The Alberta Plan for AI Research

Rich Sutton

with Mike Bowling and Patrick Pilarski

University of Alberta & DeepMind Alberta

Alberta Machine Intelligence Institute

Reinforcement Learning and Artificial Intelligence Lab

Outline (take-home messages)

- Understanding intelligence is a **grand scientific prize** that may soon be within reach; the **glory will go to the bold**
- The Alberta Plan is a direct run at this great prize
  - emphasizing intelligence as the **acquisition of new knowledge**
- My motivation for the plan stems from the **Oak** architecture, a complete conceptual design for a **Model-based RL agent**, a **Proto-AI**
- The Alberta Plan is a massive **retreat**; going back to the most basic algorithms; **re-vamping them for continual and meta-learning** (in 12 steps)

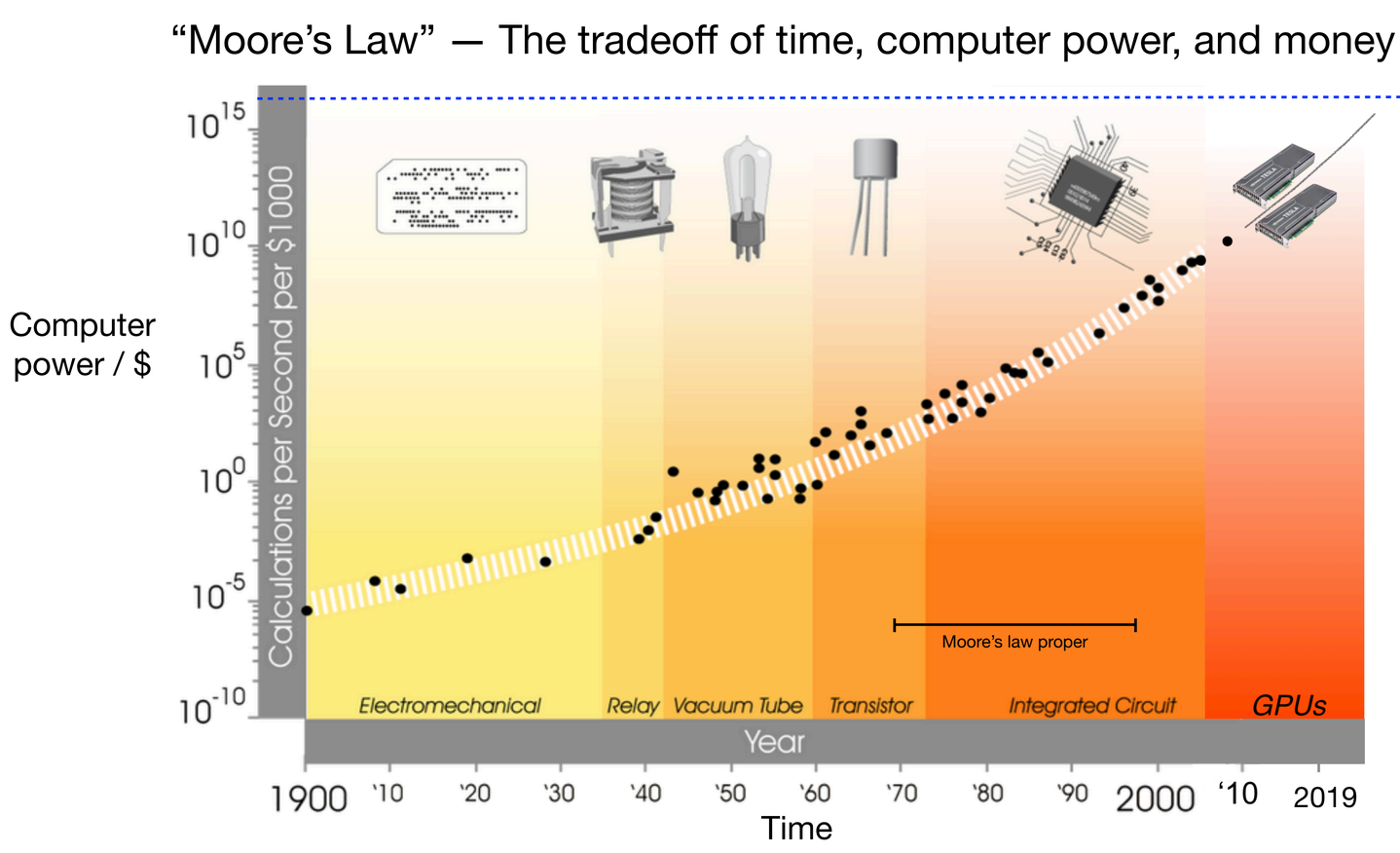
RL should keep its *Eyes on the Prize*

- The Prize is **to understand the principles of intelligence**—what it is and how it works—well enough to create (or become) beings of far greater intelligence
- The Prize is a fundamental goal of **science**, **engineering**, and the **humanities**
- Achieving it will change the way we work and play, our sense of self, and the goals we set for ourselves and our societies
- Achieving it will be comparable in significance to **the rise of life on Earth**
- It's significance is **not primarily derivative** from the benefits it will bring to humans

To understand intelligence is a great and glorious scientific goal!

Like the contributions of Einstein, Darwin, Newton, Copernicus, or Watson & Crick;
