Old AI Meets New AI in the Logic of Scientific Discovery

Ioannis Votsis (Northeastern University London)
ioannis.votsis@nulondon.ac.uk / www.votsis.org

Scientific discovery is, for the most part, a neglected topic in the philosophy of science (Langley and Arvay 2019). Since around the middle of the last century, the received view has been that discovery is not governed by logic or, more generally, by rationality, but is a largely elusive and inscrutable process (Popper [1935]1959). Thankfully, not every philosopher has been persuaded by this pessimistic view (Nersessian 2010). Given the recent cascade of developments in neural nets, this means that now more than ever we need to carefully re-evaluate our attitude towards this view. This talk aims to do precisely that by exploring how such developments affect, and how they ought to affect, the debate over the nature of scientific discovery. It will be argued that, their raw potential to make significant contributions to science notwithstanding, neural net techniques are unlikely to single-handedly reinstate the rationalist model of scientific discovery or indeed lead to mass automation. Rather, a more promising approach on both counts involves the combination of 'old AI' methods like automated theorem proving with neural nets.

The talk is structured as follows: A brief introduction to the distinction between old and new AI is followed by an overview of each tradition's strengths and weaknesses. The subject of hybrid approaches to AI is then broached, and several different variants are identified. A proposal is made for such an approach that extracts symbolic representations from neural nets and other sources and processes them using an automated theorem prover. Part and parcel of this proposal is the logical treatment of the evolution of scientific theories in terms of the addition and/or deletion of content. Two useful heuristic constraints are then discussed, namely structural correspondence and multiple testing ground consilience. The talk ends with a plea for a big, era-defining, computational scientific discovery project that integrates existing efforts into an efficient tool with wide applicability.

The main motivation for the proposed marriage between bottom-up data-driven methods like neural nets and top-down logic-driven methods like automated theorem provers has to do with their complementary strengths and weaknesses. The main strength of neural net methods is that the models they produce are highly sensitive and adaptive to the data. Their main weakness is that those models are often very difficult to interpret. By contrast, the main strength of automated theorem proving methods is they are fairly easy to interpret and even translate into natural language. Their main weakness (particularly when they involve monotonic systems) is that they are rather rigid, lacking or being deficient in the ability to adapt to new data without human input. A suitably configured amalgamation of these two kinds of methods raises the prospects of a powerful tool in the quest to both understand scientific discovery but also automate it.

Unsurprisingly, there is an increasing trend towards hybrid, also known as 'neuro-symbolic', approaches to AI. The rationale behind these is to integrate "the two most fundamental aspects of intelligent cognitive behavior: the ability to learn from experience, and the ability to reason from what has been learned" (Valiant 2003: 97). Neuro-symbolic systems come in several different guises (Kautz 2020). Some put neural computation in the driver's seat, while others bestow that honour to symbolic computation. A proposal that opts for the latter is suggested in the talk. On this proposal, models generated via neural nets are converted into symbolic representations which are then added to, and may even lead to the deletion of, content in an existing symbolic knowledge base. The modified knowledge base can then be used to derive novel hypotheses and/or expected measurements. The latter can be checked against existing measurements or, if none are present, prompt scientific experiments to produce them. Such measurements can in turn be fed into the knowledge base as well as potentially guide the construction of new neural net models. To make the proposal more concrete, some relevant work on how to extract symbolic representations from neural nets and other sources is briefly discussed (e.g. Čyras et al. 2021).

One aspect of the aforesaid proposal worth closer inspection is the addition and/or deletion of content to the knowledge base. Two quasi-logical notions are employed to perform this function: content weakening and content strengthening (Votsis forthcoming). The former involves the removal, while the latter involves the addition of content. Replacement of content is achieved by the successive operations of deletion and addition. This treatment thus shares much with operations in belief revision theory (Alchourrón, Gärdenfors & Makinson 1985; Rose & Langley 1986) but with a restricted notion of logical consequence that excludes tautological, redundant and/or irrelevant content. To illustrate the usefulness of the two notions, a historical case of theory change is reconstructed and briefly discussed.

The question then remains how to constraint the creation of content (equivalently: how to reduce the search space) to begin with. Besides the usual heuristic constraints (e.g. simplicity and hill-climbing), we propose two others: structural correspondence and multiple testing ground consilience. The structural correspondence constraint (Poincaré 1905; Russell 1927) suggests that any new theory about a domain must structurally correspond (at least in some limit form) to the well-confirmed parts of a predecessor theory about the same domain. Several successor theory pairs in the history of science seem to satisfy this constraint (Votsis & Schurz 2012; Schurz & Votsis 2014). The multiple testing ground consilience constraint suggests that to increase our confidence in which content to add and/or delete, we must consider the role they play in deriving empirical consequences in a variety of testing grounds. Whenever there is consilience that their role is overall negative/positive, the corresponding content is then removed/added.

The talk ends with a plea for a big science project on computational scientific discovery that provides economic and other efficiencies. The aim is to consolidate existing efforts into an off the rack tool that can be easily adapted for use in different domains. Such a tool should be made freely available to assist scientists all over the world in their discovery efforts.

References:

Alchourrón, C. E., Gärdenfors, P., & Makinson, D. (1985). On the logic of theory change: Partial meet contraction and revision functions. *The journal of symbolic logic*, 50(2), 510-530.

Čyras, K., Rago, A., Albini, E., Baroni, P., & Toni, F. (2021). Argumentative XAI: a survey. arXiv preprint arXiv:2105.11266.

Kautz, H. (2020). The Third AI Summer. AAAI 2020 Robert S. Engelmore Memorial Award Lecture.

Langley, P. and Arvay, A. (2019) 'Scientific Discovery, Process Models, and the Social Sciences', In *Scientific discovery in the social sciences* (pp. 173-190). Springer, Cham.

Nersessian, N. (2010). *Creating Scientific Concepts*. Cambridge, MA: MIT Press.

Poincaré, H. ([1905] 1952). *Science and hypothesis*. New York: Dover.

Popper, K. ([1935]1959). *The Logic of Scientific Discovery.* New York: Basic Books.

Rose, D., & Langley, P. (1986). Chemical discovery as belief revision. *Machine Learning*, 1, 423-452.

Russell, B. (1927). *The Analysis of Matter*. London: George Allen & Unwin.

Schurz, G., & Votsis, I. (2014). Reconstructing scientific theory change by means of frames. in T. Gamerschlag et al. (eds.), *Frames and Concept Types, Studies in Linguistics and Philosophy*, vol. 94, 93-109.

Valiant, L. G. (2003). Three problems in computer science. *Journal of the ACM* (JACM), 50(1), 96-99.

Votsis, I., & Schurz, G. (2012). A frame-theoretic analysis of two rival conceptions of heat. *Studies in History and Philosophy of Science Part A*, 43(1), 105-114.

Votsis, I. (forthcoming). Theory Change through a Logical Lens. in M. Martinez (ed.), *From Contradiction to Defectiveness to Pluralism*, Synthese Library.