:*

A quick classification of AGI original concepts integrated in our framework:

* provide intuition in some NLU instances to make sure the bot understands what the user is saying.
* machine education[[1]](#footnote-1) - give conversational bot a similar education with humans by using AI algorithms.
* self-generating policies for non-core-inputs/states.
* self-generation model for behavioral/mood instances that do not affect the answers/output.
* use a labyrinth model based on fundamentals used to structure a book, for query purposes.
* enhance NLG on the non-verbal/non-written components to be more adapted to human conversations[[2]](#footnote-2).

*Fundamentals on AGI[[3]](#footnote-3)*

We have two main and different views on AGI. There are people that don’t think in the possibility of an Artificial General Intelligence or an Artificial Narrow Intelligence and other people that are concerned that we don’t have enough safety measures in place to tackle the negative impact of AGI, if implemented. Both of these views have a negative impact on developing at least a human-level AGI that are also known as weak AI or soft AGI.

In some AI branches we understand the risk associated with AGI implementation, but fundamentally, we cannot have a conversational AI bot without Human-level AI/AGI. Why? Conversation presumes the existence of two agents. If you have on one side a human and on the other side an alien, the probability for the conversation to be incomprehensible is very high, especially in a society where informal communication and empathy prevails. Now, using neural networks can be seen as an AGI concept if you have the objective to simulate human thinking, but the intuitions behind this model, that is math based, is not quite an AGI material. This is the reason why we don’t see an interesting dynamic in NLP/DL as we see in other AI domains and therefore it is mandatory to have a hybrid model for the conversational bot. It is, however, a big question which type of AGI is proper to use inside the Human-level AI that anyway is considered to be soft.

1. Not to confuse machine education with the processes of language models and transfer learning, even if the pe-trained models are fine-tuned to cover the individualization objective. The main difference is similar with the one between knowledge and memorization, or between the concepts of short-memory and long memory (from neuroscience) [↑](#footnote-ref-1)
2. We will consider visual NLU integration as well in the future. [↑](#footnote-ref-2)
3. See also: Ragnar Fjelland: Why artificial intelligence will not be realized: <https://www.nature.com/articles/s41599-020-0494-4> [↑](#footnote-ref-3)