and less like the contributions of Guttenberg, Edison, Babbage, Page & Brin

Computer power/$ increases exponentially, with no end in sight, creating a powerful persistent pressure for understanding intelligence



≈Brain-scale computer power will cost $1000 in 2030

This estimate is rough but robust: a factor of 10 ≅ 5 years

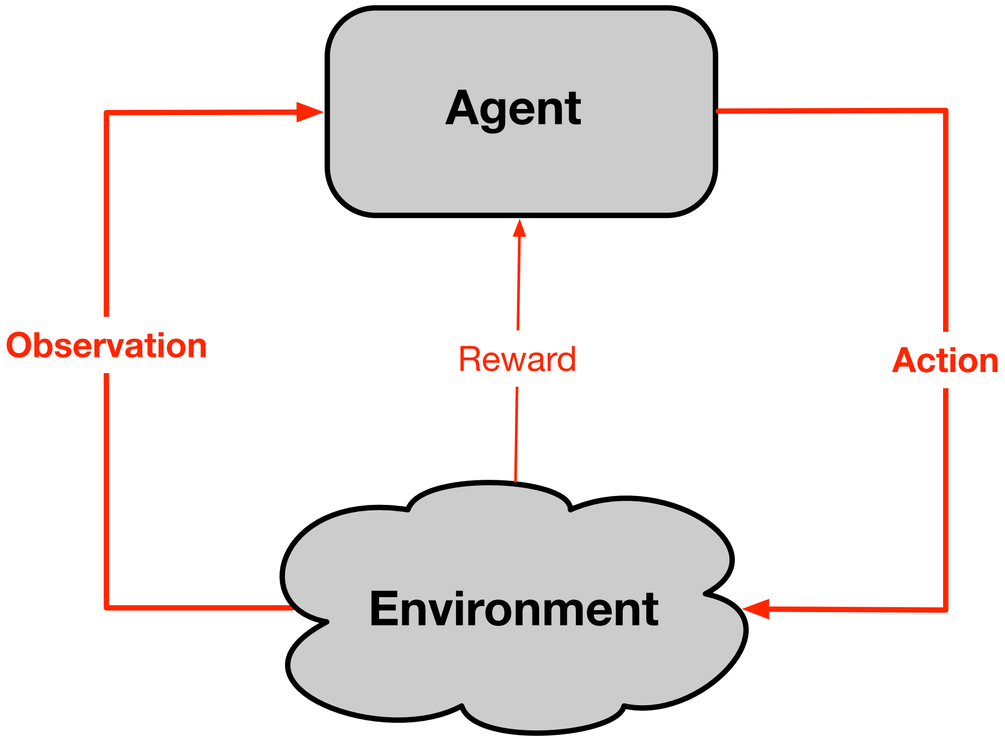
⇒ AI increases in value by a factor of 10 every 5 years

And thus so does the pressure to find the scalable AI algorithms

I estimate a 25% probability of human-level AI by 2030

[from Kurzweil AI]

The Alberta Plan takes the conventional RL interface seriously



- A complex computational agent interacts with a *vastly more complex* environment to maximize its reward
  - optimality is impossible
  - approximation is required
  - tracking is required

The "Big World" Perspective

- The interaction is continual and temporally uniform
  - there are no special training periods
  - there are no episodes, no discounting
- Actions and observations are just signals with no intrinsic meaning; they are just information
  - no language, vision, space, objects, or geometry

The Alberta Plan eschews domain knowledge

- The Plan should contain nothing specific to the human world
  - nothing about vision, space, objects, language, or other agents
  - nothing that distinguishes people from other animals
  - nothing about specific ecological niches
- No content, just algorithms for acquiring and organizing knowledge

Learning first, domain knowledge last.
Why?

