CCAI-chair Research Plan

Richard S. Sutton, May 2025

Learning from Experience

My research program is based on understanding intelligence as the process of learning from first-person experience interacting with an environment. As I have said since 1993 on my internet homepage:

> *"I am seeking to identify general computational principles underlying what we mean by intelligence and goal-directed behavior. I start with the interaction between the intelligent agent and its environment. Goals, choices, and sources of information are all defined in terms of this interaction. In some sense it is the only thing that is real, and from it all our sense of the world is created. How is this done? How can interaction lead to better behavior, better perception, better models of the world? What are the computational issues in doing this efficiently and in realtime?"*

The claim is that the proper focus of intelligence, and thus of AI research, is on the agent's first-person experience—the signals passing back and forth between it and its environment. In some sense this is undeniably true and, by definition, could be no other way. The agent's intelligence is revealed only through its half of the interaction (its actions) and can only be assessed by the effect on the other half (its observations). For the agent to have knowledge of the environment can only mean for it to have knowledge of these effects. As far as any agent is concerned, its experience stream is all there is. Whatever other ways one may think of the world—e.g., in terms of objects, physics, or other agents with all their layers of complexity—from an intelligent agent's perspective all those things are merely patterns in its experience. The experience stream is the entire input and the entire output of the agent's computations. That *intelligence is all and entirely about working with and understanding the experience stream* is the obvious yet audacious idea that my research program is based on.

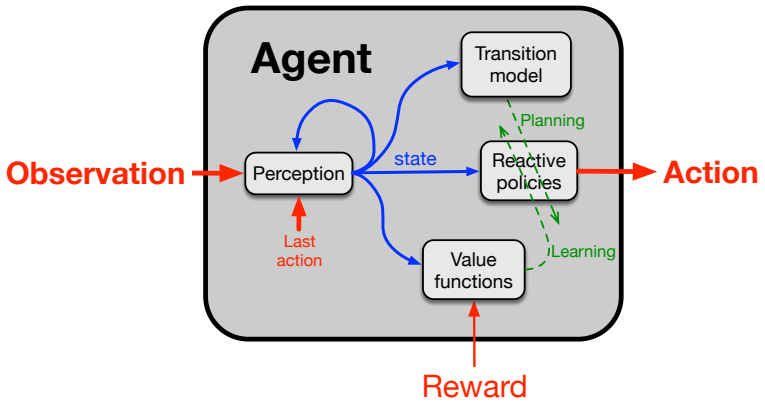
The possibility of *prior knowledge* might appear to muddy this idea, but really it does not. It is true that knowledge enters into the design of the agent prior to its interaction with the environment, but any resulting improvement in its behavior should not be viewed as due to the agent's intelligence, but, rather, as due to the intelligence of the designer. "Intelligence" as we mean it here does not mean simply performing well, but performing well because of the agent's computations on the experience stream. The upshot is that my research program eschews prior knowledge. Prior knowledge will always be present, of course; a robot must be given appropriate sensors and effectors, for example, but such prior knowledge is always just a starting point for intelligence. Researching domain knowledge is analogous to designing a better camera or a better motor. These are important tasks that must be done well in order for the agent to be effective, but they are not the focus of computational AI, or at least not of the computational AI that I am proposing here.

Without spending too much time on it, I also note that the experiential perspective on AI is very different from that of the many researchers developing large language models and other "generative AI" methods. These methods use human data to learn and fine-tune artificial neural networks that can mimic human behavior, but which are then no longer able to learn. The weights of these systems are frozen when deployed and are literally unaffected by the experience stream that occurs during their normal operation—whereas the proposed research is exclusively focused on what that effect should be! AI systems that mimic people are extraordinarily useful in many ways, but are also strongly limited. It remains possible that the early appeal of generative AI will be another instance of "the bitter lesson" [197] and that in the long run they will be eclipsed by experiential AI [47]. Both approaches should be fully pursued of course; here we propose to fully pursue experiential AI.

The Alberta Plan for AI Research

I propose to pursue experiential AI by following "the Alberta plan for AI research" [195] recently proposed by me and my fellow CCAI chairs, Michael Bowling and Patrick Pilarski. The plan presumes that intelligence is experiential and that it can be understood in terms of a relatively small number of general principles, including the basic ideas of reinforcement learning (RL), but also principles that have not yet been discovered or that are incompletely understood. The twelve steps of the plan proceed from simpler to more-complex problem settings, ending with the full problem of model-based RL with hidden state and the need for learned abstractions in state and time. Performance in the full setting is the ultimate goal, but focusing on it is probably not the best way to make progress in discovering the missing general principles. The earlier steps are meant to enable incremental progress in discovering principles by encountering each challenge in the simplest setting in which it arises. I discuss examples of this strategy later in this proposal.

