Advances in Cognitive Systems 10 (2023) 3–14 Submitted 4/2023; published 8/2023

Cognitive Systems and Theories of Open-World Learning

Pat Langley PATRICK.W.LANGLEY@GMAIL.COM
Institute for Defense Analyses, 730 East Glebe Road, Alexandria, VA 22305 USA
Center for Design Research, Stanford University, Stanford, CA 94305 USA

Abstract

In this essay, I examine challenges that arise in developing theories of open-world learning. After defining the problem, I review some theories from the history of both natural science and AI, along with the importance of inductive bias as a source of constraints in learning. Classic cognitive architectures offer one source for such guidance, but their generality means they offer little aid on this front. Instead, I argue that more constrained architectures for embodied agents have greater potential, as they make commitments about the types 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 environmental changes can lead to their revision as well. I claim that a full account of open-world learning should make commitments about the structures and processes that underlie these capabilities.

1. Autonomous Agency in Open Worlds

Advances in sensors, effectors, memories, and processors have led to autonomous agents that are far more capable and more common than those 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's controller. A radically autonomous agent would realize, in each case, that its expertise was outdated and adapt rapidly enough to survive and achieve the mission goals. Such a cognitive system would display the same flexibility and responsiveness as humans exhibit in similar situations.

Scenarios of this sort raise the challenge of *open-world learning*, a problem posed by Senator (2019), Langley (2020), and others. 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 degrade agent performance;
- *Find:* What changes have occurred and what revised expertise gives acceptable performance.

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

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, but in this light teleoperated robots and remote-controlled drones would qualify. 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. People fall at the spectrum's extreme end, because they adapt not only their world models but also, in some cases, 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; moreover, the agent must detect them and repair its expertise rapidly. Yet 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 issues related to theories of open-world learning, but the need for such theories merits some justification. In principle, we could create software artifacts that demonstrate this ability with no explicit theoretical account of their operation. However, without principles to guide the design and construction of such systems, we would have no confidence that they will behave as intended. Other fields of engineering benefit greatly from such principles, which place constraints on the design and construction of large-scale structures like bridges and mechanical devices like bicycles. The design of intelligent agents, including ones that must learn effectively in open worlds, would be aided by similar theoretical frameworks. Once available, developers could use these principles each time they construct a cognitive system for such a setting.

2. Computational Theories of 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. Scientific theories aim to explain observed phenomena in terms of a set of interconnected claims, postulates, or assumptions. For instance, Dalton's (1808) atomic theory stated that everyday objects are made from tiny molecules that comprise atoms of nondecomposable elements, whereas chemical reactions transform some types of molecules into others by rearranging their atoms. Similarly, Pasteur's (1880) germ theory of disease proposed that illnesses are caused by small organisms invading the body and that these germs spread across hosts through a process of infection. There are many analogous cases throughout the history of science from which we can draw tentative conclusions about the character of theories.

Despite many differences in form and content, scientific theories nearly always include postulates that impose *qualitative constraints*. This holds even when the account has quantitative elements, as these are typically introduced after a field has agreed about qualitative issues. 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 some scientific theories from the early days of AI and cognitive systems research that illustrate the same characteristics. Here are three examples that Langley (2018) has discussed at greater length:

- *Physical symbol systems* (Newell & Simon, 1976), which claims that mental structures consist of symbols (persistent physical patterns) and symbol structures (organized sets of 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, 1973), which elaborates on the first theory by postulating memories that contain sets of modular elements encoded as symbol structures, including a dynamic working memory and a long-term store with condition-action rules. Processing involves matching rules against working memory and altering its contents, which in turn enables new matches.
- *Heuristic search* (Newell & Simon, 1976), which also extends the first framework by declaring that problem solving uses symbol structures to denote candidate solutions, generators of candidates, criteria for acceptance, and heuristics. This theory incorporates processes for generating candidate solutions, testing them for acceptability, and using heuristics to guide search.

In summary, theoretical accounts of intelligent behavior also specify both structures and processes that operate over them, and they have the same abstract, qualitative character as theories that have been developed in the natural sciences.