To keep the problem small, the methods scalable
To keep the focus on acquiring new knowledge

The Alberta Plan's **Research vision**: Intelligence as signal processing over time

Four key features of DeepMind Alberta's AI research

1. Experience oriented. Experience is the data of AI
2. Online, temporally uniform, continually running & learning
3. Concern for computational cost, response time
4. Agents interacting with environments containing other agents

I have recently proposed that *many different fields* share common ideas about agenthood [RLDM-2022]

- Psychology
- Control theory
- Artificial intelligence
- Economics
- Neuroscience
- Operations research

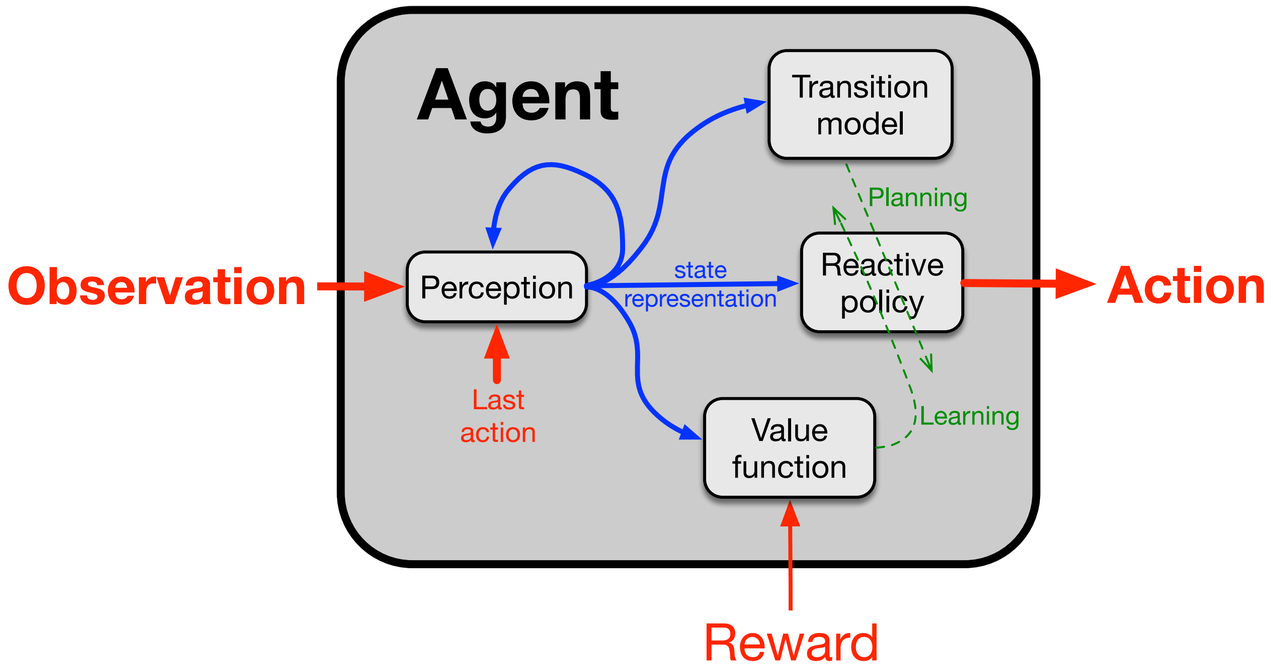
All involve d*ecision making over time to achieve a goal*

I proposed a **Common Model of the Intelligent Agent**

It's a good place for us to start in finding common ground

The common model of the intelligent agent

Common to RL, psychology, control theory, economics, neuroscience, operations research…



The agent comprises four components:

