Artificial Pedagogy: A Proposal

Jordi Bieger (jordi13@ru.is) supervised by Dr. Kristinn R. Thórisson
Center for Analysis and Design of Intelligent Agents (CADIA)
School of Computer Science, Reykjavik University

General intelligence is the ability to perform complex new tasks in a wide range of large and dynamic environments using available knowledge and resources. The knowledge to perform these tasks must be constructed throughout an entire lifetime as situations change and new challenges arise. Outside of very narrow domains it is virtually intractable to learn a complex new thing from scratch without guidance. Human children are trained, educated and raised to help them learn the basics, transfer humanity's highly sophisticated knowledge, and to grow cognitively. As far back as 1950 Alan Turing stressed the importance of teaching in "artificial intelligence" (AI) (Turing 1950). The way to get a system to operate at an adult level is not to program it directly, but rather to create a child machine and then educate it. Research in AI has mainly focused on developing algorithms with different learning mechanisms and much less on how to teach artificial systems. The goal of the proposed work is to study "artificial pedagogy" (AP): the science of how to teach an AI, with an emphasis on systems that aspire to reach or surpass human-level intelligence[1].

Successful operation in any domain requires relevant declarative, procedural and structural knowledge. This knowledge can either be innate or acquired through interaction with the environment, so to convey our knowledge to the system we have to either program or teach it. Programming is usually preferable when the system is only meant to work in a constant and well-understood task-environment, because it gives the developers great control over the system. However, there are many fixed tasks that we are able to perform without knowing exactly how we do it. One example of this is making moral judgments, which has given rise to some concern about the safety of superintelligent machines (Bieger, Thórisson, and Wang 2015). In some cases the knowledge representations that the AI uses may also be very alien to human programmers, either due to format (e.g. neural network weights are uninterpretable by humans) or content. According to the constructivist theory of learning, learners should actively construct their knowledge and its meaning from their own experiences (Liu and Matthews 2005). Since an AI might have vastly different sensors, actuators, goals and environments, the required knowledge may be unintuitive to program for a human. Teaching on the other hand can be a very convenient and natural: both teacher and learner are in their element, get to communicate in ways that their sensors and actuators allow, and can build new knowledge on top of what is already there while considering their own priorities. Another advantage of teaching is that it can be adapted and applied as needed. When new situations occur, a learner can always try to seek out a more knowledgeable teacher; even if those situations were unforeseen by the programmers. Even just programming in the knowledge for all possible foreseen situations would result in a bloated system with a lot of unnecessary knowledge for situations it will never see. Offloading knowledge on teachers means that no resources are wasted on storing unused knowledge.

[1]The notion of teaching here is broadly defined as "the intentional act of helping another system learn" and incorporates everything from simple training, to more advanced education, and even to aiding in cognitive growth by raising an artificial system like a child.

There is also a limit to how much can feasibly be programmed by a team of scientists in an acceptable amount of time. Thórisson (2012) argues that traditional constructionist software engineering methodologies are insufficient for building artificial general intelligence (AGI). A more holistic, constructivist approach is advocated where the system would largely self-construct and self-organize through interaction with the environment after starting with only a small amount of "seed" knowledge or "bootstrap code". This approach was implemented in the Autocatalytic Endogenous Reflective Architecture (AERA) (Nivel et al. 2013). Another promising cognitive architecture—NARS (which stands for Non-Axiomatic Reasoning System)—is also non-axiomatic in the sense that all of its domain knowledge comes from its experience (i.e. is not programmed in) (Wang 2013). A similar approach is taken in the growing fields of developmental AI and robotics, which focus on creating (embodied and embedded) systems that support cognitive development and growth from a childlike state through interaction with the environment (Asada et al. 2009; Guerin 2011).

Such systems start their "life" in a baby-like state with very little knowledge. Even with powerful learning mechanisms, learning from scratch can range from difficult to impossible. Even strong learners like humans cannot learn complex tasks like reading or chess completely on their own or using simple teaching methods like reward and punishment: more sophisticated education is necessary. With the increasing number of advanced cognitive architectures that support such sophisticated teaching methods and aspire to tackle complex tasks, AP becomes more and more relevant. But even traditional machine learning (ML) systems whose domains are narrowed down to the point where simple training methods suffice can benefit from more sophisticated teaching, e.g. in the form of curriculum learning, heuristic rewards or apprenticeship learning (c.f. Bieger, Thórisson, and Garrett 2014).

