Rational Agency and Radical Autonomy in Open Worlds

Pat Langley

Information and Technology Systems Division
Institute for Defense Analyses
Alexandria, Virginia

This work was supported by the Institute for Defense Analyses. Many of the ideas were inspired by DARPA's SAIL-ON Program, but it is not responsible for them. Thanks to David Aha, Dan Shapiro, Herb Simon, and others for useful discussions.

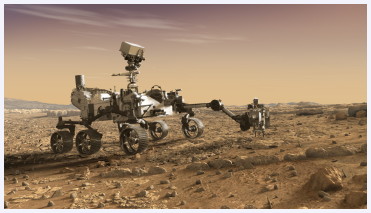
Overview: Some Questions

- When is an agent rational? How is it related to goal reasoning?
- What is open-world learning? How can an agent constrain it?
- What are cognitive architectures? Can they constrain learning?
- What are content theories? Can they aid open-world learning?
- What are motives? Can motives change in an open world?
- How might we evaluate agents that change their motives?

Autonomous Systems: Progress and Limits

Autonomous agents are becoming more common and impressive in the form of:

- Self-driving cars
- Delivery drones
- Military robots
- Planetary rovers



However, these systems depend on two critical assumptions:

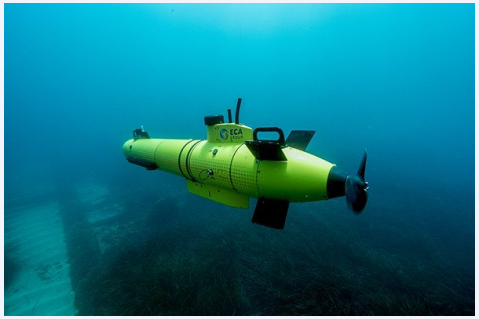
- The environment will not change in substantial ways
- Their agent's expertise will remain accurate and appropriate

These postulates will not hold in many real-world settings.

A Radically Autonomous Agent

Consider an unmanned underwater vehicle in a coastal area.

The system's expertise is accurate and its behavior good until:



- Unfamiliar kelp fouls its propellers
- A large unknown predator attacks it
- A mysterious current drags it off course
- Sonar becomes distorted in a low-visibility area
- A nearby volcano causes novel corrosive reactions

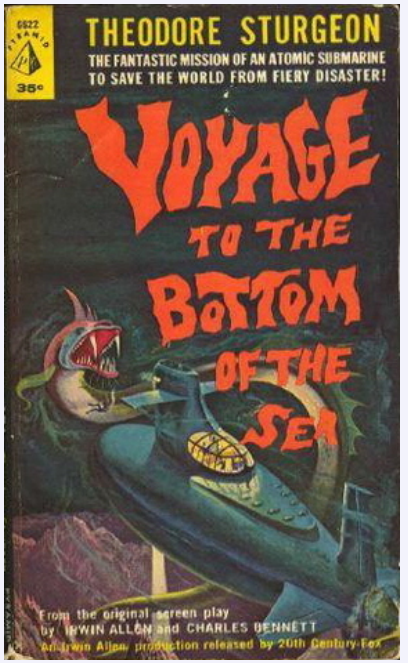
A radically autonomous system would realize these fall outside its expertise and learn rapidly enough to continue its mission.

A Radically Autonomous Agent

Consider an unmanned underwater vehicle in a coastal area.

The system's expertise is accurate and its behavior good until:

- Unfamiliar kelp fouls its propellers
- A large unknown predator attacks it
- A mysterious current drags it off course
- Sonar becomes distorted in a low-visibility area
- A nearby volcano causes novel corrosive reactions



The agent would be as responsive and adaptable as the crew of the *Seaview* in *Voyage to the Bottom of the Sea*.

Rational Agents

Naturally, we want embodied agents that respond rationally to such environmental changes.

But what does this really mean? According to Newell (1980):

- *An agent is **rational** if it selects actions that it believes will help achieve its goals.*

Thus, rationality does ***not*** require optimal choice, which Simon (1993) has argued is often ill defined.

But this framework says nothing about the *origin* of goals that guide an agent's behavior.

Simon refers to this latter aspect of cognition as ***nonrational*** decision making.

Rationality and Goal Reasoning

Fortunately, the subfield of goal reasoning (Aha et al., 2013) offers accounts of goals' origins. Examples include:

- *Creation of subgoals through backward chaining*
- *Changing goal priorities as the situation changes*
- *Generation of new goal instances using knowledge*
- *Acquiring / revising motivations through learning*

These different responses support increasingly more radical forms of agent autonomy.

The most interesting, and least explored, activity deals with ***changes to the agent's motivations***.