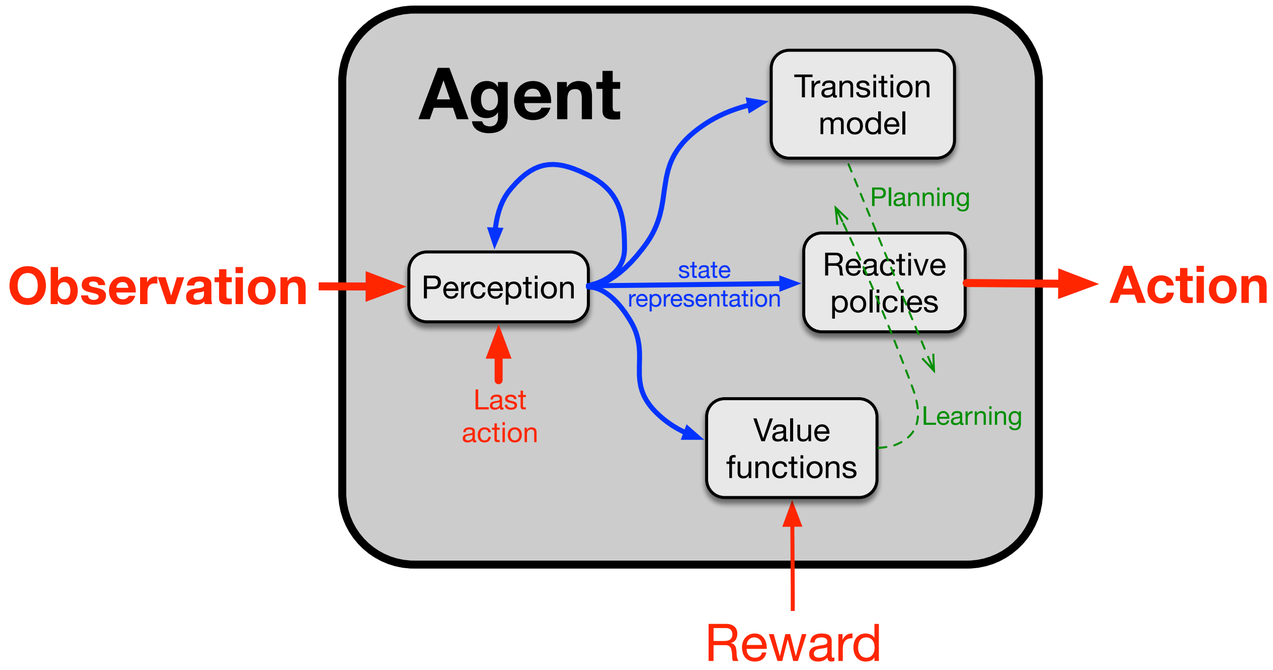
Perception produces the state representation used by all components

Reactive Policy quickly produces an action appropriate to the state

Value Function evaluates how well things are going, and changes the policy (learning)

Transition model predicts the consequences of alternate actions, and changes the policy (planning)