In a similar manner, computational theories of learning should make statements about the mental structures over which they operate, especially how they represent experience and expertise. Moreover, they should make commitments not only about these structures' acquisition, but about how performance processes use them to generate behavior. Theories of learning seldom occur in isolation; they usually incorporate assumptions about representation and performance (Langley, 1987). The reason is that we typically define learning as improvement in performance on some class of tasks, which we can characterize as the use of knowledge or expertise to pursue those tasks. Learning has no effect without a performance element that takes advantage of it, and performance requires access to stored expertise. Thus, performance is downstream from knowledge, which constrains its operation, and learning is downstream from performance. This implies that theories of open-world learning should make commitments about representation of knowledge and processes that use it.

3. Inductive Bias and Open-World Learning

Such theories of learning should acknowledge a fact that has been recognized for decades: effective learning depends on some form of *inductive bias* (Utgoff, 1984). Most AI research views learning as search through a space of hypotheses or models, but this space is potentially infinite and we need some way to limit or guide navigation through it. The computation time required to find viable candidates is an issue, but more important is the need to guard against overfitting models to the training data, which leads to poor performance on new cases.[1] This issue arises whether one carries out search through a space of discrete structures, such as rules or decision trees, or through a parameter space, such as weights in a neural network. In both cases, we require ways to constrain or direct the search for hypotheses, and stronger inductive biases mean that fewer training cases are needed to acquire acceptable models, which is crucial for successful open-world learning. Researchers have explored three general ways to incorporate such inductive biases.

One approach places limits on the *form of candidate models*, such as naive Bayesian classifiers, one-layer perceptrons, and linear equations, rather than allowing richer representations, such as decision trees and multi-layer neural networks. More constrained formalisms have less expressive power but result in smaller search spaces, which reduces the chances of overfitting the training cases; however, if they are too limiting, then they can only approximate the target model. The framework of ‘probably approximately correct’ learning (Valiant, 1984) examines this issue with the tools of complexity analysis. In contrast, the statistics community has studied the tradeoff between an induction method’s ‘bias’ (asymptotic error) and its ‘variance’ (error due to limited samples). Both reveal that more constrained representations give better results when only small training sets are available, which holds for the task of open-world learning, as defined earlier.

A second approach to inductive bias relies on the *organization of search*. For instance, standard techniques for decision-tree induction rely on greedy search, starting with an empty tree and recursively adding tests with associated branches, which produces a bias toward simpler structures. In contrast, methods for learning neural networks start with a set of random weights and pursue gradient descent through the parameter space, but they often include a term that drives many weights to zero, which imposes a different ordering over the hypothesis space. The heuristics that guide search also introduce an inductive bias, so that different evaluation functions lead to different models when combined with the same algorithm. These insights are all relevant to open-world learning, but no more so than in many other settings.

A third source of inductive bias is *knowledge*. In one powerful version of this idea – *theory revision* – the learner has a reasonably good model and need only modify it to handle new anomalous observations. This paradigm is directly relevant to open-world learning, which assumes that agents do not acquire their models from scratch, but rather that they adapt or revise their existing expertise when necessary. The approach provides a very strong inductive bias that is effective when the agent’s current model is complex and when the environmental changes require only small, piecemeal revisions to this stored content.

[1] This concern is related to Hume’s (1739) *problem of induction*, which questions how we can know that the generalizations we draw from experience are correct. The notion of inductive bias offers a practical response to this challenge without providing the logical guarantees desired by many philosophers.

However, not all expertise is domain specific, and more generic varieties can also aid the learning process. Such knowledge does not refer to domain-level predicates, but instead places constraints on the *types* of content that the agent can access, as well as how it should interpret information encoded in this manner. This concerns not the representational formalism, such as decision trees or neural networks, but rather the kinds of content that the agent acquires and how these kinds relate to each other. This approach to inductive bias has received little attention from the machine learning community, but it is very relevant to open-world settings.

4. Cognitive Architectures

One natural place to turn for the final type of 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. Briefly, a cognitive architecture specifies which facets of cognition remain unchanged across different domains and tasks. These 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 a building architecture, which specifies 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.

Most frameworks in this paradigm incorporate their key ideas from cognitive psychology. These include postulates such as: 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. Many cognitive architectures, including ACT-R (Anderson & Lebiere, 1998) and Soar (Laird, 2012), are elaborations on the *production systems* framework ourlined earlier. Learning involves creation of new production rules or revision of numeric annotations based on experience (Klahr, Langley, & Neches, 1987). Early research on production systems focused on high-level cognitive tasks like problem solving, but they have since been connected to simulated environments (e.g., Jones et al., 1999) and physical robots (e.g., Trafton et al., 2013), which requires grounding them in perception and action.