The Alberta plan builds from a base agent with four components that would be immediately familiar to most RL researchers, as diagrammed below.



The base RL agent of the Alberta plan has four components, including a *perception* component that constructs the state from low-level observations and *reactive policies* that map the state to low-level actions. Assisting in the learning of the policies are *value functions* and a *transition model* of the world used in planning. All components are intended to be learned online.

The plan details twelve steps, each increasing in challenge and complexity, with the following titles:

1. Representation I: Continual supervised learning with given features
2. Representation II: Supervised feature finding
3. Prediction I: Continual Generalized Value Function (GVF) prediction learning
4. Control I: Continual actor-critic control
5. Prediction II: Average-reward GVF learning
6. Control II: Continuing control problems
7. Planning I: Planning with average reward
8. Prototype-AI I: One-step model-based RL with continual function approximation
9. Planning II: Search control and exploration
10. Prototype-AI II: The STOMP progression
11. Prototype-AI III: Oak
12. Prototype-IA: Intelligence amplification

The steps need not be pursued in order, or one at a time. For example, in the first term of my CCAI chair, the focus was initially on how all the agent components fit together to form a complete model-based RL agent, that is, on Steps 9-11. Recall that one of the main outcomes was a journal paper [9] describing a prototype AI based on the STOMP progression (the progression from SubTasks to Options for solving them to transition Models of the options used in Planning); that work pertained to Step 10. Another major focus in the first term was on average-reward and continuing formulations of RL problems [66,65,64,144,151,d8,d12], which pertain to Steps 5 and 6. We then realized that we were being held back by limitations of our function approximator, that is, by limitations of deep-learning's backpropagation algorithm. To address these, we explored loss of plasticity and continual backpropagation in our paper in *Nature* [5], which was in large part a retreat to supervised learning and Step 2.

Continual Supervised Learning

Although the steps need not be taken in order, there are advantages to doing so, and this is what I propose to do in the renewal. In particular, I propose to start by refining *linear* supervised learning (Step 1) and then proceeding to *non-linear* supervised learning (Step 2). I feel that even with given features, continual supervised learning is in need of refinement and deserves substantially more attention prior to introducing the complexities of the non-linear case. I see potential for meaningful, clarifying refinements in three areas:

1. *Normalization* of the input signals, targets, and weight updates so that learning is fast and reliable, and to ease the setting or optimization of meta-parameters such as step sizes.
2. *Meta-learning* of per-weight step sizes, thereby autonomously assessing feature relevance. The IDBD algorithm shows that this is possible [128,196], but we have just begun to explore how it can be sped up and made more reliable [91].
3. *Elimination of all meta-parameters* from the learning and meta-learning algorithms.

If all of these goals are achieved in the linear setting, it will provide a good foundation for our real target, Step 2, the non-linear, artificial-neural-network setting.

We will address three new big challenges in the deep-learning setting of Step 2, while still considering only feedforward supervised-learning networks:

1. *Continual learning*. Linear learning with constant step sizes has no difficulty maintaining plasticity with continued learning, but this ability is lost in non-linear networks [5,194]. Continual backpropagation provides part of the answer, but further improvements are possible, and there are many more ideas to explore and test.
2. *Meta-learning* for structural credit assignment and sculpting generalization. Mechanically, this is also about optimizing step sizes, as in the IDBD algorithm, but for the full network as opposed to a single linear layer [191]. Conceptually, this will enable representation learning, in particular, the changing of how the network generalizes.
3. *Organically-grown network architectures*. Conventionally, the weights of the network are learned, but the connections (number of layers, connection pattern) is chosen by people. Really both steps should be automated; the connections should be organically grown in response to the needs of experience. See, for example, "cascade correlation networks" (Fahlman & Lebiere 1989), the "pocket" and "tower" algorithms (Gallant 1993), NEAT (Stanley & Miikkulainen 2002), columnar-constructive networks [8], and AutoGrow (Wen et al. 2020).

Note that it is #1 that unlocks #2 and #3. Only if learning continues for a long time is it possible to try different ways of learning or different connection patterns and meta-learn which of them learn faster and generalize better. The theoretical issues are complex (Wolpert 1996, Adam et al. 2019), but continual learning offers a sound new way of learning to generalize [103].

Impact on AI

My full research plan is to follow all the steps of the Alberta plan, yielding a full understanding of model-free and then model-based intelligence centered and grounded in experience. The grounding in experience offers the potential of scaling and performance beyond human abilities. If fully successful, experiential AI would transform the long-term course of AI research.

