Beyond Reward:
The Problem of Knowledge and Data

Richard S. Sutton

University of Alberta
Edmonton, Alberta, Canada

Intelligence can be defined, informally, as knowing a lot and being able to use that knowledge flexibly to achieve one's goals. In this sense it is clear that knowledge is central to intelligence. However, it is less clear exactly what knowledge is, what gives it meaning, and how it can be efficiently acquired and used. In this talk we re-examine aspects of these age-old questions in light of modern experience (and particularly in light of recent work in reinforcement learning). Such questions are not just of philosophical or theoretical import; they directly effect the practicality of modern knowledge-based systems, which tend to become unwieldy and brittle—difficult to change—as the knowledge base becomes large and diverse.

The key question for knowledge-intensive intelligent systems is 'What keeps the knowledge correct?' and there have been essentially three kinds of answers: 1) *people*—human experts understand the knowledge and ensure that it matches their beliefs, 2) *internal consistency*—the system checks that its knowledge coheres, and removes inconsistencies, and 3) *grounding in data*—the system compares its knowledge with external data in some way and changes it as needed to match the data. All of these are valid and often useful ways to maintain correct knowledge, but, in practice, relying on people to maintain correctness has been the dominant approach, supplemented by checks for internal consistency. This approach is well suited to taking advantage of existing human expertise, but ultimately limited in its ability to scale to very large knowledge bases because of its reliance on people. The essence of this approach is that knowledge is essentially *public*, describing a state of affairs in the world (separate from the intelligent system) that is at least potentially accessible to people. This might be called the *public-knowledge* approach.

In this talk we consider an alternative to the public-knowledge approach that is based on keeping knowledge correct by grounding it in data. Consider the case in which the data involved is the ordinary data available during the routine operation of the intelligent system without human intervention. This is the case of greatest interest because in it the system can correct and learn its knowledge autonomously, enabling scaling to very large knowledge bases (see Sutton 2009, 2001). If the system were a robot, this data would be simply whatever data was available through its sensors and about its motor actions. Knowledge grounded in such sensorimotor data may have no public semantics; it is tightly bound to the individual and need not be accessible to other observers in any useful way. Knowledge in this *sensorimotor* approach is essentially private, personal, and subjective, in contrast to the public-knowledge approach in which it is public,

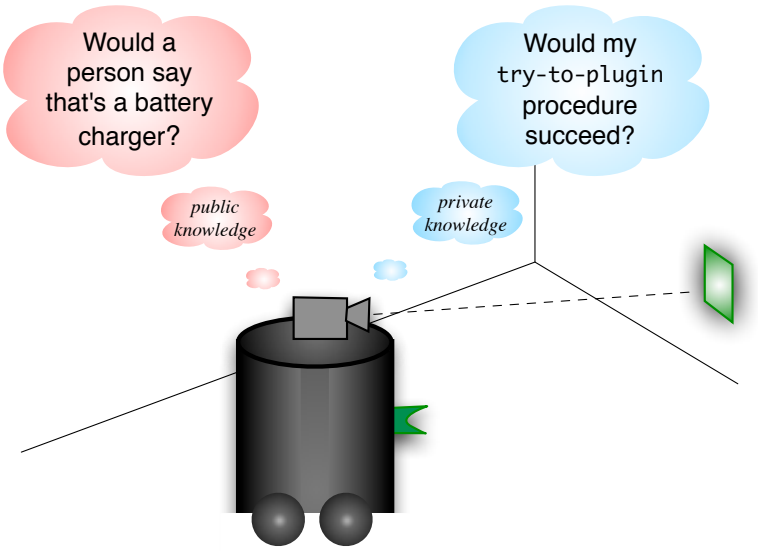

Fig. 1. A robot contemplates its camera image, trying to decide whether or not there is a battery charger on the wall. The thought bubbles on the left and right illustrate the difference between formulating this question in a public-knowledge way and in a sensorimotor-knowledge way. In the former, it is ultimately a question of what people would say, whereas, in the latter, it is question about the outcome of a sensorimotor procedure the robot could execute, in this case the procedure `try-to-plugin`, which is presumed to be some extended closed-loop procedure for trying to connect to a battery charger until success, with power trickling into the battery, or failure by running out of time.

universal, and objective. In the sensorimotor approach, knowledge is ultimately statements about the sensorimotor data stream that the system can check for itself, whereas, in the public-knowledge approach, knowledge is ultimately statements about entities in the world that can be checked by people but not typically by the system itself. An example of the contrast between the two approaches is suggested by Fig. 1.

