The value of a rationalist approach in AI

Antonio Norelli*°, Luca Moschella*°, Simone Melzi*, Giorgio Mariani*, Marco Fumero*, Arianna Rampini*, Michele Mancusi*, Luca Cosmo*, Emanuele Rodolà*

*Sapienza University of Rome, Computer Science department

Mail: surname@di.uniroma1.it

"Now I'm going to discuss how we would look for a new law.
In general we look for a new law by the following process.
First we guess it.
[Audience laugh]
Don't laugh, that's really true!"
Richard Feynman, lecture on the Scientific Method (1964)

In this work we discuss the practical value in AI research of a set of considerations from epistemology and modern linguistics.

We note how deep learning is aligned with an empiricist view on the acquisition of knowledge, which emphasizes data and conceives the learner as a tabula rasa. Such condition of data as the sole source of knowledge makes our current models extremely data-inefficient, over-fragile versus adversarially chosen samples, and incapable of generalizing out of their training distribution.

Instead, we advocate the adoption of a *rationalist* theory of knowledge for a new generation of AI models. Here data matters only as theory-laden observations, and knowledge acquisition begins with conjectures generated by a creative act. Such conjectures should then be criticized to select the best hypothesis; this is the new role of data along with other principles such as Occam's razor.

In this setting we distinguish two processes that should be implemented in an AI agent. A generative model which formulates new conjectures explaining a phenomenon, and a criticizer to discard most of them.

But how can we create a machine capable of formulating conjectures?

Crucially, we note how a new conjecture cannot but be expressed in a language, through a novel proposition that is *already* valid in that language; that is why we are able to understand it.

This new proposition is an original combination of already known concepts through the rules defining the grammar of our language. It may sound strange, even if grammatically correct, because it can express something never seen, e.g. "the blue elephant rides a unicycle on the moon".

Such a high level representation of the world through propositions is much more powerful than the traditional embeddings. The generalization power of a continuous latent space is given by the interpolation between data points, for instance the smooth transformation of a 2 in an 8 in MNIST; new elements do not differ much from the ones in the training set, and in fact we need huge datasets to be dense enough in a latent space. Conversely, with a language representation we can formulate countless propositions from a few concepts with rules to combine them (combinatorial generalization). Such propositions may describe something very different from the samples in our training set; in our elephantine example, we need to have seen only elephants, blue things, things on the moon and rideable unicycles.

This bears the question as to how can we obtain such mastering of a language in a machine, where concepts and rules are related to a physical world.

Humans acquire language through genuine learning. Children learn their first language mostly in a passive way, without receiving direct instructions.

Nevertheless, children do it also with very scarce data and without having access to negative evidence, so it seems clear that this learning cannot start tabula rasa. Following the Universal Grammar theory, we expect to master a language through a learning process with little data and strong priors, setting up a model that should encode an artificial version of the human Language Acquisition Device. This is in contrast with modern language models, which set up an unsupervised training problem upon huge text corpora and fail in grounding language to experience.

As a further contribution, we present a convenient environment to develop AI algorithms in this setting, based on the challenge posed by the game Zendo. Succeeding in this game requires explicitly to formulate conjectures to guess a secret law of nature of a small world. Importantly we know in advance that every relevant conjecture can be expressed through a tiny language, which makes Zendo an ideal environment to experiment on.

°Equal contribution