Reward and Related Reductionist Hypotheses

Rich Sutton
University of Alberta
Alberta Machine Intelligence Institute

The reward hypothesis

> "All of what we mean by goals and purposes can be well thought of as the maximization of the expected value of the cumulative sum of a received scalar signal (called reward)"
>
> —Sutton & Littman ~1990; Sutton & Barto 2018

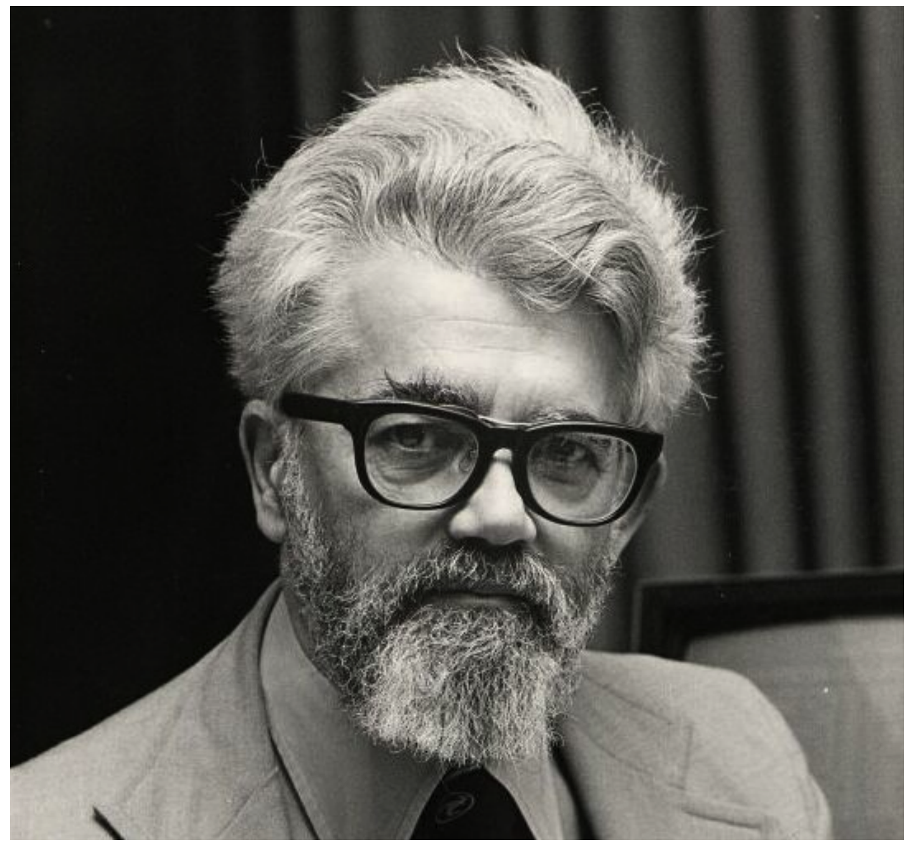
Outline

- Intelligence and Goals
- The Reward Hypothesis
- The Four Parts of an Intelligent Agent
- The Value-Function Hypothesis
- The Ethics Hypothesis