The two approaches have different strengths. Public knowledge is easily communicated to and from people, and is naturally abstract and expressive, whereas sensorimotor knowledge is more easily maintained without human intervention. The latter is a key strength bearing directly on one of the most important problems facing modern knowledge based systems. A second motivation for exploration of the sensorimotor approach is that it is much less developed; there has been very little effort expended trying to extend it to encompass abstract and high-level knowledge. It is not clear if this can be done or even exactly what it might mean. In this talk I summarize recent work trying to explore the uncharted challenges of the sensorimotor approach to knowledge.

The first challenge to be addressed in pursuing the sensorimotor approach is the obvious mismatch between sensorimotor data, which is typically low-level, fine-grained, and rapidly changing, and knowledge, which is typically high-level, abstract, and persistent. Some abstraction can be achieved merely by introducing new terms or features corresponding to sets or regions of the sensorimotor data space. A more important conceptual innovation is dealing with *temporal* abstraction. The key solution idea here is expressing knowledge in terms of predictions about the outcome of temporally extended procedures. The basic idea is exemplified in Fig. 1, in which the outcome of the `try-to-plugin` procedure is used to learn to recognize battery chargers. This procedure might execute over several seconds or minutes before succeeding with charging or failing by giving up. To be able to predict which of these will occur, say from the visual image, is an important kind of knowledge, useful for planning and localization. It provides a way of abstracting over camera images, grouping many different images all into the class of those that predict success in plugging in. This general idea has been formalized in a theory of closed-loop macro-actions (a.k.a. "options") and of planning with predictive models of their outcomes, developed and used by many researchers over the last decade or so (e.g., Parr 1998; McGovern & Sutton 1998; Sutton, Precup & Singh 1999; Precup 2000; Stolle & Precup 2002; Mannor et al. 2004; Singh, Barto & Chentanez 2005; Rafols 2006; Koop 2007; Konidaris & Barto 2007)

Somewhat surprisingly, a second major challenge in pursuing the sensorimotor approach has been finding a sound algorithm for learning knowledge of the procedure-prediction form. On the surface the learning problem appears to be exactly that solved by reinforcement learning algorithms, particularly temporal-difference (TD) methods (e.g., see Sutton & Barto 1998). Just as TD methods can be used to learn to predict whether you will win a backgammon game if you try, they can also be used to predict whether you will succeed in plugging into the charger if you try. The problems are directly analogous but, surprisingly, there is a major technical problem applying standard reinforcement-learning algorithms such as Q-learning, TD(λ), or Sarsa in conjunction with function approximation. The convergence results for these methods do not apply when experience comes in incomplete trajectories. These methods are sound as long as one always plays backgammon games to conclusion, but if one processes incomplete games, then they may diverge (Baird 1995; Tsitsiklis & Van Roy 1997). In applications like recognizing the battery charger, it is important to be able to learn from incomplete trajectories. One wants to be able to guess that something is a charger, then take a few steps closer to confirm that guess, but not be obligated to go all the way and plug in. Learning from incomplete fragments of sensorimotor experience is essential to obtaining the key strength of the sensorimotor approach to knowledge, but it requires new reinforcement learning algorithms. Technically, it requires algorithms with the ability to learn from "off-policy" training. The search for off-policy TD algorithms has been ongoing since 1995, but has been essentially unsuccessful until the last couple of years when Hamid Maei, Csaba Szepesvári, and I developed a new family of learning

algorithms, called *gradient-TD* methods. These methods appear to make learning temporally-abstract knowledge from sensorimotor data practical for the first time (Sutton, Szepesvári & Maei 2009; Sutton et al. 2009; Maei 2011; Maei et al. 2009, 2010; Maei & Sutton 2010).

A final major aspect of our pursuit of the sensorimotor approach to knowledge has been to develop a significant in-house robotics effort. Robots have unambiguous sensorimotor data and force one to address issues of real-time computation, but can also have major practical drawbacks. Fortunately, in recent years the costs, difficulties, and overheads of using physical robots have come down substantially. In the last year we have been able to demonstrate a robot learning thousands of non-trivial, temporally-extended facts about its sensorimotor interface in real time without human training or supervision (Sutton et al. 2011; Modayil, White & Sutton 2012, Degris & Modayil 2012).

There is still a long way to go, but so far it appears possible to make steady progress in expanding the range and scale of knowledge that can be grounded in sensorimotor data and maintained by an intelligent system without human intervention.

Acknowledgements

This work was done with many colleagues, including particularly Anna Koop, Mark Ring, Joseph Modayil, Thomas Degris, Satinder Singh, Doina Precup, Csaba Szepesvari, Hamid Maei, Leah Hackman, David Silver, Mike Sokolsky, Patrick Pilarski, Marc Bellemare, and other members of the Reinforcement Learning and Artificial Intelligence (RLAI) laboratory. The RLAI laboratory is supported by Alberta Innovates – Tech Futures, the Alberta Innovates Center for Machine Learning, NSERC, Mprime, and the Glenrose Hospital.