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 for learning, namely that they must involve the incremental, piecemeal acquisition of knowledge elements and that they must be interleaved with the performance they improve. 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 a desire to provide *general* accounts of intelligence, but I maintain that it goes too far in this direction. To offer the inductive bias needed for open-world learning, we must focus more seriously on frameworks for embodied agents that operate in physical environments.

5. Content-Laden Architectures

We have established that constraints are necessary to make open-world learning effective and also that traditional cognitive architectures, despite making 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 theoretical constraints not about specific environments, but rather about their generic characteristics.

This response is reminiscent of proposals by philosophers like Aristotle and Kant that people do not enter the world with innate content, but that they are endowed with abstract types or templates, which they later populate with content based on experience. The idea is reflected in AI research on planning (Ghallab, Nau, & Traverso, 2004), which distinguishes between *states*, which are composed of environmental relations, *goals*, which specify a set of desired states, and *operators*, which describe the conditional effects of actions on states. The standard notation for operators is similar to that for production rules and the formalism for states is much like that for short-term memories, but planning systems make stronger commitments than production systems, in that their structures denote situations and activities in the physical world rather than purely mental entities. This commitment has important implications for open-world learning, as it means that discrepancies between expectations and observations can drive detection and adaptation.

Of course, one can imagine a cognitive architecture that incorporates strong constraints about the nature of its environment and the structures that it uses to represent them. For instance, such a framework might postulate that long-term memory includes:

- *Concepts*, which define categories of objects in terms of their observed features (e.g., vehicles, buildings) or generic relations among objects based on their configurations (e.g., lines of vehicles, blocks of buildings). These would provide the terms that the agent uses to describe states of the environment it encounters over time.
- *Maps*, which specify locations, places, and regions in the physical environment in terms of objects and their layouts, along with spatial relations among such places (e.g., in topological networks). These could also encode *fields* with attribute values that vary over a spatial region, such as wind velocity or radiation level.
- *Skills*, which describe how the agent should set one or more control attributes (e.g., linear or angular force) as a function of the current state (e.g., distance from a target) to achieve an objective (e.g., reaching the target). Composite skills would specify how to combine simpler skills to produce higher-level organized activities.
- *Processes*, which specify natural mechanisms (e.g., acceleration, temperature exchange) that alter attribute values of objects (e.g., velocity, temperature) or maps as a function of the current environmental state. These would occur at different rates that depend on both constant parameters and changing attribute values.
- *Constraints*, which indicate what situations or activities can or cannot occur in the environment (e.g., that two objects cannot occupy the same location), as well as structural or numeric conditions (e.g., that an object is solid) under which they apply. These would limit the set of possible states that can arise in the agent's environment.

Most of these mental structures are *descriptive* in the sense that they aim to characterize the environment and its behavior, accurately or not.[2] Skills have a *prescriptive* aspect in that they specify how the agent should respond to the situation in which it finds itself, although we will see later that other factors can also come into play.

In addition, an embodied agent needs short-term structures to describe past, present, and current situations and events. One simplifying assumption is that each such dynamic element must be an instance of some long-term knowledge structure (e.g., Choi & Langley, 2018). For example, the agent might encode a belief that it 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.

Langley and Katz (2022) have described PUG, a cognitive architecture that incorporates many of these assumptions about mental structures, but there are certainly other ways to organize an agent's knowledge about the environment. For example, Fox and Long (2006) present an extension to the widely adopted PDDL formalism, which many AI planning systems use to describe domains. Their framework distinguishes between *operators*, which describe the conditional effects of an agent's actions, and a different encoding of natural processes, which describe the effects of nonvolitional mechanisms. Although it remains unclear which scheme will best support open-world learning, they both provide content-laden architectures that promise to impose substantially greater constraints on the space of environment models than classic frameworks.

