Attention Based Learning as a Foundation for Conscious Agents

Joscha Bach

Harvard Program for Evolutionary Dynamics, Cambridge, MA
bach@fas.harvard.edu

Abstract. This position paper addresses the area of reflection and metacognition to extend it into learning. It suggests the outline of an attentional learning paradigm that replaces or augments backpropagation learning by using a protocol of past behavior control configurations that associates changes with expected outcomes and observational triggers that allow the application of a learning signal. The generation of an integrated protocol of the contents of an agent's attention may also be responsible for the functionality and phenomenology of consciousness.

Keywords: Artificial General Intelligence, Needs, Purposes, Attention, Attentional Learning, Consciousness, Cognitive Architectures, Common Model of Cognition

From Classical AI to Compositional Function Approximation

The last five years in Artificial Intelligence research have seen general a shift from "classical" approaches towards connectionist solutions (a.k.a. *deep learning* [1]), which might obscure a more fundamental change in perspective. Deep learning is often colloquially framed as a training (mostly) feed-forward ANNs by a variant of backpropagation with stochastic gradient descent, which limits its scope in the eyes of some researchers (c.f. [2]), but it may encompass many more techniques, and should perhaps better be understood as *compositional function approximation*. In this sense, Artificial Intelligence may have left an era in which practitioners identified and implemented algorithms that produced intelligent functionality directly ("first order AI") for one in which we create algorithms that automate the search for this functionality ("second order AI"). It is tempting to think that this heralds a third stage, in which we predominantly look for algorithms that automate the search for learning algorithms ("third order AI", or meta-learning).

In the context of cognitive architectures and the search for a *common model of cognition* [3], these developments might mean that we shift our focus from modularized models ("box and arrow architectures") and a multitude of combined functions to architectures for general learning/function approximation that are driven by a suitable multitude of rewards.

Today, our search for a possible common model is characterized by examining cognition as a set of disparate areas of functionality: emotion and motivation, perception and learning, knowledge and reasoning, reflection and metacognition, motor control, and language. This separation is partially pragmatic and due to the specialization and focus of individual groups of researchers, but it also reflects a modularizing view of the mind. And while this view is certainly accurate when we regard the performance of humans and many higher animals, it may not be the most productive when it comes to recreating such performances from a minimal set of orthogonal principles within a developmental paradigm.

A model of the minimal generative components of a cognitive system might well force us to take a more unified and minimalist perspective: a mind emerging over a set of regulation problems (represented as motivational system that defines suitable *needs* of the system, and can generate corresponding *rewards*), and a general learning paradigm that can produce models based on the rewards, in the service of regulation of the needs. The multitude of different modular functions we observe in a fully developed agent will be the result of a general [cortical] learning system, a suitable embedding of that system in an environment that delivers rich and cohesive training data, a developmental route (usually starting from proprioception and spatial cognition towards meta cognition, abstract reasoning and eventually the derivation of a complete model of the mind itself and the nature of its environment), and a variety of innate biases that guide the allocation of cognitive resources (attention) to speed up the convergence of the developing cognitive architecture.

The idea of a general learning system implies that—somewhat similar to the class of universal computers (i.e. the set of computable functions that can compute all computable functions [within the bounds of their resources], and which contains itself)—there exists a class of efficient universal function approximators, which can approximate all efficiently computable functions that can be efficiently approximated by a universal computer [within its resource bounds]. If this class turns out to contain itself, it describes Artificial General Intelligences (AGI), and if we succeed in developing an AGI, we are arguably generally intelligent, too.

While the currently dominant deep learning paradigms are obviously expressive enough to implement arbitrary algorithms, it is not clear that they can efficiently discover all the classes of algorithms we are interested in. Training a multi-layer network with stochastic gradient descent usually requires that the network is either globally differentiable, or set up in very specific ways [4, 5] that allow us to deal with recurrent and lateral connections (as they are common in our own brains). This means that we may need to explore different learning paradigms, such as *attention based learning*.

