A Neurobiologically Inspired Plan Towards Cognitive Machines

Jeffrey L. Krichmar[0000−0003−0739−2468]

Department of Cognitive Sciences, Department of Computer Science, University of California, Irvine, Irvine, CA 92697-5100 USA

Abstract. Despite incredible recent progress in artificial intelligence, current systems fall short of what we would consider to be intelligent, thinking machines. This paper presents a neurobiologically inspired path towards creating cognitive machines. It suggests that incorporating aspects found in biological organisms, such as flexible learning, efficient processing, embodiment, value systems, and predictive coding could lead to systems that are truly cognitive.

Keywords: Cognition · Embodiment · Neuroscience · Neurorobotics.

1 Introduction

In this paper, I describe a pathway towards designing and constructing intelligent, cognitive machines. It stems from the goals of neurorobotics [18], where 1) deploying embodied agents could lead to a holistic understanding of how the nervous system gives rise to cognitive behavior, and 2) following the brain's architecture and dynamics may lead to truly cognitive machines. I feel the latter goal is necessary for artificial intelligence since the brain can serve as a working model for intelligence, cognition, and possibly consciousness.

In a recent article, Jeff Hawkins stated that intelligent systems must incorporate these aspects of the brain [10]: 1) Learning by rewiring, 2) Sparse representations, and 3) Embodiment. I would add 4) Value and 5) Prediction to this list. Furthermore, he stated that future thinking machines can ignore many aspects of biology, but not these. Taking the point of view of a neuroroboticist, I will expand on each of these aspects in the remainder of the paper.

2 Aspects of the Brain for Designing Future Machines

2.1 Learning By Rewiring

Brains exhibit some remarkable learning properties that have not been replicated by artificial intelligence or machine learning to date. Organisms learn quickly, sometimes with only one presentation of a new stimulus or situation. It's not just a human ability, rats can learn new contexts in a single experience [30]. Compare this to a deep learning system or neuroevolutionary algorithm that

takes thousands of iterations to learn a task. It could be argued that humans build up years of experience and that one-shot learning leverages this experience. However, Hawkins makes the point that learning is incremental. We can learn something new without retraining the entire brain or forgetting what we learned before. This is an open issue in artificial systems, in which catastrophic forgetting or catastrophic interference are active areas of research [29, 14]. Furthermore, most artificial systems are trained to some criterion and then learning is frozen for deployment. In contrast, biological organisms learn throughout their lifetime, while maintaining old memories.

In the brain, rapid learning by the hippocampal formation and its interaction with the neocortex are key to learning and memory [13, 19]. In our own work, we showed that a biologically plausible neural network model, with interactions between the hippocampus and the medial prefrontal cortex, was able to learn and consolidate memory schemas over time, as well as quickly assimilate new information if it was consistent with a prior schema [12]. The neural network was also able to learn multiple schemas without catastrophic forgetting. In robotic studies, we showed that the interactions between a simulated hippocampus and neocortex during goal directed behavior could lead to the formation of episodic memories [17, 7]. These simulation and neurorobot experiments suggest that the brain's architecture has evolved a means to support lifelong learning in a way that is different from current artificial approaches.

2.2 Sparse Representation

Biological organisms are under tight metabolic constraints, and the brain utilizes a number of means to reduce energy expenditure, while maximizing performance. One way to conserve energy is to reduce the amount of neural activity and neurons necessary to represent information. Indeed, sparse coding and dimensionality reduction is a common coding strategy across multiple brain regions. In our own simulations, we have shown that dimensionality reduction and sparse coding is an efficient coding strategy that is prevalent throughout the brain [3].

Many sensory and cortical representations in the brain can be recovered by applying dimensionality reduction and sparsity constraints to their inputs. For example, a sparse, parts-based representation of visual motion emerged, which showed a remarkable resemblance to receptive fields observed cortical area MSTd, by applying a dimensionality reduction technique known as Non-negative Matrix Factorization (NMF) to MSTd's inputs [2]. When we applied NMF to neurophysiological recordings of the retrosplenial cortex during a rodent navigation task [36], we were able to replicate neural activity during the experiment and predict the rat's behavior. In both cases, stimuli were represented by only a small number of neurons (population sparsity), and any given neuron was activated by only a small number of stimuli (lifetime sparsity). These simulations suggest the brain has evolved ways to represent information efficiently without loss of information.

2.3 Embodiment

Brains do not work in isolation; they are closely coupled with the body acting in its environment. The brain is embodied and the body is embedded in the environment. In fact, there is compelling evidence that the *Body Shapes the Way We Think* [23], rather than the brain telling the body how to act. Biological organisms perform morphological computation, that is, certain functions performed by the body alleviate costly brain processing. For example, bipedal locomotion is a difficult control problem that we carry out with ease and without even thinking. Passive walker robots, by exploiting gravity and friction, demonstrate natural walking gaits that have simple control policies and utilize orders of magnitude less energy than conventional walking robots [4, 5].

