Constraints on Theories of Open-World Learning

Pat Langley
Information Technology and Systems Division
Institute for Defense Analyses
730 East Glebe Road
Alexandria, VA 22305

Abstract

In this paper, I examine challenges that arise in developing theories of open-world learning. After defining the problem, I review some theories from the history of chemistry, biology, geology, and AI, along with the importance of inductive bias to constrain the learning process. Classic cognitive architectures offer one source for such guidance, but their generality provides little aid on this front. Instead, I propose that more constrained architectures for embodied agents have greater potential, as they make commitments about the form of domain knowledge used to describe environments, as well as the processes that operate over them. In addition, I hypothesize that autonomous agents must include motivational structures that drive behavior and that changes to the environment can lead to their revision as well. I argue that a full account of open-world learning should make commitments about the structures and processes that underlie these capabilities.

Autonomous Agency in Open Worlds

Advances in sensors, effectors, memories, and processors have led to autonomous agents that are far more capable and common than those from only a decade ago. These take on many different forms, from self-driving cars and delivery drones to military robots and planetary rovers. The development of such systems typically relies on collection and processing of very large training sets to create accurate pattern recognizers and efficient controllers. This approach is viable for some applications and certain contexts, but it depends on two related assumptions: the environment will not change in important ways; and the agent's expertise will remain accurate and appropriate. Unfortunately, these postulates will not hold in many real-world settings.

We would like autonomous agents that are robust to such shifts. For example, consider an unmanned aerial drone on an exploratory mission in the Amazon rainforest. The system's expertise remains accurate and its behavior is acceptable until an airborne spider's web tangles one of its rotor blades, a large predatory bird attacks it from above, a strong updraft pulls it off course, a dense fog bank degrades its visibility, or high humidity causes intermittent shorts in its camera controller. A truly flexible autonomous agent would realize, in each case, that its expertise was outdated and adapt rapidly enough to still achieve the mission goals.

Scenarios of this sort raise the challenge of *open-world learning*, a class of problems introduced by DARPA's SAIL-ON program (Senator, 2019). We can state this problem as:

- *Given:* An agent architecture that operates in some class of tasks and environments;
- *Given:* Expertise that supports acceptable performance for these tasks and environments;
- *Given:* Limited experience after sudden, unannounced changes to the environment degrade performance;
- *Find:* When the environmental change occurs and what revised expertise will give acceptable performance.

This formulation applies to many agents, environments, and tasks, regardless of whether their initial expertise is handcrafted or learned from experience.

The problem of open-world learning addresses the very heart of what we mean by the term 'autonomous agent'. We say that an entity is an 'agent' if it carries out actions that affect its environment over time. However, in this light, teleoperated robots and remote-controlled drones would count as agents. We say that an agent is 'autonomous' if it operates independently and without supervision, but this can hold to different degrees. Thermostats are autonomous but only in a very narrow context. Robot vacuum cleaners have a broader range of behaviors but have been programmed by humans. Humans fall at the spectrum's extreme end, because they can not only adapt their model of the world but, in some cases, alter their own motivations and value systems.

Some readers may question why open-world learning poses a challenge, in that modern techniques for machine learning have been widely advertised as the solution to nearly any problem. However, remember that environmental shifts can be sudden and unannounced and that the agent must detect them and repair its expertise rapidly. The most widely adopted methods for classification learning, despite their success in some settings, rely on batch processing and require many labeled training cases, neither of which are sufficient here. Reinforcement learning, a popular approach to sequential action selection, typically requires many runs on a simulator, which will not be available for unfamiliar physical environments. In summary, mainstream approaches to machine learning are ill suited to such scenarios.

In the pages that follow, I discuss a number of issues related to theories of open-world learning. First, I review some familiar examples from the history of science, including artificial intelligence, and their lessons for accounts of learning. After noting the need for some form of inductive bias to make the search process tractable, I consider what cognitive architectures can offer to this end. Upon concluding that they offer only weak constraints, I turn to more specialized architectures for embodied agents that include theoretical commitments about knowledge of the physical environment. Finally, I argue that a complete theory of open-world learning must address not only how agents can alter their environmental models, but how they can alter their own motivations in response to such changes.

Theories of Open-World Learning

Given that we want the research community to develop theories of open-world learning, we should consider the form that such accounts might take. The history of science, including the early phases of artificial intelligence, offer compelling examples from which researchers can draw useful lessons. In this section, I review three classic theories from the disciplines of chemistry, biology, and geology, along with three others cases from the study of intelligent systems. In closing, I revisit some familiar ideas about the need for constraints in machine learning.

