

Grand Challenges for Computing Research: SELF-REFLECTIVE MACHINE LEARNING

Dimitar Kazakov
CS Dept., University of York
kazakov@cs.york.ac.uk

21 Oct 2002

“Know thyself”. Inscription on the Oracle of Apollo at Delphi.

“...one of their Eyes [was] turned inward, and the other directly up to the Zenith.” Jonathan Swift, Gulliver’s Travels.

“Self-reference will eventually turn out to be at the core of artificial intelligence and the focus of all attempts to understand how human minds work” - Douglas Hofstadter, Gödel, Escher, Bach.

1. The Challenge

Design a machine learning program that, in addition to the ability of learning models from the data sets supplied, can acquire (learn) a model of itself by observing its input and output for the available data, and, possibly, by sending additional data it has generated to its input in order to explore its own behaviour. Should we use the now popular notion of *agents*, the task would be to build a learning agent that can study its own perceptions (inputs) and actions (outputs) in order to build a model of itself.

2. Background and Motivation

Since the Ancient Greeks, the faculty of intelligence has always been connected with and measured by the ability of self-observation. Montaigne’s “I am myself the subject of my book” has been paraphrased by many a philosopher who have seen in the study of oneself the ultimate challenge for the daring intellect. The ability for self-reflection as a general property that appears in and characterises natural and formal systems of a certain level of complexity is also discussed at length in Douglas Hofstadter’s book “Gödel, Escher, Bach: An Eternal Golden Braid”. This work, which focuses on understanding self-reference as a way of explaining consciousness, is a demonstration of how this apparently technical subject can attract the imagination of the lay public and experts (philosophers, linguists, logicians, mathematicians and computer scientists) alike to become a bestseller and win a Pulitzer Prize.

The challenge being launched here copies the open challenge that human intelligence has set to itself: *Is any artefact able to understand its true nature and, if at all possible, to what extent can it do so?* In the language of Systems Theory, one can speak of a system aiming at *identifying* itself. To a Machine Learning (ML) expert, the task will be to construct a learner able to learn its own model. Of course, the artefact would not have direct access to its own description (program, blueprint), but would only see itself as a black box, to which it could supply the inputs, and observe the outputs. These pairs of input-output observations would constitute the data for self-reflective learning.

The attempt at knowing oneself is part of the aim of bettering oneself. The decision to hide the description of the learner from itself could mistakenly be taken for a useless formal exercise, which takes the attention away from the more important issue of self-improvement. It is not so, and the reason is known to any designer of complex artefacts, who has implemented his project only to find that the real system shows some unexpected behaviour. In the case of a program, such discrepancies can appear between the programmer's intentions and the code he wrote, between the semantics of the program in theory and what the compiler actually generated, etc. This line of thought could be extended to assume possible bugs in the operating system, the design of the microprocessor, and even the systematic influence of the computer's environment. Such *reverse engineering* reconstructing design from observed behaviour can be performed as part of the standard test process on entities with static characteristics. In artefacts possessing the ability of learning and modifying themselves, it would be indispensable to include the property of self-reflection if any guarantee of their functionality throughout their development is to be preserved.

3. Evaluation

One can measure the ability of self-reflection by comparing the output of the self-reflective model the learner has acquired with the actual output of the learner for a given set of inputs.

While this criterion may seem self-centred, and not measuring the learner's ability to learn about its environment, one can expect that the ability of the learner to acquire models that are complex enough to represent its own algorithm would require that this algorithm should be of certain minimal complexity. This, in turn, would make the learning task harder. In this way, it is expected that an artefact able to achieve a certain level of self-reflection (as measured by the above criterion) would also be guaranteed the general ability to learn models of certain complexity, which would increase with the level of self-reflection achieved.

It may be the case that the ability of self-reflection is sufficient to measure the learner's general ability to build any models. Should this turn out not to be the case, the (classes of) learning algorithms capable of the same level of self-reflection would also be compared and their relative quality judged by the complexity of systems other than themselves that these algorithms could model successfully. This second criterion could be used only when the first is not sufficient or the two could be used independently, each introducing a partial order among the competing artefacts.

4. State of The Art

The existing work in Mathematics on self-referencing formal systems and the issues related to them (e.g., Russell's and Gödel's work), formal language theory since the seminal work of Chomsky (1957), work on meta-level architectures and reflection (Maes, 1987, see also ESPRIT LTR Project METAL on Meta-Learning and REFLECT, ESPRIT Basic Research Project on Reflective Expertise in Knowledge-Based Systems), along with the achievements in the areas of ML and Intelligent Agents have all prepared the ground for the challenge raised here. The task, although hard, and unachievable at

present, is conceivable in terms of the existing ML approaches, for instance: use ILP, a machine learner producing models in first-order logic (Prolog), to learn its own definition (code) originally written in the same language. The interest in the challenge is guaranteed by the high international interest in the areas of ML and learning agents as a whole. For instance, the SIG on Agents that Learn, Adapt and Discover (ALAD, <http://www.cs.york.ac.uk/~kudenko/alad/>) of the European Network AgentLink has 47 nodes, both from academia and industry. It is expected that such a challenge would boost both co-operation and healthy competition among all interested participants.

5. Discussion

Any self-referential model would necessarily be incomplete, as the infinite number of meta-levels of knowledge it prescribes (a model containing its own model, which contains its own model, etc. *ad infinitum*) would in practice be limited by the capacity of the physical medium (hardware) in which the model is stored. The challenge would open a discussion on languages allowing self-reference. Such languages can generate undecidable statements, such as the Liar's paradox, yet exactly this property may make them valuable for building artificial intelligence similar to our own.

The task can be decomposed and simplified by supplying the learner with partial models of its design. What could initially be a 'disembodied' algorithm may also be defined as an embodied agent interacting with a simulated or physical environment, possibly containing other agents. Once an – approximate – model of oneself is built, it can also be used to study one's own imperfections or emulate the actions of a *team* of such individuals and therefore bootstrap one's own abilities. The challenge can go beyond self-reflection, and aim at modelling systems of which the agent is part.

Research on ML oscillates between aiming at autonomous ML tools and such that assist experts, with the latter representing most of the commercially driven work on Scientific Discovery at present. Research on Learning Agents has only recently taken off, with very little work on self-reflection. This challenge is timely, as its first important achievements, expected within the next 10 years, will coincide with the emerging needs of the industry in both fields, which would otherwise only be starting to explore the area.

Work on the challenge would go beyond the initial goal to provide the necessary results for research on self-repairing and self-improving artefacts. The mental models of oneself could be replicated, combined and used to increase one's ability (and complexity). The algorithm would have to be able to evaluate its performance (including on the task of self-reflection!), realise its own limitations and attempt to surpass them by modifying itself and bootstrapping the abilities of the original design. Again, when – and if – achieved, the task will produce a new type of intelligent entity that is able to surpass itself in the way humankind did when it created the computer or, to use an even better analogy, in the way it will hopefully evolve when a complete understanding of the human genome content is achieved, and put to good use. It will add a new dimension to the term *artificial intelligence*, and give humankind the chance for the first time to study the challenges of self-reflection as an external observer.