In our own neurorobotics work, where we construct large complex neural networks to control behavior, embodiment is still a strong driving force. For example, the timing of whisker activations allowed our robot to construct spatiotemporal representations of textures [27]. In our soccer playing Segway robot, a simple plastic tubing, which resembled a Hula hoop, alleviated our detailed visual cortex model from constructing trajectories, by trapping the ball to its body [8]. In general, there is always some aspect of the interaction between the neural network (brain), the robot (body) and the environment that leads to unexpected results and more intelligent behavior.

2.4 Value

Organisms adapt their behavior through value systems that signal contextual information, trigger learning, and select actions. Neuromodulatory systems act as value systems by signalling rewards, costs, surprises and other important event to the rest of the brain [15, 1]. The neuromodulatory systems are subcortical regions in the brain that have a strong influence on a number of brain areas thought to be involved in cognition. These neuromodulatory regions send their signals through different neurotransmitters; Dopamine signals reward, saliency, novelty, and invigoration. Serotonin signals harm aversion, anxious states, and withdrawal. Norepinephrine maintains a vigilance signal and tracks unexpected uncertainty. Acetylcholine is critical for memory consolidation, attention, and tracking expected uncertainty.

In robotics, neuromodulatory value systems can control behavior by changing the agents cognitive state. For example, in a robotic version of the open field test, a robot mimicked rodent behavior by staying near walls or near a nest when it was anxious about an unfamiliar environment [16]. However, once it sensed the environment was safe, curiosity took over and the robot explored novel objects in the middle of the environment. Simulated acetylcholine and norepinephrine allowed the robot to respond quickly to novel events and habituate to uninformative events. Increasing serotonin levels in the model led to risk averse behavior (i.e., staying near the walls or nest), whereas increasing dopamine levels led to invigorated curious behavior (i.e., examining objects in the middle of the environment).

2.5 Prediction

Prediction is crucial for fitness in a complex world. The main functions of the brain are predicting and planning for the future, and adaptation when the result does not meet expectations. The central nervous system is rather slow to respond, too slow and cumbersome to keep up with environmental change. The body or peripheral nervous system can handle much of the rapid sensing and motor actions necessary via morphological computation. However, a predictive engine leads to planning, imagery and quite possibly consciousness.

Prediction requires the construction and maintenance of an internal model. The brain maintains internal models for a wide range of behaviors; from motor control to language processing [28, 11]. There is evidence for neural correlates of model-based reinforcement learning in the prefrontal cortex, where an internal model is maintained to predict the value of future decisions [9]. In the rodent hippocampus, neural traces have been observed while mentally evaluating different paths before taking action [24, 26]. Prediction and inference are fundamental computations in cortical systems [25]. These predictive models in the brain allow the organism to plan for the future and are advantageous when deliberation before action is possible. In robotics, these strategies have inspired robot controllers that develop internal models to predict movement of objects and of other robots [21, 20].

Prediction can lead to deliberation, mental simulation and mental imagery, all important aspects of cognition. It is compatible with the ability to create a scene in one's mind, which has been called the 'remembered present' or primary consciousness [6]. Moreover, prediction is important for having a theory of mind; the ability to understand and predict the intentions of others [22]. This awareness of one's self and others would be a critical component for any conscious machine.

3 Conclusion

Artificial systems have made great progress in recent times, but currently fall short of what we would call cognitive or conscious machines. Using the brain as an existence proof, it is argued here that aspects of neural computation could bridge this gap. Specifically, 1) Learning, 2) Efficient information processing, 3) Embodiment, 4) Value signaling, and 5) Predictive coding are aspects of the brain that should be included in future systems. Biological organisms are the ultimate learning machines. They learn quickly, incrementally, and over a lifetime. Much is now known about different neurobiological learning rules and the roles different brain regions play in encoding and recalling diverse memories. Biology is under tight energy constraints and the brain is amazingly power efficient. This leads to efficient information processing in the form of sparse, reduced representations of environmental features and actions. Not only will this lead to power efficient cognitive machines, it will also lead toward rapid decision making. Brains do not work in isolation. Much of what is considered cognitive is a close coupling between brain, body, and environment. Such a coupling requires multimodal sensorimotor integration and morphological computation. Future

cognitive machines need to take this into account. Taken together, these aspects of the brain may provide a design pathway for future cognitive machines that may have some degree of what we would call consciousness.

