# AI Risk and Reasoning in Neurosymbolic AI

Artur d'Avila Garcez[1]

[1]City, University of London, UK

## Abstract

Current advances in Artificial Intelligence (AI) and Machine Learning (ML) have achieved unprecedented impact across research communities and industry. Nevertheless, serious concerns around trust, safety, interpretability and accountability in AI were raised by influential thinkers. Many identified the need for well-founded knowledge representation and reasoning to be integrated with ML systems. Neurosymbolic AI has been an active area of research for many years seeking to do just that, bringing together robust learning in neural networks with reasoning and explainability via symbolic representation and description. In this talk I will review the research in neurosymbolic AI and computation, and discuss how it can help shed light into the increasingly prominent role of safety, trust, interpretability and accountability in AI. AI has become the focus of large-scale research endeavours and has changed businesses. This led to an important debate about the impact of AI on education and society. It has been argued that the building of a rich AI system, semantically sound, explainable and ultimately trustworthy, will require a sound reasoning layer in combination with deep learning. Parallels have been drawn between Daniel Kahneman's research on human reasoning and decision making and so-called AI systems 1 and 2. I will revisit early theoretical results of fundamental relevance to shaping the latest research, such as the proof that recurrent neural networks compute the semantics of logic programming. I will also seek to identify bottlenecks and the most promising technical directions for the sound representation of learning and reasoning in neural networks. I will conclude by discussing the key ingredients for sustainable AI going forward, identifying directions and challenges for the next decade of research in the field.

*CILC 2024: 39th Italian Conference on Computational Logic, June 26-28, 2024, Rome, Italy*

a.garcez@city.ac.uk (A. d'Avila Garcez)
0000-0001-7375-9518 (A. d'Avila Garcez)

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073