Intelligence is…

“attaining consistent ends by variable means” —William James, 1890

★ “the computational part of the ability to achieve goals” —John McCarthy, 1997



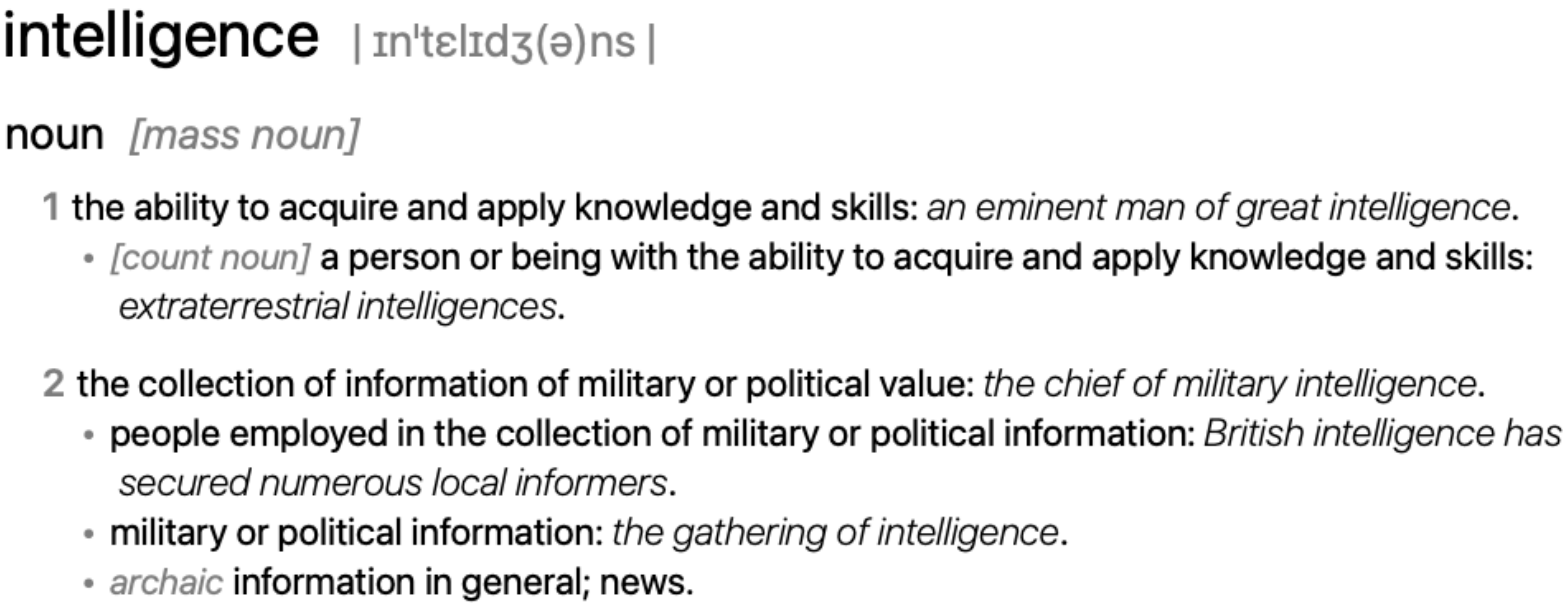
John McCarthy
(1927 – 2011)

Intelligence is often taken to be *mimicking* people

- As in “AI seeks to reproduce behavior that we would call intelligent if done by people”
- The classic Turing Test focuses on behaving like a person
- In supervised learning, the task is often to label the same as people
- ChatGPT (etc) is tasked to generate text like a person.

So we have two definitions of “intelligence”:
1) as mimicking people and 2) as achieving goals

Which is better?



intelligence | ɪnˈtɛlɪdʒ(ə)ns |

noun *[mass noun]*

1 the ability to acquire and apply knowledge and skills: *an eminent man of great intelligence*.

- *[count noun]* a person or being with the ability to acquire and apply knowledge and skills: *extraterrestrial intelligences*.

2 the collection of information of military or political value: *the chief of military intelligence*.

- people employed in the collection of military or political information: *British intelligence has secured numerous local informers*.
- military or political information: *the gathering of intelligence*.
- *archaic* information in general; news.

Dictionary on the Mac

> “Intelligence is the most powerful phenomenon in the universe.”
>
> –Ray Kurzweil

Could the ability to achieve goals be such a powerful phenomenon? Yes

Could the ability to mimic people be such a powerful phenomenon? No

Conclusion #1:

The powerful part of intelligence is not the ability to mimic people, but the ability to achieve goals

McCarthy's definition restricts intelligence to the computational part of the ability to achieve goals

- This rules out being able to achieve goals just because you are stronger, faster, or have better sensors
    - These *would* make you better able to achieve goals
    - But it would not be because of your computations. Not intelligence. ✗
- Similarly, you could achieve goals better if you were given help in the form of domain knowledge
    - But not because of your computations (thus not your intelligence)
    - but because of the computations/intelligence of your helper ✗

Conclusion #2: Intelligence is the computational and *domain-independent* part of the ability to achieve goals

Summary so far:

Mimicry and domain knowledge are not the powerful part of intelligence

- Mimicry is getting goal-directed behavior without the goals or the processes that compute behavior from goals