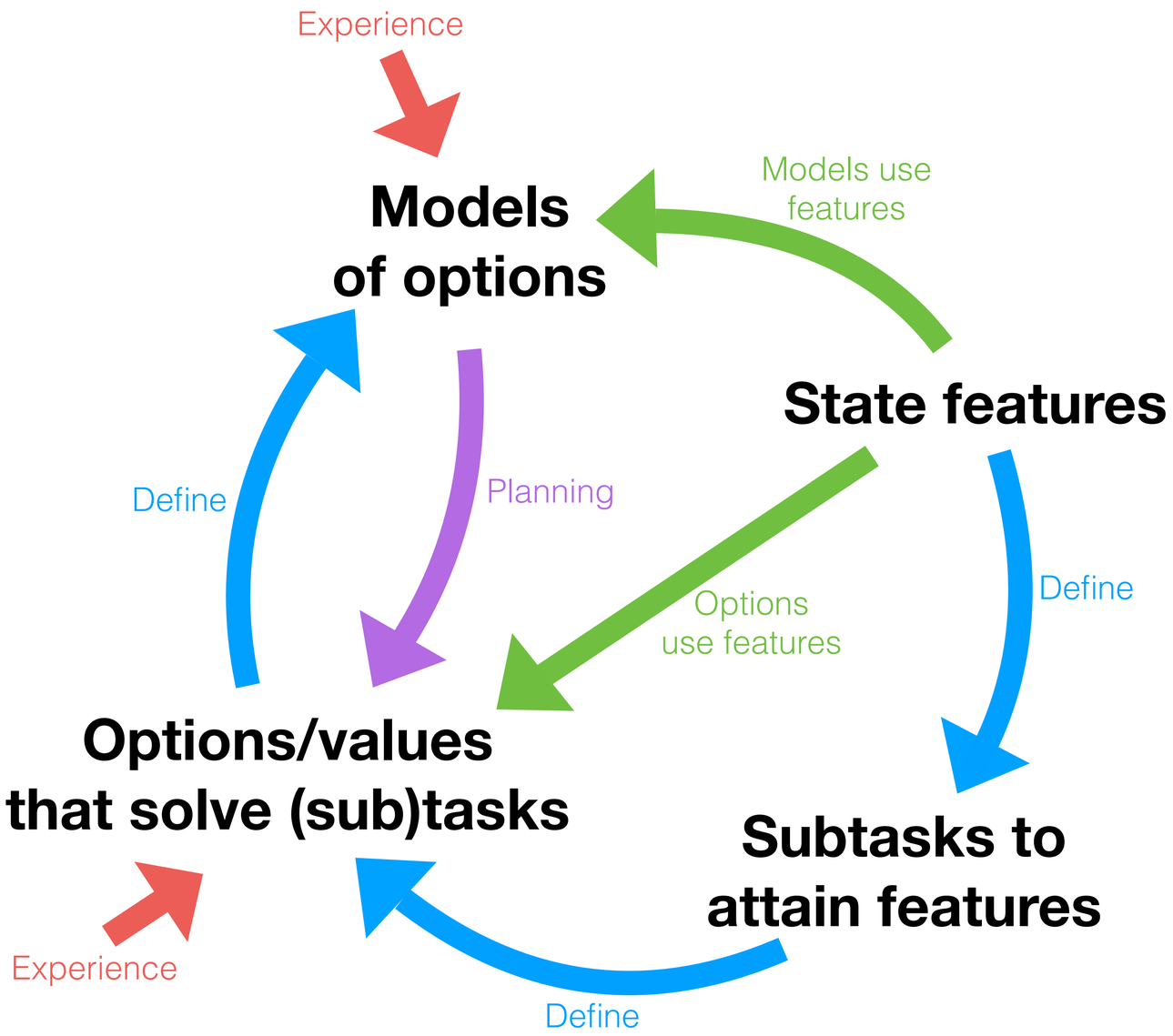
The Alberta Plan's model of the intelligent agent

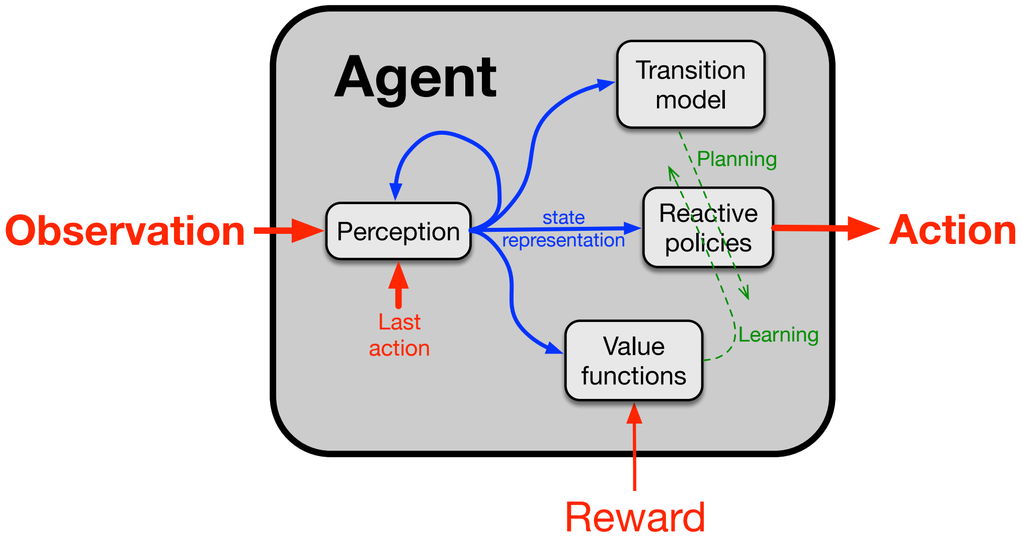


Adds SubTasks (≈feature-control subtasks)