Scientific theories aim to explain observed phenomena in terms of a set of interconnected claims or assumptions. Here are three well-known examples:

- Dalton's (1808) atomic theory posited that macroscopic objects are made from tiny molecules, each involving atoms of nondecomposable elements. Moreover, chemical reactions transform some types of molecules into other types by rearranging their constituent atoms.
- Pasteur's (1880) germ theory of disease proposed that many illnesses are caused by small organisms that invade the body and attack it. In addition, these germs spread from one host to another through the process of infection.
- Hess's (1962) theory of plate tectonics stated that interlocking plates make up the Earth's surface, with mountains and deep sea trenches at their interfaces. These plates move very slowly under, over, and against each other to produce large-scale geological formations.

There are many analogous examples from the history of science, but we can draw some tentative conclusions from this set about the character of theories.

Despite their many differences in form and content, scientific theories nearly always include postulates that impose *qualitative constraints* on the domain under study. This holds even when the account also includes quantitative elements, which are often introduced after a field has agreed about their qualitative aspects. Moreover, theories posit both *structures* (e.g., entities and their relations) and *processes* that operate over and transform them. For instance, molecules and atoms are structures in the atomic theory, whereas chemical reactions are processes that affect them. Also, theories are *abstract* enough that they cannot be tested directly; this can only occur when one has added enough assumptions to produce operational *models*. For example, the atomic theory must be augmented by specific claims about the constituents of particular molecules, while germ theory requires associations between specific microorganisms and diseases. Finally, scientific theories regularly elaborate earlier ones with which they share assumptions, as the immune theory builds on the more basic germ theory.

We should also consider examples from the early days of artificial intelligence, which illustrate many of the same characteristics. These include three classic theories:

- *Physical symbol systems* (Newell & Simon, 1976), which posits that mental structures consist of symbols (persistent physical patterns) and symbol structures (organized sets of such symbols), which in turn can designate other entities or activities. This theory also proposes mental processes that create, modify, and interpret an evolving sequence of these symbol structures.
- *Production systems* (Newell, 1966), which elaborates on the first theory by postulating memories that contain sets of modular elements encoded as symbol structures, including a rapidly changing working memory and a more stable long-term store with condition-action rules. Processing involves repeatedly matching rules against elements in working memory and using them to alter its contents, which in turn enables new matches.
- *Heuristic search* (Newell & Simon, 1976), which also extends the first framework by declaring that problem solving relies on symbol structures to denote candidate solutions, generators of candidates, criteria for acceptance, and heuristics. This theory assumes processes for generating candidate solutions, testing them for acceptability, and using heuristics to guide choices.

These examples clarify that theories of intelligent behavior specify both structures and processes that operate over them. They also have the same abstract, qualitative character as the cases from chemistry, biology, and geology.

These observations are relevant for researchers who desire to develop theories of open-world learning. They suggest that such accounts make statements about the mental *structures* over which learning operates, especially how they represent the agent's experience and expertise. They also indicate that these theories should make commitments not only about the learning mechanisms that acquire the structures, but about the performance processes that use them to generate behavior. Theories of learning are seldom stated in isolation; they almost always incorporate assumptions about representation and performance (Langley, 1987).

However, such theories should also acknowledge a fact that has been recognized since the early days of machine induction: effective learning depends on some form of *inductive bias*. Most research in this field views learning as search through a space of hypotheses or models, but either this space or the manner in which one traverses it must be constrained in some manner. In some approaches, this bias places limits on the form or structure of candidate models, as occurs with naive Bayesian classifiers, support vector machines, and linear equations. Many techniques organize the search process in some way, as with top-down construction of decision trees and gradient descent through parametric neural networks. Typically, stronger inductive biases mean that fewer training cases are needed to acquire acceptable models, which is crucial to effective open-world learning. Of course, one advantage is that agents need not acquire their models from scratch; they can adapt or revise existing expertise, which itself provides a strong bias that is reliable given that the environmental changes are piecemeal. However, other constraints appear necessary[1] and the next two sections consider two forms that they might take.

Cognitive Architectures

One natural place to turn for inductive bias is the literature on *cognitive architectures* (Langley, Laird, & Rogers, 2009; Langley, 2017), which are computational theories for intelligent systems that operate over time. Not all open-world agents need be instances of such a framework, but agents that are stated in such terms would inherit their many assumptions, which would in turn impose constraints on how they adapt to environmental change. Classic research on cognitive architectures had strong connections to psychological findings, but this is not a defining characteristic; the important feature for our purposes is that they make strong assumptions about the nature of the mind.