Needs and Purposes

The learning and decision making of cognitive agents can be described as driven by reward maximization. In the author's own MicroPsi architecture, rewards are generated by needs [6], which yield reinforcement signals when satisfied ("pleasure")

or frustrated ("pain"). Needs may be physiological (nutrition, rest, physical integrity etc.), cognitive (competence, exploration, aesthetics), or social (affiliation, dominance, nurturing, legitimacy etc.). However, while needs provide rewards, behavior needs to be driven by *anticipated rewards*, i.e. models of the needs. The model of the anticipated rewards of an agent are its *purposes*. Unlike needs (which may have different relative strengths), purposes form hierarchies: Maslow's famous "hierarchy of needs" [7] is better understood as a hierarchy of purposes. This hierarchy itself is the result of the requirement of making conflicting purposes accountable to each other. In humans, this hierarchy does not have a single, universal structure, but may be subject to interpersonal differences: some people may work so they can eat, and others eat so they can work. Even if human behavior eventually turns out to be quite similar, because we have to serve similar needs, the purposes of individuals may differ considerably.

We may understand purposes as the goals of clusters of learned behavior programs: the purpose of brushing your teeth is having clean teeth, the purpose of visiting a restaurant may be to feed or to meet friends, and the ego of a person is a purpose that integrates their expected reward over their anticipated lifetime. We may also serve purposes outside of our person, such as relationship goals, and ideological, patriotic or religious purposes. Non-transactional cooperation between individuals is built on the recognition of shared extra-personal purposes. While needs are associated with an actual reward (which acts as a reinforcement), the anticipated reward of purposes does not have to manifest itself to guide behavior.

At each point in time, a cognitive agent may be serving a variety of purposes, but its actual behavior usually needs to select a single course of action. This focus on a single behavioral program or strategy is facilitated by the attentional system. A disruption in the control of attention makes it difficult for the agent to serve its purposes: we may understand "free will" as the ability to serve one's purposes. In this sense, the opposite of free will is not determinism or coercion, but compulsion, i.e. the disruption of purposeful behavior by behaviors that opportunistically serve needs instead of the hierarchy of the agent's purposes. Attentional deficits are often best understood as deficits in the regulation of attention, i.e. in the selection of behaviors, according to the regulation model that the agent has formed over its needs, or a deficit in the ability to acquire a suitable hierarchy by learning how to best satisfy the multitude of the agent's needs.

A Strategy for Attentional Learning

How can we train individual behaviors in a neural architecture that cannot be treated as differentiable, such as our neocortex? Without this property, it is difficult to propagate a learning signal through a large number of processing layers, as it is currently done in most deep learning applications. A possible solution might involve a specific variant of attentional control: during a trial (such as the attempt of getting better at playing tennis), the agent performs a targeted, conditional change on the part of the behavioral architecture that it expects to have the greatest influence on the

outcome, and stores the latent variables required to recreate the corresponding partial brain configuration later on (for instance, the preconditions of the motor program to produce a particular swinging motion), together with the expected result (e.g., the ball moving in a particular way), and the conditions that are expected to make it possible to evaluate the outcome of the action (such as the ball being intercepted by the other player a few moments later, or the match being won or lost a few minutes later). These data are stored in an attentional protocol, indexed by the relative order of events and the conditions that should trigger the recreation of the brain state and attentional pointer on the action.

The agent is now free to instantiate new behaviors and perform new observations, thereby changing its mental configuration, until the triggering condition is being met (for instance, the tennis match is being won or lost). The agent may now restore the state of the behavior program that was responsible for the action, and (depending on the outcome) either reinforce the change, or undo its effects.

This learning strategy requires the maintenance of an integrated protocol of the agent's contents of attention (see [8] for a more detailed description). Because it narrowly targets a single variable within a behavior program instead of many weights over several layers of control, it may often converge much faster to the desired result than backpropagation, but it relies on an attentional system that can pinpoint the relevant behavior programs and their variables. In other words, the agent needs to learn and update a model of the structure its own evolving cognitive architecture. I suspect that the best way of doing that may be to train the attentional learning system by itself, i.e. using reflexive attention.

Attentional learning may present a solution to the problem of assigning credit to actions with arbitrarily delayed outcomes, but I suspect it may also be used to interactively construct and modify analytic models, whereby the outcome of a representational change can be observed without delay. Attentional learning with a reflexive protocol may be instrumental in analytic reasoning.