Satisficing and Aspirations

An important, but seldom discussed, element of goal reasoning is Simon's (1956) notion of ***satisficing***:

- *An agent halts problem solving when it finds an option that it considers good enough.*

This requires that we provide a criterion for 'good enough', which Simon linked to the agent's ***aspiration level***.

Most satisficing analyses assume that aspirations are constant, but a radically autonomous agent might alter them.

We will return to this idea later, to discuss the conditions under which it might occur.

Open-World Learning

Now let us return to the problem of *open-world learning*, which we can specify as:

- *Given:* An agent architecture that can operate in some class of environmental settings
- *Given:* Expertise that supports acceptable performance in these environments
- *Given:* ***Limited*** experience after ***sudden***, ***unannounced*** changes to the environment degrade performance
- *Find:* ***When*** the environmental change occurs and ***what*** revised expertise gives acceptable performance

This formulation applies to many agents and situations, whether initial expertise is handcrafted or learned.

The Technical Challenge

But why is this a challenge? Modern learning techniques can do almost anything, right? ***Unfortunately, no***.

Remember environmental shifts are *sudden* and *unannounced*, and expertise repair must be *rapid*.

- Statistical supervised induction? X
  - *Nonincremental, requires too many labeled cases*
- Reinforcement learning? X
  - *Requires far too many trials, no simulator available*

Mainstream approaches are ill suited for open-world learning.

The Need for Inductive Bias

To make open-world learning tractable, we must constrain its operation without overly restricting it.

In machine learning terms, we must provide a strong *inductive bias* that limits search through the model space.

- Hypothesis: *A theory of physical environments and transforms over them is necessary for effective open-world learning.*

Such a theory would reduce the time needed to detect changes and to repair the agent's expertise in response.

Note: A theory can be *domain independent* and very general yet still provide strong constraints.

Theories of Environmental Change

What form should a theory of environment change take?

It might be procedural and implicit, but this has disadvantages; in contrast, a *declarative* theory would explicitly specify:

- Elements of environments that agents may encounter
  - With each environment being a collection of such elements
- Transformations that can alter these environments
  - Composable operators that alter some elements but not others

Such specification would define a *space* of environments and possible trajectories through the space.

In classic machine learning, this is known as a *declarative bias*.

Cognitive Architectures

A *cognitive architecture* (Newell, 1990) is an infrastructure for intelligent systems that:

- Specifies those facets of cognition that remain *constant* across different domains;
- Including memories and representations of elements in those memories, but *not* their content, which changes over time;
- Provides a *programming language* with high-level syntax that reflects its theoretical assumptions.

A cognitive architecture moves beyond isolated capabilities, as it aims to provide a *unified* account of the mind.

Assumptions of Cognitive Architectures

Most cognitive architectures incorporate key postulates from psychological theories:

- *Short-term* memories are distinct from *long-term* stores
- Memories contain *modular* elements cast as *symbol structures*
- Long-term structures are accessed through *pattern matching*
- Cognitive processing occurs in *retrieval/selection/action cycles*
- Cognition involves *dynamic composition* of mental structures
- Learning is *monotonic* and *interleaved with performance*

Many architectures share these assumptions. ***Might they provide an inductive bias for open-world learning?***

Architectures and Open-World Learning

Unfortunately, typical cognitive architectures provide very few constraints on what can be learned.

- They specify how to encode short-term and long-term elements
- But they make few commitments about what elements describe

E.g., the *Common Model of Cognition* does not mention ***goals*** as a special type of mental structure.

- The framework allows goals, but they have no special status

To support the declarative bias necessary to guide open-world learning, we need ***content theories*** of embodied agency.

Computational Theories of Learning

Learning involves acquisition of expertise in order to *improve performance* on some class of tasks.

Theories of learning postulate mechanisms responsible for this improvement, but also make commitments about:

- *Mental structures* on which learning operates, especially the representation of experience and expertise;
- *Performance processes* that operate over those structures, not only the mechanisms that acquire them.

Accounts of learning are seldom stated in isolation; they often build on theories of representation and performance.

Can we imbue them with content that provides inductive bias?

Embodied Agency: Mental Structures

Agents that operate over time in external environs incorporate long-term (stable) structures:

- *Conceptual knowledge* – Categories of objects, relations, events
- *Motivational knowledge* – Value / desirability of situation classes
- *Procedural knowledge* – How to achieve desirable changes

Such systems also manipulate short-term (dynamic) structures:

- *Percepts / Beliefs* – Specific instances of conceptual structures
- *Goals / Values* – Desired beliefs and their importance to agent
- *Intentions* – Instances of procedures to be carried out

Although still general, these commitments focus on the ***content*** an embodied agent accesses and generates.