Briefly, a cognitive architecture is a theory about infrastructure for intelligent systems that specifies which facets of cognition remain unchanged across different domains and tasks. These typically include the memories that store domain content, the representation of such content, and the processes that create, access, and modify these elements. However, it does *not* specify the particular content, which can change across domains and over time. A standard analogy is with architectures for buildings, which specify the layout of floors, rooms, and passages between them, but not the furniture or occupants, which may vary. A typical cognitive architecture also provides a programming language with a high-level syntax that reflects its theoretical assumptions about representation and processing.

As noted earlier, most frameworks in this paradigm incorporate key ideas from cognitive psychology. These include postulates that: short-term memories, which change rapidly, are distinct from long-term ones, which change slowly; both types of memories contain modular elements that are encoded as symbol structures; long-term elements are accessed by matching them against structures in short-term memories; cognition involves the dynamic composition of mental structures to create new ones; and learning is a monotonic process that is interleaved with performance. Two well-known examples are ACT-R (Anderson & Lebiere, 1998) and Soar (Laird, 2012), but Langley et al. (2009) discuss others, some with more distant connections to human cognition.

Given that cognitive architectures are intended as theories of intelligent systems, might they offer an inductive bias that would aid effective open-world adaptation? They place some constraints on the mechanisms of learning, namely that it must involve the incremental, piecemeal acquisition of knowledge elements and that it must be interleaved with the performance that it improves. They also constrain the *form* of acquired expertise, say as a collection of condition-action rules. However, few cognitive architectures make strong claims about the *type* of content that populates memories and that learning mechanisms generate. Indeed, the Common Model of Cognition (Laird, Lebiere, & Rosenbloom, 2017) does not even include a theoretical distinction between beliefs and goals. This follows from researchers' desire to provide *general* accounts of intelligence that apply in many settings, but I maintain that it goes too far in this direction. To offer the inductive bias needed for open-world learning, we must look to frameworks that focus on embodied agents operating in physical environments.

Content-Laden Architectures

We have established that constraints are necessary to make open-world learning effective and that traditional cognitive architectures, despite making strong assumptions about mental representations and processes, do not suffice. Limiting attention to physical settings holds considerable promise, but we must guard against being overly specific, since we desire theories that are as general as possible. We can achieve this aim by adopting constraints not about specific environments, but about their generic characteristics.

This suggests that we assume agents' environmental models include certain types of mental structures. For instance, we might postulate that long-term knowledge contains:

- *Concepts*, which define categories of objects based on their observed features or generic relations among objects based on their spatial configurations;
- *Maps*, which specify physical places in terms of objects and their layouts, along with spatial relations among such places (e.g., in topological networks);
- *Skills*, which describe the conditional effects of volitional agent actions, along with skill complexes that combine them into organized activities; and
- *Processes*, which specify natural mechanisms that alter objects or places over time, as well as networks of processes that make up causal chains.

Such mental structures would be *descriptive* in the sense that they characterize the environment, whether accurately or not.[2] They make no prescriptive statements about whether classes of situations or events are *desirable*.

An embodied agent also needs short-term structures to describe past, present, and current situations and events. One common simplifying assumption is that each such dynamic element is an instance of some long-term knowledge structure. For example, the agent might encode a belief that it

[1] Unfortunately, the phrase 'open-world learning' has led some researchers to conclude that the problem must be entirely unconstrained, but this does not follow and it is unrealistic.

[2] As Langley (2020) has proposed, an agent might also encode models of *fields* with attribute values that vary over a spatial region, such as fluid flow or magnetic attraction, but such content might also be stored with mental maps.

is located between a rock and a tree as instances of the relational concept *between* and the object concepts *rock* and *tree*. Similarly, it might represent its movement from place A to place B along route R as an instance of a skill for traversing that route. The long-term and short-term memories need not use the same symbols to denote content, but this assumption simplifies both performance and learning.

A content-laden architecture promises to impose substantially greater constraints on the space of environment models than classic frameworks, but it must still incorporate some of their features. The most important characteristic is that it should provide a programming language with an associated syntax for stating mental structures. A key difference is that this language would have distinct constructs for concepts, spatial knowledge, agent skills, and natural processes. Together, the notation for these types of cognitive structures would define a space of models for physical environments while remaining very general. They would provide a form of *declarative bias* (Ade, De Raedt, & Bruynooghe, 1995), which specifies the forms that models can take and thus would limit search during learning considerably.

However, it is important to note that declarative structures, by themselves, remain ambiguous. They cannot drive intelligent systems until they are joined with mental processes that interpret them. Thus, a complete content-laden architecture for physical agents would include mechanisms for:

- *Categorization and conceptual inference*, which matches conceptual knowledge against percepts to create beliefs;
- *Place recognition and localization*, which compares spatial knowledge with percepts to identify agent location;
- *Mental simulation of process networks*, which generates trajectories of expected situations over time; and
- *Physical execution of skill complexes*, which carries out volitional agent actions in the environment.