Multiple policies and value functions

The **Oak** Architecture (Options And Knowledge) is based on a **cycle of discovery**





- Arrows show the direction of primary flow
  - Each arrow has a slower *backward* flow of credit
  - "I use you, so don't change too much"
- The consumers of state and time abstractions support, evaluate, and shape the abstractions
- Cycle of discovery produces abstractions tailored to the environment and ultimately tied to reward

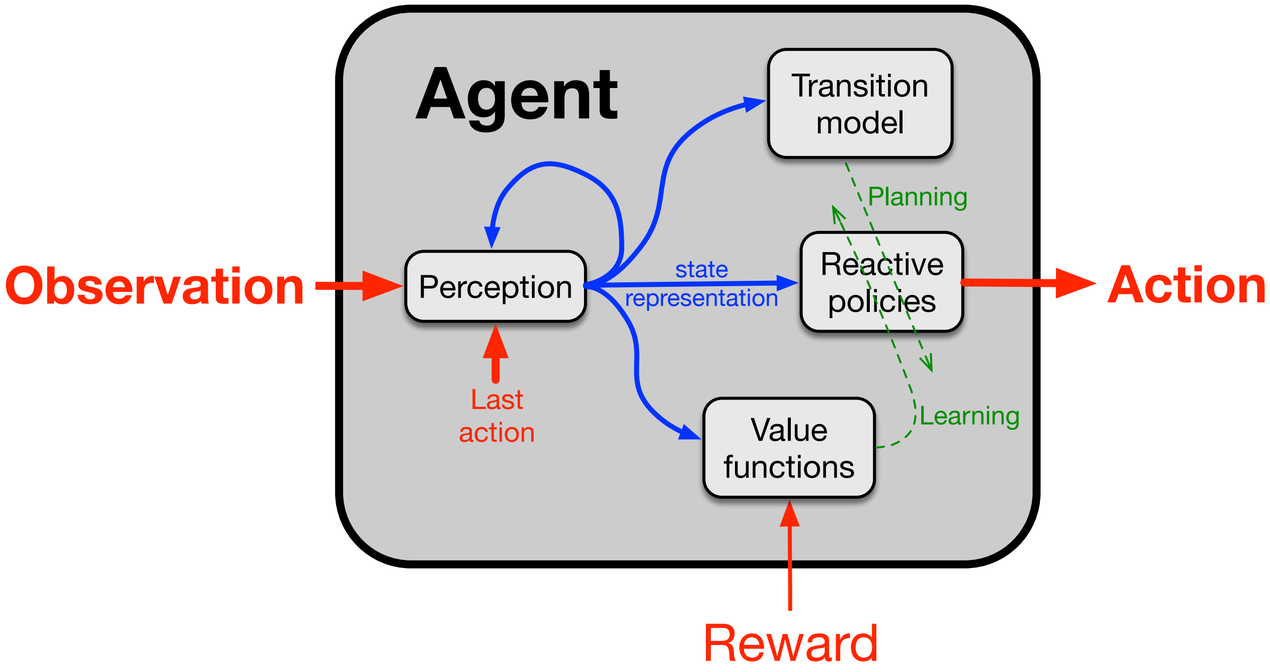
["Reward-respecting Subtasks for Model-based Reinforcement Learning," Sutton, Machado, Holland, Szepesvari, Timbers, Tanner, White, submitted]

Outline (take-home messages)

- Understanding intelligence is a **grand scientific prize** that may soon be within reach; the **glory will go to the bold**
- The Alberta Plan is a direct run at this great prize
  - emphasizing intelligence as the **acquisition of new knowledge**
- My motivation for the plan stems from the **Oak** architecture, a complete conceptual design for a **Model-based RL agent**, a **Proto-AI**
- The Alberta Plan is a massive **retreat**; going back to the most basic algorithms; **re-vamping them for continual and meta-learning** (in 12 steps)