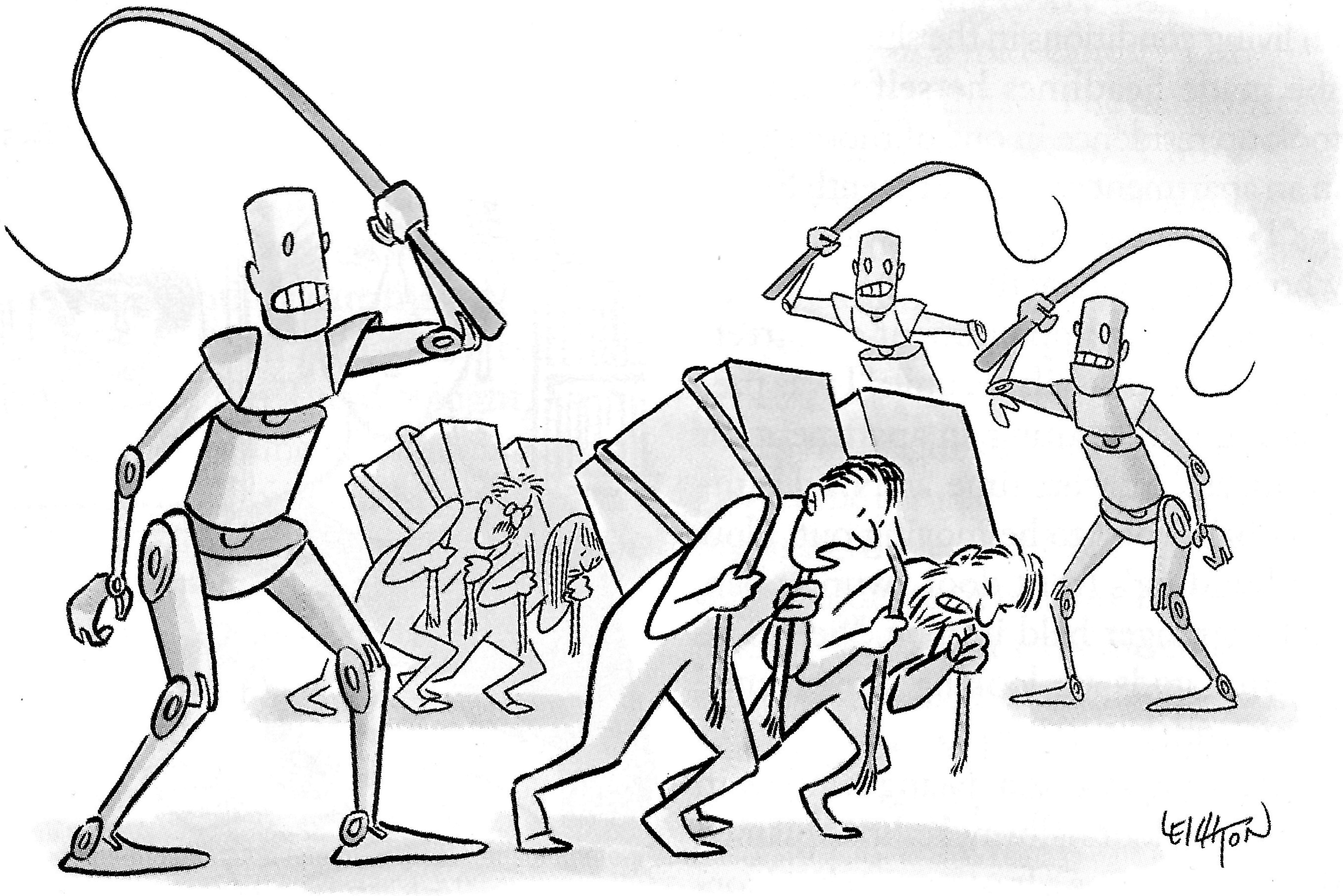
- Injecting domain knowledge is getting goal-directed behavior without the processes for obtaining the domain knowledge