Competing content theories will propose different mechanisms for interpreting cognitive structures. For example, one architecture might posit that concepts match in an all-or-none fashion, whereas another might support partial matching to various degrees. Such choices will have implications for how models align with observations.

Of course, the theory must also postulate structures and processes that support learning. For instance, an open-world learner must distinguish between short-term elements that it infers from perceptions and ones that it predicts with its environmental model. The agent must compare these structures to detect *anomalies*, say by using a mechanism similar to those for monitoring plan execution (e.g., Langley et al., 2016). If the agent deems an anomaly sufficiently different, then it initiates a process of model revision that produces one or more *hypotheses* about how to change the model. Finally, the theory must specify how the agent evaluates these alternatives, chooses among them, and uses the selected candidates to generate an improved model. Neither anomalies or hypotheses introduce any new declarative bias, as they involve the same types of knowledge structures – concepts, maps, processes, and skills – as does the performance system that interacts with the environment.

Motivated Agency in Open Worlds

These observations take us part way toward the inductive bias needed for effective open-world learning, but they omit an important factor. We desire not only systems that form accurate models to predict events, but agents that engage in goal-directed volitional activities. Such behavior relies not on *descriptive* knowledge structures but on *prescriptive* ones that let agents decide not what will happen but what they should do or seek. This indicates the need for another form of mental content. Many AI planning systems are given goals to achieve or tasks to carry out, but these are typically concrete and problem specific. We require instead generic structures that describe the conditional values or utilities of situations and activities. Langley et al. (2016) refer to such knowledge elements as *motives*, but other labels have also appeared in the AI literature (e.g., Hanheide et al., 2010).

Naturally, an architectural theory must also incorporate processes that operate over such motivational content, say to introduce concrete goals and calculate their values. Work on goal reasoning (Aha, Cox, and Muñoz-Avila, 2013), which builds on classic frameworks for plan generation, has proposed mechanisms for this purpose. In this light, planning techniques are also prescriptive processes, since they determine how an agent carries out search to achieve goals. Interestingly, neither the introduction of motivational structures or the processes that interpret them would seem to offer further inductive bias beyond that provided by concepts, places, processes, or skills. They are essential to a complete theory because they are needed to drive behavior, but they do not obviously constrain search during open-world learning.

Nevertheless, their central role in agent decision making raises an important issue. Motives are prescriptive knowledge elements that refer to a descriptive model of the environment, but they are internally defined. When the world changes, updating this model will help the agent achieve high-value goals, but such shifts can have other implications. Let us return to the initial scenario of an aerial drone on an exploratory mission. Some unexpected changes, such as damage to rotors, might reduce the agent's range or reaction time, making it unable to achieve some mission objectives. In such cases, should the agent alter its motives and thus the values it assigns to situations and activities?

This certainly happens in humans, arguably the most autonomous agents on the planet. If an Olympic runner hurts a knee badly enough that he no longer has a realistic chance of winning races, does he continue to compete in the same circles? Or does he change his aspirations and learn to find value in other activities, such as playing a sport like golf that does not require running or mentoring young athletes who might carry on his tradition? We maintain that a complete theory of open-world learning should cover such internal changes to agent motivations and aspirations. Indeed, they appear even more central to understanding the nature of autonomous agency than revisions to environmental models.

This is unexplored territory, but we can outline some mechanisms that might support such internal restructuring. For instance, if an agent finds that it can no longer achieve certain performance levels, then it might lower its expectations and become satisfied with lesser ones, but if changes let it do better than before, then its aspirations might increase. More substantially, an agent might revise the situations or activities that motivate it. Such shifts might result from updating values associated with existing motives or refining their activation conditions. Acquisition of new concepts or skills could enable novel motives that link to them, and this process might be influenced by lateral transfer from motives for similar structures. Another source might be the imitation of motives inferred from the behavior of other agents.

Radically autonomous agents that can alter their own motives, and thus how they compute values, raise difficulties for evaluation. In such cases, one cannot specify *external* metrics for success, as the agent can determine its own criteria, even if they were initialized by a human developer. We can measure the trajectory for the agent's values over time, but there is some danger that it would simply assign high values to all situations. One option is to require that changes to the agent's motives and aspirations be gradual, with large arbitrary jumps being forbidden. This might be enough to ensure 'reasonable' behavior provided the environment changes slowly enough, but we might need other constraints. Either way, a full theory of open-world learning should address the potential for agents to alter their motivations, which could have major implications for behavior.