The 12 steps of the Alberta Plan

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



First, the Big World Perspective

- The **world is vastly more complex** than the agent
    - After all, the world contains many other agents,
      and what is going on in many of their minds matters for yours
- It follows that you can never accurately know the state of the world,
  much less its transition dynamics; **your knowledge is a gross approximation** of reality
- This alone makes **prior knowledge a never-ending fool's errand**, a mistake
    - The agent must adapt its approximations to the particular part of the world that it encounters
    - And do this continually as the part encountered changes
- The agent's approximations will be **non-stationary**; they must **track**; this is **continual learning**
- When learning is continual, **extensive repetitive experience is obtained with learning**;
  this makes meta-learning—learning how to learn—possible and essential
- Thus continual learning and meta-learning (e.g., representation learning) are naturally linked

Step 1:

Representation I: Continual supervised learning with given features

- This step exemplifies the **primary strategy of the Alberta Plan**: Retreat, retreat retreat, so as to **study each issue in isolation in the simplest setting** in which it arises
- In Step 1 the issue is **continual learning** and **meta-learning**
- The simplest setting in which they arise is **linear supervised learning** with a **fixed set of features**
- **Normalization** of the features (changing them only by translation and scaling) has a **powerful effect on the speed of learning**; online, continual normalization, has been little studied
- To study **meta-learning** in this simple setting **is** to study **feature relevance**
    - In particular, to set **separate step sizes** (learning rates) **for each feature** and adapt them by **meta-gradient** methods
    - This has been studied before (e.g., IDBD) but not extensively; researchers have been too quick to move on to more complex cases (nonlinearity, new features, sequential data)
- This step is being pursued now by Thomas Degris, Khurram Javed, et al.; it may be half done

Step 2:

Representation II: Supervised feature finding

- Now we bring in the ability to add **new features**
  that are **non-linear functions** of existing features (as in regular deep learning)

- While retaining everything from Step 1:
  normalization, continual learning and meta learning

- **Backprop** (SGD) remains key, but it's **different** with per-weight online normalization
  (for continual learning) and step-size adaptation (for meta learning)

    - In addition, an element of **random creation** (as in random forests and
      generate-and-test feature creation) is thought to be key

- This step, if pursued successfully, will produce a **new form of deep learning**
  designed from the beginning for continual learning and meta learning

[Mahmood 2011; "Maintaining Plasticity in Continual Deep Learning," Dohare, Hernandez-Garcia, Rahman, Sutton, & Mahmood, in preparation]

Step 3:

Prediction I: Continual GVF prediction learning

- Now, for the first time, we consider a **non-i.i.d.** setting
  - We consider a setting in which the data has a **Markov state dependency**
  - But we don't yet consider full RL (control); **we only do prediction**
- A natural, general, prediction setting is that of **off-policy learning of General Value Functions** (GVFs)
- We retain all aspects of Steps 1 & 2 and apply them to this new setting
- There will need several **new elements** having to do with **eligibility traces** and **delayed errors**
- But the **big challenge** will be to extend the new feature creation methods from the i.i.d. case; this will involve **introducing short-term memory** into the new features
- All these issues should be **totally resolved** before going to the next step

[Samani, 2022; Rafiee, Abbas, Ghiassian, Kumaraswamy, Sutton, Ludvig, White, *Adaptive Behavior* 2022]

Step 4:

Control I: Continual actor-critic control

- Only now do we introduce **control**
  - retaining the novelties from Steps 1-3
- First in a **model-free setting**
- Probably in the form of an **episodic actor-critic algorithm**
- There will be some new challenges
  - e.g., setting the two kinds of step sizes to get good continual learning of non-stationary approximations

Step 5:

Prediction II: Average-reward GVF learning

- We return to the prediction case to deal with the new challenges of **average reward**
- The setting here is that we are given the GVF question, and we have to learn the **approximate differential value** from **off-policy data**
- Some preliminary work has been done here by Naik and Wan
- This work must be combined with the new algorithms from Steps 1-3 for continual learning, meta learning, and feature finding

Step 6:

Control II: Continuing control problems

- Now we address the full case of **model-free control**,
  that is, without episodes or discounting, but instead with **average reward**