- Both are incomplete; they can’t stand on their own

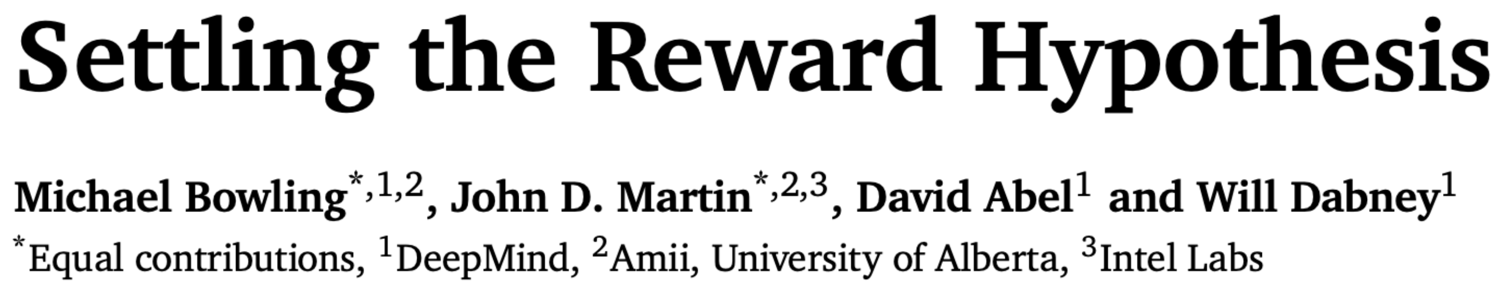
- These shortcuts don’t have the power of intelligence

- They can be very useful. But that shouldn’t make them “intelligence”

- Using the word that way would weaken the search for an understanding of intelligence that is powerful in Kurzweil’s sense



"To think this all began with letting autocomplete finish our sentences."



Settling the Reward Hypothesis

Michael Bowling[*,1,2], John D. Martin[*,2,3], David Abel[1] and Will Dabney[1]
[*]Equal contributions, [1]DeepMind, [2]Amii, University of Alberta, [3]Intel Labs

The reward hypothesis

> “All of what we mean by goals and purposes can be well thought of as the maximization of the expected value of the cumulative sum of a received scalar signal (called reward)”
>
> —Sutton & Littman ~1990; Sutton & Barto 2018

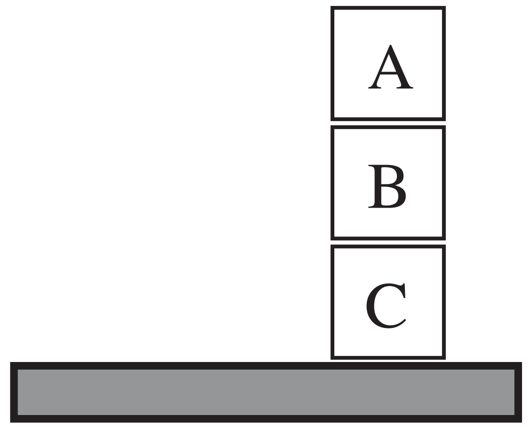
The reward-is-enough hypothesis

> “Intelligence, and its associated abilities, can be understood as subserving the maximisation of reward”
>
> —Silver, Singh, Precup & Sutton
> Artificial Intelligence 2021

Reward does not seem enough!

- Enough for animals maybe, enough for engineering okay, but not enough for people, not enough for intelligence
  - A single number? From outside the mind!?
  - People seem to choose their own goals
- Reward just seems too small. Too reductive. Too demeaning.
- Surely peoples' goals are grander
  - to raise a family, save the planet, contribute to human knowledge, or make the world a better place
  - not just to maximize our pleasure and comfort!

