ACADEMIA EUROPAEA

The Academy of Europe

FROM KNOWLEDGE TO ACTION APPLICATION

Title: Improving Machine Self-Assessment by Algorithmic Turing Learning at the LLM Frontier

MAE/YAE Investigator(s):
Imre Lendák, PhD, email: lendak@inf.elte.hu
Orsolya Vásárhelyi, PhD; email: orsolya.vasarhelyi@uni-corvinus.hu

Other investigators: Zsófia Hajnal; email: zsofia.hajnal@stud.uni-corvinus.hu

Category: From Science to Society

Problem (max. 50 words):
The phenomenon of benchmark saturation (forms of artificial intelligence surpassing the performance standards humans can establish through tests) means that it is no longer only up to ourselves to safeguard against deception by AI.

Unmet Need (max. 50 words):
Beyond watermarking, one promising primary tool against deception through AI (scams, synthetic content, fake bots, etc.) may be the given AI model itself, provided that it is better at self-detection than deception. To fulfil this latter condition ought to be a cybersecurity imperative, but the methods are not straightforward.

Project Description (max. 200 words):
The project has a multifold exploratory and translational goal. One is the mapping and testing of unique challenges to large language models along the threads of computing, logic, and geometry. Our respective starting points are the halting problem (as proposed by Alan Turing), the application of the liar paradox (as proposed by the Ancient Greek philosopher Epimenides), and operations on non-orientable surfaces, such as the Klein bottle and the Möbius strip. These examples contain elements with the potential to perplex LLMs. By being able to describe the characteristics of the challenges at the frontier of LLM capabilities in dynamic terms, we aspire to generalise the method of proposing such challenges, essentially building a Turing learning (a machine learning method for two competing systems) based algorithm for AI to be able to self-assess and challenge itself.

Hypothesis (25 words):
There exists an algorithm for challenges at the frontier of LLM capabilities, which improves LLMs' self-assessment abilities through Turing learning.

Implication for Practice (50 words):
Finding and describing an LLM frontier Turing learning algorithm would mean an enormous cybersecurity leap, implying increases in trust, productivity, and efficiency in both business and state domains.

Implication for Research (50 words):
Knowledge science can benefit from the lessons to be learned throughout the endeavour to find and describe an LLM frontier Turing learning algorithm, in terms of both insights and opportunities.