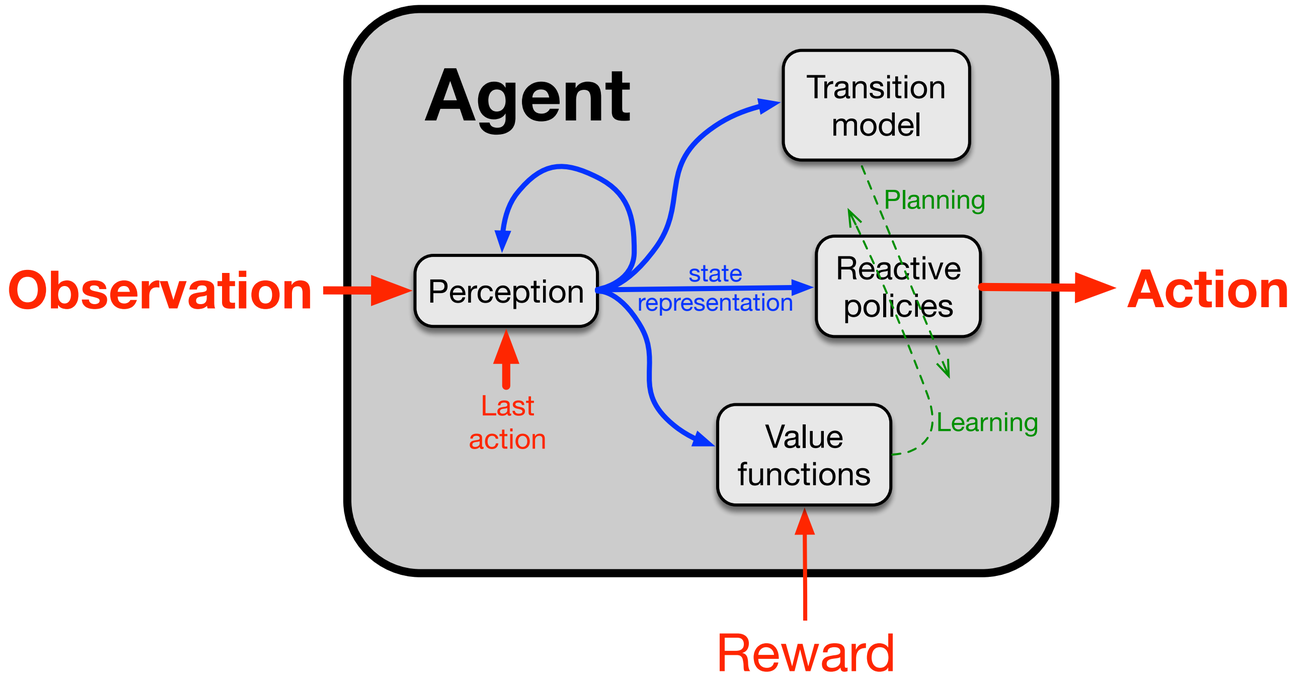
- This step will be the development of **new environments**—a new suite of problems like OpenAI Gym but for the continuing case (“C-suite”)

Step 6 concludes the revamping of all our essential model-free RL algorithms for continual learning and meta learning

Steps 3–6, if done well, will revolutionize model-free RL
and prepare us for a full model-based RL (as in Oak)

The first 6 steps of the Plan re-vamp all 4 components of the common model of the intelligent agent

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



And Steps 9–11 (planning and integration) we have already talked about

Step 12: Prototype-IA: Intelligence Amplification

- In part, this is about the effort spear-headed by Patrick Pilarski to integrate machines and people working together, as in intelligent prostheses
- But in part it is about how machines and people can work together *more generally*
    - the integration may be tight, as in another lobe of your brain,
      or richer sense organs and actuators, created by technology
    - or the integration may be looser, as in a personalized intelligent assistant
    - or still looser, as in Free AIs taking their proper full place in our society and economy
- We are optimistic that these things can be done well, to everyone's mutual benefit,
  but we should all take responsibility, and play a part, in making this likely and widespread

Finally,

Take-home messages (outline)

- Understanding intelligence is a **grand scientific prize** that may soon be within reach; the **glory will go to the bold**
- The Alberta Plan is a direct run at this great prize
  - emphasizing intelligence as the **acquisition of new knowledge**
- My motivation for the plan stems from the **Oak** architecture, a complete conceptual design for a **Model-based RL agent**, a **Proto-AI**
- The Alberta Plan is a massive **retreat**; going back to the most basic algorithms; **re-vamping them for continual and meta-learning** (in 12 steps)

Thank you for your attention