From Attention to Consciousness

The integration of the contents of attention into an indexed protocol that allows the recreation of past configurations of an agent's representational and behavioral models, and the reflexive access to these contents may possibly explain the functionality and purpose of access consciousness and reflexive consciousness in humans, as for instance suggested by Graziano [9], Dennett [10], or Drescher [11]. In this sense, the purpose and functionality of conscious attention would be the creation of indexed memories, and the purpose and functionality of conscious recall would be the application of a learning signal to the recalled behavioral or representational configuration. Conversely, behaviors that do not require additional learning (according to the expectations generated by the agent's attentional system) will be executed and orchestrated without conscious attention and/or recall, and won't become part of the agent's consciously remembered stream of events.

But how does such a paradigm account for the phenomenal experience of consciousness? How is it possible for a computational system, such as an AGI, to *experience what it is to be like* in a given situation? This difficulty has led some researchers, such as Christof Koch, to suggest that consciousness can only occur in biological brains, not in a computational simulation [12]. I believe that this interpretation has it exactly backwards: no physical system can have conscious experience. Biological neurons, electronic switches or combinations thereof cannot know what it is like be a person, to experience having a self, or to experience what it is like to be confronted with a universe, because they don't form persons, and don't have direct access to a physical reality. Instead, our minds form simulation models of the universe, dynamical virtual realities that are tuned to predict the progression of patterns of sensory data our nervous systems are confronted with, and simulation models of persons, to evaluate how a person would react to a universe that is relevant to its needs and purposes.

In this way, we may picture our personal self as a simulacrum within a simulation, a fictional character embedded in the virtual, multimodal narration of a dynamic universe. It is not the brain (or an AGI's hardware or computational architecture) that experiences being aware of the universe. Instead, the simulated persona inside of the system's simulation of the world recalls these experiences, and by giving this simulated person access to the language center, the system gives rise to the surprised utterances of fictional beings that find themselves in a dream generated by their host's hardware. Physical simulations cannot experience consciousness—only simulations can.

Summary

Our field tends to view the mind as a set of complex and somewhat modular functions that form a distinct architecture, and recreating mental functionality is seen as investigating and modeling its various subdomains. (This includes the author's own efforts in the context of the MicroPsi architecture.) A simpler and more unified perspective may start from the minimal set of principles required to generate a cognitive system like ours, and it includes algorithms for general learning, and a motivational system that can generate rewards to give rise to perception, the creation of a dynamic, cohesive world and situational awareness, the abstraction of perceptual models into knowledge, the creation of a physiological, agentic, social and metacognitive self, the acquisition of language and so on. This contribution sketches the abstraction of needs into purposes, and a strategy for attention based learning that may help us to overcome the architectural limitations of stochastic gradient descent learning.

References

1. LeCun, Y., Bengio, Y., Hinton, G. Deep Learning. Nature 521, 436–444 (28 May 2015)
2. Marcus, G. Deep Learning: A Critical Appraisal. arXiv:1801.00631 (2 January 2018)
3. Laird, J., Lebiere, C., Rosenbloom, P. A Standard Model of the Mind: Toward a Common Computational Framework across Artificial Intelligence, Cognitive Science, Neuroscience, and Robotics. AI Magazine. 38. 13. 10.1609/aimag.v38i4.2744 (2017)
4. Hochreiter S., Schmidhuber, J. Long short-term memory. Neural Computation. 9 (8) 1997: 1735–1780
5. Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), October 2014, 1724—1734
6. Bach, J. Modeling Motivation in MicroPsi 2. Artificial General Intelligence, 8th International Conference, AGI 2015, Berlin, Germany: 3-13
7. Maslow, A.H. A theory of human motivation. Psychological Review. 50 (4) 1943: 370–96.
8. Bach, J. The Cortical Conductor Theory: Towards Addressing Consciousness in AI Models. Proceedings of BICA 2018, Prague
9. Graziano, M. S. A., Webb, T. W. A Mechanistic Theory of Consciousness. International Journal on Machine Consciousness (2014)
10. Dennett, D. C. Consciousness Explained. Back Bay Books, New York (1992)
11. Drescher, G. Good and Real. MIT Press (2006)
12. Tononi, G., Boly, M., Massimini, M., Koch, C. Integrated information theory: from consciousness to its physical substrate. Nature Reviews Neuroscience. 17 (7) 2016: 450–46