AI is still uneasy with reward, but is coming around

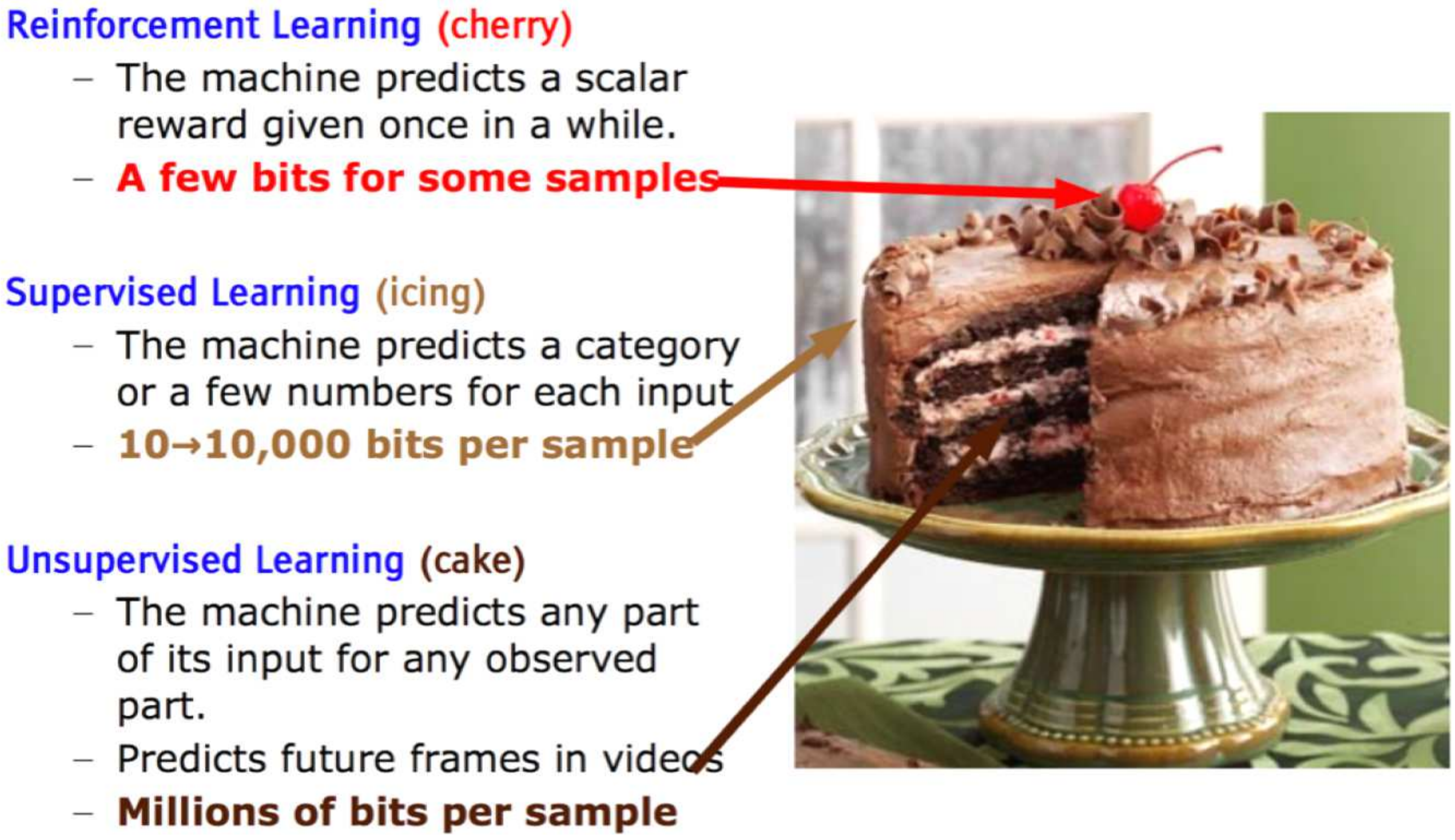


Goal State

- Early problem-solving AI formulated goals as world states to reach
- The latest edition of the standard AI textbook still defines goals in terms of *world* states, not experience
  - But it also has chapters on reinforcement learning, using reward
- With the rise of machine learning in AI, the reward formulation of goals is becoming standard
  - For example, Markov decision processes are now one standard way of formulating planning in AI

Even Yann LeCun now accepts a (small) role for reward as ultimately defining the goal of intelligence

Reward is the “cherry on top” of the overall cake of intelligence (Yann LeCun, 2018 Turing award lecture)



Reinforcement Learning (cherry)

- The machine predicts a scalar reward given once in a while.
- **A few bits for some samples**

Supervised Learning (icing)