Like traditional architectures, a content-laden framework provides a programming language with an associated syntax for stating mental structures. A key difference is that it offers distinct constructs for content like concepts, spatial knowledge, agent skills, and natural processes. Together, the notation for such cognitive elements defines a space of models for physical environments while remaining very general. They provide a form of *declarative bias* (Ade, De Raedt, & Bruynooghe, 1995) that limits search during learning considerably. However, most work on this topic provides domain-specific knowledge to constrain induction, which hinders its application. To ensure generality, open-world learning should incorporate more abstract declarative encodings, like that in work on logical analyses of temporal reasoning (e.g., Allen, 1983), spatial cognition (e.g., Cohn et al., 1997), and causal inference (e.g., Nam & Baral). Similar generality occurs in the upper levels of ontologies for common-sense reasoning (e.g., Spear, Ceusters, & Smith, 2016).

However, declarative structures, by themselves, are 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 inference*, which matches concepts against percepts to create beliefs;
- *Place recognition*, which compares spatial knowledge with percepts to identify agent location;
- *Mental simulation of processes*, which generates expected trajectories of situations over time;
- *Physical execution of skills*, which carries out volitional actions to achieve agent ends; and
- *Constraint interpretation*, which eliminates alternatives in inference, planning, and execution.

[2] Agents can use the same types of structures to encode models of other agents' mental states, as in research on collaborative AI and dialogue systems (e.g., Gabaldon, Langley, & Meadows, 2014).

Competing content theories will propose different mechanisms for processing such 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 environmental observations.

Of course, the theory must also postulate structures and processes that support learning. For instance, an open-world learner should distinguish between short-term beliefs that it infers from perceptions and ones that it predicts with its environmental model. When these two disagree, then the agent must encode these relationships as *anomalies* that indicate problems with its knowledge base (Dannenhauer et al., 2021; Muñoz-Avila et al., 2019). These in turn will lead to *hypotheses* about how this expertise may be incorrect, each of which will suggest possible changes to the agent's model of the environment. There may be multiple hypotheses for a given anomaly, so each one may have an associated score that indicates its ability to explain the unexpected situations or events.

To support general open-world learning, the architectural mechanisms that generate and interpret these short-term structures should be domain independent. They should include processes that let the agent respond to unexpected changes by:

- Detecting anomalies by comparing its model's predictions to its observations, using methods similar to those for monitoring plan execution;
- If an anomaly is deemed sufficiently important, generating hypotheses (e.g., through causal analysis) for how to change its model to address the error; and
- Evaluating hypotheses from this candidate set, selecting among them, and using the selected candidate to revise its knowledge base in response.

Neither the detection of anomalies or the generation and evaluation of hypotheses introduce new declarative bias, for they refer to the same types of knowledge structures as the performance system. But these processes can take advantage of representational assumptions that encode such a bias.

In addition, it seems likely that open-world learners will need separate mechanisms to detect anomalies and repair models for each of type of expertise. For example, revising conceptual knowledge might rely on something as simple as detecting mismatches between observed objects and known categories. Similarly, noting flaws in mental maps and updating them might be reasonably straightforward. In contrast, identifying and repairing problems with volitional skills (e.g., Benson, 1995) and natural processes (e.g., Arvay & Langley, 2016) will be more challenging because state trajectories can be influenced by multiple factors, so that understanding anomalies will require credit assignment. An important hurdle will be handling discrepancies that involve multiple types of knowledge, which in turn will require their joint identification and revision.

6. 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 agents that form accurate models to predict events, but ones that engage in goal-directed volitional activities. Many AI planning systems are given goals to achieve, but these are typically concrete and problem specific. The same holds for the knowledge encoded in hierarchical task networks, in skills that have associated control equations, and in reactive systems that use expected values to select actions. Agents that must survive on long-term missions in complex environments also require mental structures that move beyond domain-specific concerns to encode more general prescriptions. This indicates the need for another form of mental content. The psychology literature often refers to such elements as *motives* and posits them as a primary factor in determining human behavior.

Work on goal reasoning (Aha, 2018), which builds on classic frameworks for plan generation, is centrally concerned with representing and using such motivational content. For example, Langley and Katz's (2022) PUG framework represents motives as rules that specify the conditional values or utilities of certain relations, and others have used different labels to denote similar structures (e.g., Choi & Langley, 2018; Hanheide et al., 2010). Naturally, such efforts also specify mechanisms that operate over motivational structures, say by introducing concrete goals that the agent should pursue and calculating their associated values. Motives are essential to a complete theory of intelligence because they drive behavior, but they also provide another target for open-world learning, and assumptions about their representation place constraints on this process.