Computational learning theory defines the "teaching dimension" to analyze the minimum number of samples a specific kind of supervised learning algorithm needs to be shown to learn a concept (Goldman and Kearns 1995). With his proposal for "machine teaching" Zhu (2015) wants to move beyond looking at the cardinality of the optimal training set and to consider a wider class of traditional (supervised) machine learning algorithms. AP aims to extend this further to the full range of artificial learning systems, while placing an emphasis on those that aspire to artificial general intelligence. This theory of teaching should furthermore be practical and not make unrealistic assumptions like knowing all the details of the learner, its knowledge and its implementation. Such teaching needs to take place in the same kind of complex environment as learning, and is itself a highly nontrivial process that requires real-time interaction and even learning about the learner(s). Since no AP theory of teaching exists, teaching methods are usually developed on a case-by-case basis without a solid understanding of what works in what situation.

In the proposed work we will investigate possible teaching methods and how they affect system learning and utility over time. To this end, we will seek to formally define the central questions of AP as well as specify a framework for modeling the interaction between teacher, learner, task and environment. This will serve to develop a common terminology, which we can then use to place existing literature in AI, ML, developmental psychology and pedagogy in the context of AP. We aim to find what relevant characteristics of learners, teachers, and task-environments affect the availability and efficacy of different teaching methods. Since teachers will often lack detailed knowledge of their learners initially, they will need to obtain this information through interaction. This requires methods for (interactive) assessment, which could also be used to evaluate the success of our teaching methods (Garrett, Bieger, and Thórisson 2014; Thórisson et al. 2015).

The following list provides a rough overview of the plan:

- Define Artificial Pedagogy more rigorously
  - Formalize central questions of the field
  - Specify the interaction process between learner, teacher and environment
- Invent a classification system for teaching methods
  - Relate them to fields like developmental psychology, pedagogy, epigenetic robotics and AI
  - Identify use cases
  - Identify requirements on learning systems
- Invent new ways of evaluating the effects of teaching methods
  - Derive an evaluation framework
  - Investigate interactions with learners and task-environments

References[2]

Asada, Minoru, Koh Hosoda, Yasuo Kuniyoshi, Hiroshi Ishiguro, Toshio Inui, Yuichiro Yoshikawa, Masaki Ogino, and Chisato Yoshida. 2009. "Cognitive Developmental Robotics: A Survey." *Autonomous Mental Development, IEEE Transactions on* 1 (1): 12–34.

Bieger, Jordi, Kristinn R. Thórisson, and Deon Garrett. 2014. "Raising AI: Tutoring Matters." In *Proceedings of AGI-14*, 1–10. Quebec City, Canada: Springer.

Bieger, Jordi, Kristinn R. Thórisson, and Pei Wang. 2015. "Safe Baby AGI." In *Proceedings of AGI-15*, 46–49. Berlin: Springer-Verlag.

Garrett, Deon, Jordi Bieger, and Kristinn R. Thórisson. 2014. "Tunable and Generic Problem Instance Generation for Multi-Objective Reinforcement Learning." In *Proceedings of the IEEE Symposium Series on Computational Intelligence 2014*. Orlando, Florida: IEEE.

Goldman, S. A., and M. J. Kearns. 1995. "On the Complexity of Teaching." *Journal of Computer and System Sciences* 50 (1): 20–31.

Guerin, Frank. 2011. "Learning like a Baby: A Survey of Artificial Intelligence Approaches." *The Knowledge Engineering Review* 26 (02): 209–36.

Liu, Charlotte Hua, and Robert Matthews. 2005. "Vygotsky's Philosophy: Constructivism and Its Criticisms Examined." *International Education Journal* 6 (3): 386–99.

Nivel, E., K. R. Thórisson, B. R. Steunebrink, H. Dindo, G. Pezzulo, M. Rodriguez, C. Hernandez, et al. 2013. "Bounded Recursive Self-Improvement." Technical RUTR-SCS13006. Reykjavik, Iceland: Reykjavik University Department of Computer Science. http://arxiv.org/abs/1312.6764.

Thórisson, Kristinn R. 2012. "A New Constructivist AI: From Manual Methods to Self-Constructive Systems." In *Theoretical Foundations of Artificial General Intelligence*, 145–71. Springer.

Thórisson, Kristinn R., Jordi Bieger, Stephan Schiffel, and Deon Garrett. 2015. "Towards Flexible Task Environments for Comprehensive Evaluation of Artificial Intelligent Systems & Automatic Learners." In *Proceedings of AGI-15*, 187–96. Berlin: Springer-Verlag.

Turing, Alan M. 1950. "Computing Machinery and Intelligence." *Mind* 59 (236): 433–60.

Wang, Pei. 2013. *Non-Axiomatic Logic: A Model of Intelligent Reasoning*. Singapore: World Scientific Publishing.

Zhu, Xiaojin. 2015. "Machine Teaching: An Inverse Problem to Machine Learning and an Approach Toward Optimal Education." In *The Twenty-Ninth AAAI Conference on Artificial Intelligence (Senior Member Track, AAAI)*. Palo Alto, CA.

[2] Many references have been omitted due to space limitations; the official proposal submitted to and approved by my PhD committee has more than 6 pages of references.