Knowledge Representation for Automated Scientific Discovery

Larisa N. Soldatova[1], Gabriel K. Reder[2], Alexander H. Gower[2], Filip Kronstrom[2], Rushikesh Halle[3], Vinay Mahamuni[3], Amit Patel[3], Harshal Hayatnagarkar[3], Ross D. King[2,4,5]

[1] Department of Computing, Goldsmiths, University of London, London, United Kingdom,
[2] The Department of Computer Science and Engineering, Chalmers University of Technology, Gothenburg, Sweden, [3] ThoughtWorks Technologies, Pune, India, Department of Computing,
[4]Department of Chemical Engineering and Biotechnology, University of Cambridge, United Kingdom,
[5]Alan Turing Institute, London, United Kingdom.

The more an AI system 'knows' – the more intelligent and capable it can be.

AI Scientists require knowledge to be represented in a machine processable form: mathematical equations, First Order Logic statements, knowledge graphs - to name but a few. Whilst the research community has made impressive progress in recording data and knowledge in-forms suitable for computation [1,3], many pressing challenges remains:

1. **The area of automated scientific discovery requires a new generation of data and knowledge resources.**

AI Scientists rely on processing relevant domain data and knowledge [4]. However, most existing databases and knowledge models are primarily designed for human users, not AI agents. Human researchers may employ a variety of tools, e.g. database queries, information retrieval scripts, search engines - nevertheless, they are primary users.

In AI-based discovery systems, the main data and knowledge users are software agents. They query data and knowledge with higher intensity, they depend on stable and reliable data and knowledge sources, and flawless logic in data schemas to support reasoning.

2. **'Scientific knowledge is probabilistic and non-monotonic, and a representation system shall be able to reflect this reality'** [5].

Most databases and knowledge models are designed to capture facts. For example, the ontology world has been dominated by the rationalist view where only concepts that have correspondence to entities in the real world could be represented. The world of scientific discovery is not like that, it is populated by hypothetical entities, connections, and plausible events. Data and knowledge resources for AI Scientists need to reflect that.

3. **Continuous dynamic updating of data and knowledge resources**.

AI discovery systems and laboratory automation are exacerbating the big data challenges, driving the ever-expanding production of data and pushing the boundaries of data and knowledge representation. Keeping data and knowledge resources up to date is a formidable challenge. Reliable closed-loop AI systems require dynamic maintenance of its background knowledge [2]. The underpinning knowledge model should be updated every time a new data point, a new hypothesis, or a new model become available.

4. Modelling of AI discovery systems.

The research community benefits from many relevant and well-developed standards, e.g. for the representation of biomedical models [3], for biomedical investigations [1]. However, there are no standards for the representation, recording and sharing information for the whole scientific discovery process: what components are essential, how they are connected, what information need to pass between them.

AI Scientist is likely to be a multi-agent system. Data and knowledge exchange between its components and their synchronisation are pivotal for functioning a complex system. A holistic representation of automated scientific discoveries will contribute to a better understanding the nature of scientific discoveries.

5. Re-usable data and knowledge components for Scientific Discovery.

It is rather an exception, if a scientific discovery system can employ a readily available high-quality data and knowledge models.

Thousands of knowledge models, ontologies, databases are in existence, covering numerous scientific domains. Unfortunately, they are often fragmented and inconsistent. Often, the developers of AI scientific discovery systems have to modify, extend, combine existing sources, or develop data and knowledge models from scratch. This slows down the progress of the area and does not promote consistency of data and knowledge representations.

6. Representation of AI systems and its discoveries in a form humankind can comprehend.

Accurate and comprehensive recording of all the necessary information about AI Scientists supports the transparency, interpretability, openness, re-use, and trustworthiness of AI systems. Knowledge and data about the automated discovery process can speed up the development of AI Scientists and enhance their performance.

References:

[1] Bandrowski A., Brinkman R, Brochhausen M, et al. (2016) The ontology for biomedical investigations. PloS ONE 11/4 e0154556.

[2] Coutant A., Roper K., Trejo-Banos D., et al. (2019) Closed-loop cycles of experiment design, execution, and learning accelerate systems biology model development in yeast. PNAS 116 (36) 18142-18147.

[3] Juty, N. Le Novère, N., Waltemath, D., et al. (2010) Ontologies for use in Systems Biology: SBO, KiSAO and TEDDY. Nat Prec. doi.org/10.1038/npre.2010.5122.1

[4] King, R. D. Rowland, J., Oliver, S.G., et al. (2009) The automation of science. Science 324, 85–89.

[5] Kitano, H. Nobel Turing Challenge. NPJ Syst Biol Appl 7, 29 (2021).