The paradigm of reinforcement learning usually assumes that rewards come from the environment, but this idea has always rested on thin ice. Motives, which we might view as separate elements of a factored reward function, are *internal* cognitive structures and, in people at least, they can shift over time. They provide prescriptive knowledge that links to a descriptive model of the environment, and changes to the latter can lead human-like agents to revise the former. Let us return to the initial scenario of an aerial drone on an exploratory mission. Suppose that unexpected changes, such as damage to rotors, reduce the agent's range or speed, making it unable to achieve mission objectives that are distant or urgent. In such cases, should the agent alter its motivational structures and thus assign different values to situations and activities?

This certainly happens in humans, arguably the most autonomous agents on the planet. If an Olympic runner injures 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 follow in his footsteps? I maintain that a complete theory of open-world learning should cover such internal changes to an agent's motivations in response to environmental shifts, including those in its own body. Indeed, they appear even more central to understanding the nature of autonomous agency than the revisions to environmental models that research on open-world learning has emphasized to date.

An important, but seldom discussed, facet of intelligent agency is Simon's (1956) notion of *satisficing*: An agent halts decision making when it finds an option that it considers 'good enough', which Simon linked to the agent's *aspiration level*. Many satisficing analyses assume that aspirations are constant, but a radically autonomous agent might alter them. This territory is poorly explored, but we can outline some mechanisms that might support such internal restructuring:

- If an agent finds it can no longer achieve certain performance levels, then it might lower its expectations and become satisfied with lesser ones, as in the drone example. If changes let it do better than before, then its aspirations might increase and lead it to more audacious pursuits.
- If an agent can no longer achieve certain types of aims, then it might revise the situations or activities that drive it. This could occur by refining the activation conditions for motives based on what it can still accomplish or revising their utility functions based on outcome quality.

- If an agent acquires new concepts or skills, then it can use them to construct novel motives, possibly influenced by lateral transfer from similar structures. Another source might be motives inferred from other agents' behavior, as in work on imitation learning, but at a deeper level.

Mechanisms of this sort could let radically autonomous agents update their motivational knowledge in response to environmental changes, giving them more accurate views of what they can and cannot accomplish and thus which goals and activities are practical for them to pursue.

However, endowing agents with the ability to alter their own motives, and thus how they compute values, raises difficulties for evaluation. We cannot invoke traditional *external* performance measures because the agent determines its own criteria for success, even when those are initialized by a human developer. We could measure the trajectory for the agent's computed utilities over time, but there is danger that it would simply assign the same values to all situations. This could lead to complacent agents that are satisfied with whatever transpires or depressed agents that view all situations and events as undesirable. One option is to allow only gradual changes to motives and aspirations, with large arbitrary jumps being forbidden. This might be enough to ensure 'reasonable' behavior provided the environment changes slowly enough, but we must first define what this term means for adaptive systems. Either way, a full theory of open-world learning should address the potential for agents to alter their motivations and its implications for behavior.

7. Concluding Remarks

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

Instead, I argued in favor of content-laden architectures for embodied agents, which make stronger commitments about the types of knowledge stored in memory and the processes that operate over them. These constrained architectures retain domain independence but, by providing more details about how agents describe and reason about their environment, they offer a stronger inductive bias and hold greater potential to guide learning. Finally, I suggested that open-world agents should alter their motives in response to changes by revising the objectives they consider important and their associated aspiration levels. However, such extreme adaptivity in turn raises issues about how to evaluate behavior when agents can define for themselves what constitutes success.

The essay introduced some important questions that arise in the study of open-world learning and suggested different ways that the community might address them, but the aim was not to offer final answers. Nevertheless, I hope that the observations about scientific theories, inductive bias, cognitive architectures, content-based constraints, and motivated agency will help guide future research on adaptive cognitive systems that must operate in open worlds. Science itself involves the exploration of a changing landscape, filled with surprises that require us to reconsider and revise cherished assumptions before we can advance our understanding.