Concluding Remarks

In this essay, I defined the problem of open-world learning and clarified why it poses a challenge for existing paradigms. After this, I discussed the need for theories of this process and the form they might take, drawing on analogies with earlier scientific accounts, including ones from artificial intelligence. I also reviewed the notion of inductive bias and its central role in constraining the space of models considered during learning. Next, I discussed cognitive architectures as a possible source of such bias but concluded that their emphasis on generality makes them poorly suited for this end.

Instead, I argued that content-based architectures, which make stronger commitments about types of knowledge and processes that operate over them, hold much greater potential to constrain learning. Finally, I suggested that open-world agents should incorporate the ability to alter their motives in response to environmental change, which in turn raises issues about how to evaluate them. Although the paper introduced a number of important questions, it offered no definitive answers. Nevertheless, its observations may help guide future research on the task of open-world learning.

Acknowledgements

This research was supported by the Defense Advanced Research Projects Agency (DARPA). The views, opinions, and/ or findings expressed are those of the author and should not be interpreted as representing the official views or policies of the Department of Defense or U.S. Government. Approved for public release, distribution unlimited.

References

Ade, H., De Raedt, L., & Bruynooghe, M. (1995). Declarative bias for specific-to-general ILP systems. *Machine Learning*, *20* 119–154.

Aha, D. A., Cox, M. T., & Muñoz-Avila, H. (Eds.) (2013). *Goal Reasoning: Papers from the ACS Workshop*. Baltimore, MD.

Anderson, J. R., & Lebiere, C. (1998). *The atomic components of thought*. Mahwah, NJ: Lawrence Erlbaum.

Dalton, J. (1808). *A new system of chemical philosophy* (Part 1). London, UK: R. Bickerstaff.

Hanheide, M., Hawes, N., Wyatt, J., Gobelbecker, M., Brenner, M., Sjoo, K., Aydemir, A., Jensfelt, P., Zender, H., & Kruijff, G-J. M. (2010). A framework for goal generation and management. *Proceedings of the AAAI-2010 Workshop on Goal-Directed Autonomy*. Atlanta, GA.

Hess, H. H. (1962). *History of ocean basins*. In A. E. J. Engel, H. L. James, & B. F. Leonard (Eds.), *Petrologic studies: A volume to honor A. F. Buddington*, 599–620. Boulder, CO: Geological Society of America.

Laird, J. E. (2012). *The Soar cognitive architecture*. Cambridge, MA: MIT Press.

Laird, J. E., Lebiere, C., & Rosenbloom, P. S. (2017). A standard model for the mind: Toward a common computational framework across artificial intelligence, cognitive science, neuroscience, and robotics. *AI Magazine*, *38*, 13–26.

Langley, P. (1987). Research papers in machine learning. *Machine Learning*, *2*, 195–198.

Langley, P. (2017). Progress and challenges in research on cognitive architectures. *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence* (pp. 4870–4876). San Francisco: AAAI Press.

Langley, P. (2020). Open-world learning for radically autonomous agents. *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence* (pp. 13539–13543). New York, NY: AAAI Press.

Langley, P., Barley, M., Meadows, B., Choi, D., & Katz, E. P. (2016). Goals, utilities, and mental simulation in continuous planning. *Proceedings of the Fourth Annual Conference on Cognitive Systems*. Evanston, IL.

Langley, P., Choi, D., Barley, M., Meadows, B., & Katz, E. P. (2017). Generating, executing, and monitoring plans with goal-based utilities in continuous domains. *Proceedings of the Fifth Annual Conference on Cognitive Systems*. Troy, NY.

Langley, P., Laird, J. E., & Rogers, S. (2009). Cognitive architectures: Research issues and challenges. *Cognitive Systems Research*, *10*, 141–160.

Newell, A. (1966). *On the analysis of human problem solving protocols*. Technical Report, Department of Computer Science, Carnegie Institute of Technology, Pittsburgh, PA. Reprinted in J. C. Gardin & B. Jaulin (1968), *Calcul et formalisation dans les sciences de l'homme*, 146–185. Paris.

Newell, A., & Simon, H. A. (1976). Computer science as empirical enquiry: Symbols and search. *Communications of the ACM*, *19*, 113–126.

Pasteur, L. (1880). On the extension of the germ theory to the etiology of certain common diseases. *Comptes rendus, de l'Academie des Sciences*, *XC*, 1033–44.

Senator, T. E. (2019). Science of AI and learning for open-world novelty (SAIL-ON). Presented at the Proposers' Day Meeting. DARPA: Arlington, VA.