References

Baird, L. C. (1995). Residual algorithms: Reinforcement learning with function approximation. In *Proceedings of the Twelfth International Conference on Machine Learning*, pp. 30–37.

Degris, T., Modayil, J. (2012). Scaling-up knowledge for a cognizant robot. In notes of the *AAAI Spring Symposium on Designing Intelligent Robots: Reintegrating AI*.

Konidaris, G., Barto, A. G. (2007). Building portable options: Skill transfer in reinforcement learning. *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, pp. 895–900.

Koop, A. (2007). *Investigating Experience: Temporal Coherence and Empirical Knowledge Representation*. University of Alberta MSc. thesis.

Maei, H. R. (2011). *Gradient Temporal-Difference Learning Algorithms*. University of Alberta PhD. thesis.

Maei, H. R., Sutton, R. S. (2010). GQ(λ): A general gradient algorithm for temporal-difference prediction learning with eligibility traces. In *Proceedings of the Third Conference on Artificial General Intelligence*.

Maei, H. R., Szepesvári, Cs., Bhatnagar, S., Precup, D., Silver, D., Sutton, R. S. (2009). Convergent temporal-difference learning with arbitrary smooth function approximation. In *Advances in Neural Information Processing Systems 22*. MIT Press.

Maei, H. R., Szepesvári, Cs., Bhatnagar, S., Sutton, R. S. (2010). Toward off-policy learning control with function approximation. In *Proceedings of the 27th International Conference on Machine Learning*.

Mannor, S., Menache, I., Hoze, A., Klein, U. (2004). Dynamic abstraction in reinforcement learning via clustering. *Proceedings of the Twenty-first International Conference on Machine Learning*.

McGovern, A., Sutton, R.S., (1998). Macro-actions in reinforcement learning: An empirical analysis. Technical Report 98-70, University of Massachusetts, Department of Computer Science.

Modayil, J., White, A., Sutton, R. S. (2012). Multi-timescale nexting in a reinforcement learning hobot. To appear in: *Proceedings of the 51st IEEE Conference on Decision and Control*.

Parr, R. (1998). *Hierarchical Control and Learning for Markov Decision Processes*. PhD thesis, University of California at Berkeley.

Precup, D. (2000). *Temporal Abstraction in Reinforcement Learning*. University of Massachusetts PhD thesis.

Rafols, E. J. (2006). *Temporal Abstraction in Temporal-difference Networks*. University of Alberta MSc. thesis.

Singh, S., Barto, A. G., Chentanez, N. (2005). Intrinsically motivated reinforcement learning. In: *Advances in Neural Information Processing Systems 17*, pp. 1281-1288.

Stolle, M., Precup, D. (2002). Learning options in reinforcement learning. *Abstraction, Reformulation, and Approximation*, pp. 212–223.

Sutton, R. S. (2001). "Verification" and "Verfication, the key to AI". Online at `http://richsutton.com/IncIdeas/Verification.html` and `http://richsutton.com/IncIdeas/KeytoAI.html`.

Sutton, R. S. (2009). The grand challenge of predictive empirical abstract knowledge. In: *Working Notes of the IJCAI-09 Workshop on Grand Challenges for Reasoning from Experiences*.

Sutton, R. S., Barto, A. G. (1998). *Reinforcement Learning: An Introduction*. MIT Press.

Sutton, R. S., Maei, H. R., Precup, D., Bhatnagar, S., Silver, D., Szepesvári, Cs., Wiewiora, E. (2009). Fast gradient-descent methods for temporal-difference learning with linear function approximation. *Proceedings of the 26th International Conference on Machine Learning*.

Sutton, R. S., Modayil, J., Delp, M., Degris, T., Pilarski, P. M., White, A., Precup, D. (2011). Horde: A scalable real-time architecture for learning knowledge from unsupervised sensorimotor interaction. *Proceedings of the 10th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*.

Sutton, R. S., Precup, D., Singh, S. (1999). Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artificial Intelligence 112*, pp. 181–211.

Sutton, R. S., Szepesvári, Cs., Maei, H. R. (2009). A convergent $O(n)$ algorithm for off-policy temporal-difference learning with linear function approximation. In *Advances in Neural Information Processing Systems 21*. MIT Press.

Tsitsiklis, J. N., and Van Roy, B. (1997). An analysis of temporal-difference learning with function approximation. *IEEE Transactions on Automatic Control 42*:674–690.