Acknowledgements

This essay incorporates feedback from Dustin Dannenhauer and expands on a paper presented at the Spring Symposium on Designing Artificial Intelligence for Open Worlds, reporting research funded by the Defense Advanced Research Projects Agency. The views, opinions and/or findings are the author's and should not be interpreted as representing the official views or policies of the Department of Defense or the US 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. W. (2018). Goal reasoning: Foundations, emerging applications, and prospects. *AI Magazine*, *39*, 3–24.

Allen, J. F. (1983). Maintaining knowledge about temporal intervals. *Communications of the ACM*, *26*, 832–843.

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

Arvay, A., & Langley, P. (2016). Heuristic adaptation of quantitative process models. *Advances in Cognitive Systems*, *4*, 207–226.

Benson, S. (1995). Inductive learning of reactive action models. *Proceedings of the Twelfth International Conference on International Conference on Machine Learning* (pp. 47–54). Tahoe City, CA: Morgan Kaufmann.

Choi, D., & Langley, P. (2018). Evolution of the ICARUS cognitive architecture. *Cognitive Systems Research*, *48*, 25–38.

Cohn, A. G., Bennett, B., Gooday, J., & Gotts, M. M. (1997). Qualitative spatial representation and reasoning with the Region Connection Calculus. *GeoInformatica*, *1*, 275–316.

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

Dannenhauer, D., Muñoz-Avila, H., & Cox, M. T. (2021). Expectations for agents with goal-driven autonomy. *Journal of Experimental & Theoretical Artificial Intelligence*, *33*, 867–889.

Fox, M., & Long, D. (2006). Modelling mixed discrete-continuous domains for planning. *Journal of Artificial Intelligence Research*, *27*, 235—297.

Gabaldon, A., Langley, P., & Meadows, B. (2014). Integrating meta-level and domain-level knowledge for task-oriented dialogue. *Advances in Cognitive Systems*, *3*, 201–219.

Ghallab, M., Nau, D., & Traverso, P. (2004). *Automated planning: Theory and practice*. San Francisco, CA: Morgan Kaufmann.

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.

Hume, D. (1739). *A treatise of human nature*. Oxford, UK: Oxford University Press.

Jones, R. M., Laird, J. E., Nielsen P. E., Coulter, K., Kenny, P., & Koss, F. (1999). Automated intelligent pilots for combat flight simulation. *AI Magazine*, *20*, 27–42.

Klahr, D., Langley, P., & Neches, R. (Eds.) (1987). *Production system models of learning and development*. Cambridge, MA: MIT Press.

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, CA: AAAI Press.

Langley, P. (2018). Theories and models in cognitive systems research. *Advances in Cognitive Systems*, *6*, 3–16.

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., & Katz, E. P. (2022). Motion planning and continuous control in a unified cognitive architecture. *Proceedings of the Tenth Annual Conference on Advances in Cognitive Systems*. Arlington, VA.

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

Muñoz-Avila, H., Dannenhauer, D., & Reifsnyder, N. (2019). Is everything going according to plan? – Expectations in goal reasoning agents. *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence* (pp. 9823–9829). Honolulu, HI: AAAI Press.

Nam, T. H. & Baral, C. (2004). Encoding probabilistic causal model in probabilistic action language. *Proceedings of the Nineteenth National Conference on Artificial Intelligence* (pp. 305–310). San Jose, CA: AAAI Press.

Newell, A. (1973). Production systems: Models of control structures. In W. G. Chase (Ed.), *Visual information processing*. New York, NY: Academic Press.

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.

Simon, H. A. (1956). Rational choice and the structure of the environment. *Psychological Review*, *63*, 129–138.

Spear, A., Ceusters, W., & Smith, B. (2016). Functions in Basic Formal Ontology. *Applied Ontology*, *11*, 103–128.

Trafton, J. G., Hiatt, L. M., Harrison, A. M., Tamborello, F., Khemlani, S. S., & Schultz, A. C. (2013). ACT-R/E: An embodied cognitive architecture for human robot interaction. *Journal of Human-Robot Interaction*, *2*, 30–55.

Utgoff, P. E. (1984). *Shift of bias for inductive concept learning*. Doctoral dissertation, Department of Computer Science, Rutgers University, New Brunswick, NJ.

Valiant, L. (1984). A theory of the learnable. *Communications of the ACM*, *27*, 1134–1142.