However, a major impact of this research could also come much more quickly than that. Consider just the first two steps of the Alberta plan as discussed in the previous section. The inability of deep-learning networks to fully ingest new data without retraining from scratch at the cost of millions of dollars is a pain point in current AI practice. And if the continually learning network could meta-learn better ways of generalization—rather than relying on human programmers to do this—this again could be remarkably useful in practice. If just Steps 1 and 2 were done well, then I think the resulting deep learning algorithms would become widely used in generative AI and transform the practice of the current AI industry.

References

Adam, S. P., Alexandropoulos, S. A. N., Pardalos, P. M., & Vrahatis, M. N. (2019). No free lunch theorem: A review. *Approximation and optimization: Algorithms, complexity and applications*, 57-82. Springer Nature Switzerland.

Fahlman, S., & Lebiere, C. (1989). The cascade-correlation learning architecture. *Advances in neural information processing systems*, *2*.

Gallant, S. I. (1993). *Neural network learning and expert systems*. MIT press.

Stanley, K. O., & Miikkulainen, R. (2002). Evolving neural networks through augmenting topologies. *Evolutionary computation, 10*(2), 99-127.

Wen, W., Yan, F., Chen, Y., & Li, H. (2020). Autogrow: Automatic layer growing in deep convolutional networks. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining* (pp. 833-841).

Wolpert, D. H. (1996). The lack of a priori distinctions between learning algorithms. *Neural computation 8*(7), 1341–1390 (1996)

d8 Yi Wan, "Learning and Planning with the Average-Reward Formulation," University of Alberta, 2023. Currently Research Scientist at a startup

d12 Abhishek Naik, "Reinforcement Learning for Continuing Problems using Average Reward," University of Alberta, 2024. Currently Research Scientist at a Canadian Space Agency

5. Dohare, S., Hernandez-Garcia, J.F., Lan, Q., Rahman, P., Sutton, R. S., Mahmood, A.R. (2024). Loss of plasticity in deep continual learning. *Nature 632*:768-774.

8. Javed, K., Shah, H., Sutton, R. S., White, M. "Scalable real-time recurrent learning using columnar-constructive networks. *Journal of Machine Learning Research 24*(256):1-34, 2023.

9. Sutton, R. S., Machado, M. C., Holland, G. Z., Timbers, D. S. F., Tanner, B., White, A. "Reward-respecting subtasks for model-based reinforcement learning." *Artificial Intelligence*, 2023.

47. Silver, D., Sutton, R. S. (in press). "Welcome to the Era of Experience." To appear in *Designing an Intelligence*, G. Konidaris, Ed. MIT Press.

64. Wan, Y., Naik, A., Sutton, R. "Average-reward learning and planning with options." *Advances in Neural Information Processing Systems 34*, 22758-22769, 2021.

65. Zhang, S., Wan, Y., Naik, A., Sutton, R. S., Whiteson, S., "Average-reward off-policy policy evaluation with function approximation." In *Proceedings of the International Conference on Machine Learning*, 2021.

66. Wan, Y., Naik, A., Sutton, R. S., "Learning and planning in average-reward Markov decision processes." In *Proceedings of the International Conference on Machine Learning*, 2021.

91. Mahmood, A. R., Sutton, R. S., Degris, T., Pilarski, P. M., "Tuning-free step-size adaptation." In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing,* Kyoto, Japan, 2012.

103. Sutton, R. S., Koop, A., Silver, D., "On the role of tracking in stationary environments," *Proceedings of the 24th International Conference on Machine Learning*, 2007.

128. Sutton, R. S. "Adapting bias by gradient descent: An incremental version of Delta-Bar-Delta," *Proceedings of the Tenth National Conference on Artificial Intelligence*, pp. 171–176. MIT/AAAI Press, 1992.

144. Wan, Y., Sutton, R. S. (2022). On convergence of average-reward off-policy control algorithms in weakly-communicating MDPs. *14th International OPT Workshop on Optimization for Machine Learning*.

151. Naik, A., Sutton, R. S., "Multi-step average-reward prediction via differential TD(λ)." In *Multi-disciplinary Conference on Reinforcement Learning and Decision Making*, 2022.

191. Sharifnassab, A., Salehkaleybar, S., Sutton, R. "Metaoptimize: A framework for optimizing step sizes and other meta-parameters." ArXiv:2402.02342, 2024

194. Dohare, S., Hernandez-Garcia, J.F., Rahman, P., Sutton, R. S., Mahmood, A.R. "Loss of plasticity in deep continual learning." ArXiv:2210.14361, 2023.

195. Sutton, R. S., Bowling, M., Pilarski, P. M. "The Alberta plan for AI research." ArXiv:2208.11173, 2023.

196. Sutton, R. S. "A history of meta-gradient: Gradient Methods for meta-learning." ArXiv:2202.09701, 2022

197. Sutton, R., "The bitter lesson." Influential blog entry, 2019.