Embodied Agency: Mental Processes

Intelligent systems that operate over time in external environs also incorporate processes for:

- *Inference*, which matches conceptual knowledge to percepts and existing beliefs to generate new beliefs.
- *Goal processing*, which compares motivational knowledge to percepts and beliefs to activate and deactivate goals.
- *Plan generation*, which constructs / revises plans (sets of linked intentions) that should produce situations that satisfy goals.
- *Plan execution*, which carries out plans and their intentions in the external environment.

Each mechanism operates over the results of earlier ones in a closed-loop meta-cognitive cycle.

Embodied Agency: Open-World Learning

We can extend the theory of embodied agency to open-world learning by assuming new cognitive structures:

- *Anomalies* – Observations that are inconsistent with models
- *Hypotheses* – Candidate reasons that anomalies have occurred

We can also posit new processes that operate on these structures:

- *Detecting anomalies* by comparing models to observations
- *Generating hypotheses* to explain causes of detected anomalies
- *Evaluating* and *selecting hypotheses* from this candidate set
- *Revising knowledge structures* based on selected hypotheses

Agents can draw on this updated knowledge in future situations to improve their performance.

The PUG Architecture: Structures

PUG is an architecture for embodied agents (Langley et al., 2016) that incorporates three types of knowledge:

- *Concepts* – Define generic relations among objects and their associated numeric attributes
- *Motives* – Specify when goal instances should become active and their utilities
- *Skills* – Encode actions' immediate effects and their final results under specified conditions

States, plans, and search trees organize short-term memory elements into larger configurations.

PUG makes far stronger content assumptions than do many cognitive architectures.

The PUG Architecture: Processes

The architecture involves three different processing levels, each with its own cognitive cycle:

- Inferring beliefs, goals, and associated utilities
- Mentally executing candidate intentions and trajectories
- Carrying out heuristic search through a space of plans

Planning builds on mental execution, which in turn builds on inference, much as in ICARUS (Choi & Langley, 2018).

An extension, PUG/X, supports plan execution, monitoring, and replanning on detecting anomalies.

The architecture does not exhibit open-world learning, but it includes many necessary building blocks.

Open Worlds, Motives, and Aspirations

Open worlds require agents to revise not only concepts and skills, but also their motives and aspirations.

- Consider an Olympic athlete who loves to compete in races.
- However, an accident leaves him with a serious limp in one leg.
- On realizing that he can no longer win competitions, he may:
  - Become satisfied to *finish* races rather than winning them;
  - Shift to a *new sport* like golf that does not require running.

Similar changes can reduce the abilities of autonomous vehicles, which must then adapt in response.

A complete theory of open-world learning should cover such changes to agent motives and aspirations.

Mechanisms for Motivational Change

Radical autonomous agents should adapt to changes in ability.

How might an embodied agent alter its aspiration levels?

- *Increasing / decreasing them based on received value*
- *Learning conditions on what values are achievable*

How might an embodied agent revise what motivates it?

- *Update value functions on existing motives*
- *Refining conditions on existing motives*
- *Creating motives for new concepts or skills*
- *Inferring the motives of others and imitating them*

These processes would be purely internal and thus would run counter to classic notions of reinforcement learning.

Challenges for Evaluation

Agents that can alter their own motives, and thus how they compute values, pose challenges for evaluation.

- We cannot specify *external* metrics for success, as the agent determines its *own* criteria.
- We can measure the agent's trajectory of values over time, but what keeps it from giving high value to *everything*?

One response is to require that changes in agent motives and aspirations be *gradual* and *incremental*.

- The agent could only alter its value system slowly, with large, arbitrary jumps being forbidden.

We can even imagine convergence proofs on this scheme, at least if environments / abilities change slowly.

Summary: Some Answers

- When is an agent rational? How is it related to goal reasoning?
  - *Goals and goal reasoning are **central** to rational behavior.*
- What is open-world learning? How can an agent constrain it?
  - *When world changes unexpectedly. **Inductive bias** is essential.*
- What are cognitive architectures? Can they constrain learning?
  - *Unified theories of the mind. They offer **limited** constraints.*
- What are content theories? Can they aid open-world learning?
  - *Theories about mental content. They offer **stronger** constraints.*
- What are motives? Can motives change in an open world?
  - Motives are the ***source*** of goals / values. They ***can*** change.
- How might we evaluate agents that change their motives?
  - That remains unclear, but it must be on the ***agent's own terms***.

References

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

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

Langley, P. (2020). Open-world learning for radically autonomous agents. *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence*. 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.

Newell, A. (1982). The knowledge level. *Artificial Intelligence*, *18*, 87–127.

Newell, A. (1990). *Unified theories of cognition*. Cambridge, MA: Harvard University Press.

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

Simon, H. A. (1993). Decision making: Rational, nonrational, and irrational. *Educational Administration Quarterly*, *29*, 392–411.