References

1. Avery, M.C., Krichmar, J.L.: Neuromodulatory systems and their interactions: A review of models, theories, and experiments. Frontiers in Neural Circuits **11**(108) (2017)
2. Beyeler, M., Dutt, N., Krichmar, J.L.: 3d visual response properties of mstd emerge from an efficient, sparse population code. The Journal of Neuroscience **36**(32), 8399–8415 (2016)
3. Beyeler, M., Rounds, E.L., Carlson, K., Dutt, N., Krichmar, J.: Sparse coding and dimensionality reduction in cortex. bioRxiv (2017)
4. Bhounsule, P.A., Cortell, J., Grewal, A., Hendriksen, B., Karssen, J.G.D., Paul, C., Ruina, A.: Low-bandwidth reflex-based control for lower power walking: 65 km on a single battery charge. The International Journal of Robotics Research **33**(10), 1305–1321 (2014)
5. Collins, S., Ruina, A., Tedrake, R., Wisse, M.: Efficient bipedal robots based on passive-dynamic walkers. Science **307**(5712), 1082–5 (2005)
6. Edelman, G.M.: Remembered Present: A Biological Theory Of Consciousness. Basic Books (1990)
7. Fleischer, J.G., Gally, J.A., Edelman, G.M., Krichmar, J.L.: Retrospective and prospective responses arising in a modeled hippocampus during maze navigation by a brain-based device. PNAS **104**(9), 3556–3561 (2007)
8. Fleischer, J., Szatmary, B., Hutson, D., Moore, D., Snook, J., Edelman, G., Krichmar, J.: A neurally controlled robot competes and cooperates with humans in segway soccer. In: IEEE International Conference on Robotics and Automation (2006)
9. Glascher, J., Daw, N., Dayan, P., O'Doherty, J.P.: States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. Neuron **66**(4), 585–95 (2010)
10. Hawkins, J.: What intelligent machines need to learn from the neocortex. IEEE Spectrum (2017)
11. Hickok, G., Houde, J., Rong, F.: Sensorimotor integration in speech processing: computational basis and neural organization. Neuron **69**(3), 407–22 (2011)
12. Hwu, T.J., Krichmar, J.L.: A neural model of schemas and memory consolidation. bioRxiv (2018)
13. van Kesteren, M.T., Beul, S.F., Takashima, A., Henson, R.N., Ruiter, D.J., Fernandez, G.: Differential roles for medial prefrontal and medial temporal cortices in schema-dependent encoding: from congruent to incongruent. Neuropsychologia **51**(12), 2352–9 (2013)
14. Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A.A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., Hassabis, D., Clopath, C., Kumaran, D., Hadsell, R.: Overcoming catastrophic forgetting in neural networks. Proceedings of the National Academy of Sciences **114**(13), 3521 (2017)
15. Krichmar, J.L.: The neuromodulatory system - a framework for survival and adaptive behavior in a challenging world. Adaptive Behavior **16**, 385–399 (2008)

16. Krichmar, J.L.: A neurorobotic platform to test the influence of neuromodulatory signaling on anxious and curious behavior. Front Neurorobot **7**, 1–17 (2013)
17. Krichmar, J.L., Nitz, D.A., Gally, J.A., Edelman, G.M.: Characterizing functional hippocampal pathways in a brain-based device as it solves a spatial memory task. Proc Natl Acad Sci U S A **102**(6), 2111–6 (2005)
18. Krichmar, J., Wagatsuma, H.: Neuromorphic and Brain-Based Robots. Cambridge University Press (2011)
19. Kumaran, D., Hassabis, D., McClelland, J.L.: What learning systems do intelligent agents need? complementary learning systems theory updated. Trends Cogn Sci **20**(7), 512–534 (2016)
20. Murata, S., Yamashita, Y., Arie, H., Ogata, T., Sugano, S., Tani, J.: Learning to perceive the world as probabilistic or deterministic via interaction with others: A neuro-robotics experiment. IEEE Trans Neural Netw Learn Syst **28**(4), 830–848 (2017)
21. Park, J.C., Lim, J.H., Choi, H., Kim, D.S.: Predictive coding strategies for developmental neurorobotics. Front Psychol **3**, 134 (2012)
22. Pearson, J., Clifford, C.W., Tong, F.: The functional impact of mental imagery on conscious perception. Curr Biol **18**(13), 982–6 (2008)
23. Pfeifer, R., Bongard, J.: How the Body Shapes the Way We Think: A New View of Intelligence. The MIT Press, Cambridge, MA (2006)
24. Pfeiffer, B.E., Foster, D.J.: Hippocampal place-cell sequences depict future paths to remembered goals. Nature **497**(7447), 74–9 (2013)
25. Richert, M., Fisher, D., Piekniewski, F., Izhikevich, E.M., Hylton, T.L.: Fundamental principles of cortical computation: unsupervised learning with prediction, compression and feedback. arXiv:1608.06277 (2016)
26. Schmidt, B., Papale, A., Redish, A.D., Markus, E.J.: Conflict between place and response navigation strategies: effects on vicarious trial and error (vte) behaviors. Learn Mem **20**(3), 130–8 (2013)
27. Seth, A.K., McKinstry, J.L., Edelman, G.M., Krichmar, J.L.: Active sensing of visual and tactile stimuli by brain-based devices. International Journal of Robotics and Automation **19**(4), 222–238 (2004)
28. Shadmehr, R., Krakauer, J.W.: A computational neuroanatomy for motor control. Exp Brain Res **185**(3), 359–81 (2008)
29. Soltoggio, A., Stanley, K.O., Risi, S.: Born to learn: The inspiration, progress, and future of evolved plastic artificial neural networks. Neural Networks **108**, 48–67 (2018)
30. Wagatsuma, A., Okuyama, T., Sun, C., Smith, L.M., Abe, K., Tonegawa, S.: Locus coeruleus input to hippocampal ca3 drives single-trial learning of a novel context. Proc Natl Acad Sci U S A **115**(2), E310–E316 (2018)