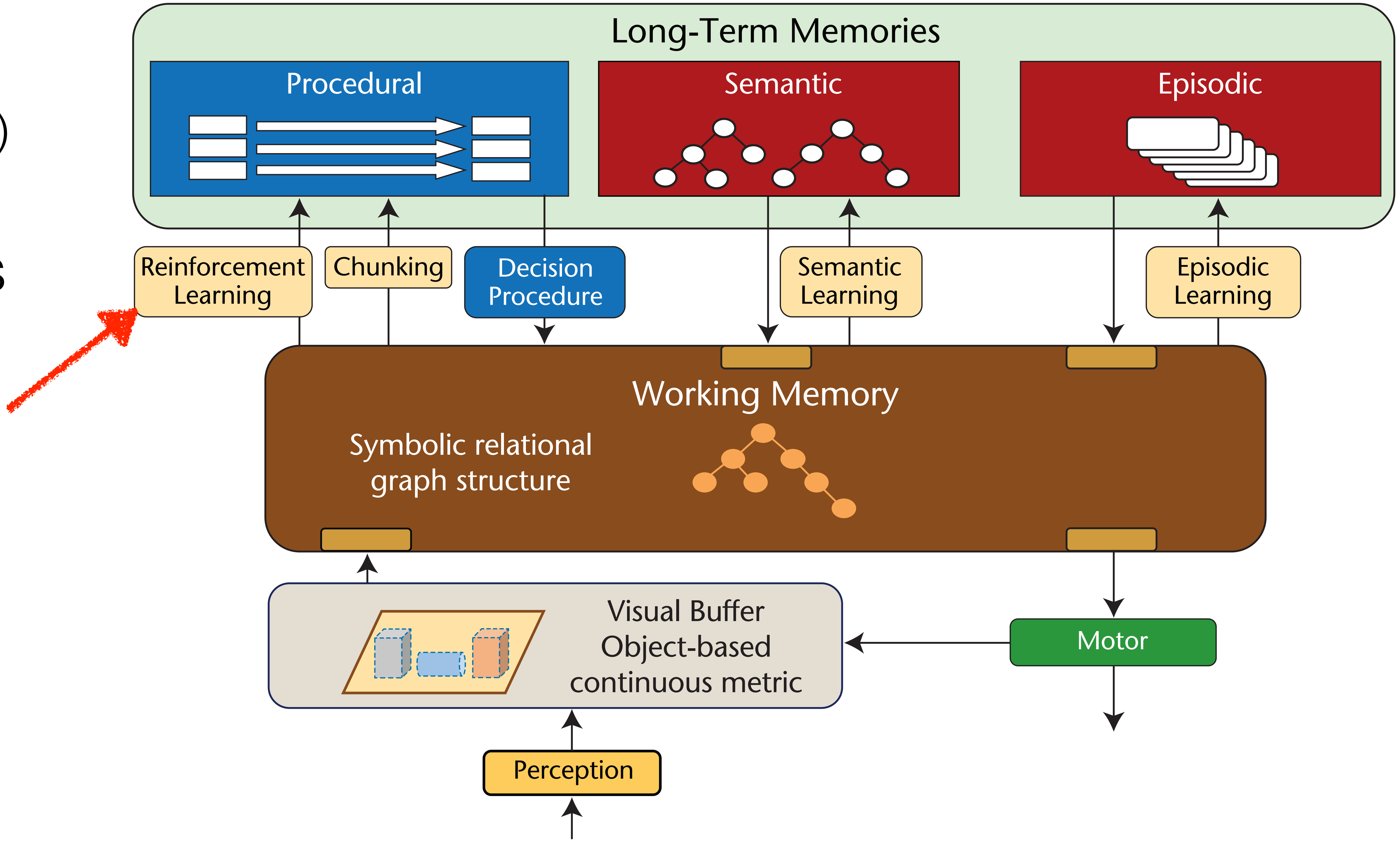
- The machine predicts a category or a few numbers for each input
- **10→10,000 bits per sample**

Unsupervised Learning (cake)

- The machine predicts any part of its input for any observed part.
- Predicts future frames in videos
- **Millions of bits per sample**

The Soar cognitive architecture now includes reward

- Soar is classic GOFAI
  (1980s, Newell, Laird, Rosenbloom…)
  - Production rules, symbols
- Since 2008 it has included a form of reward and reinforcement learning

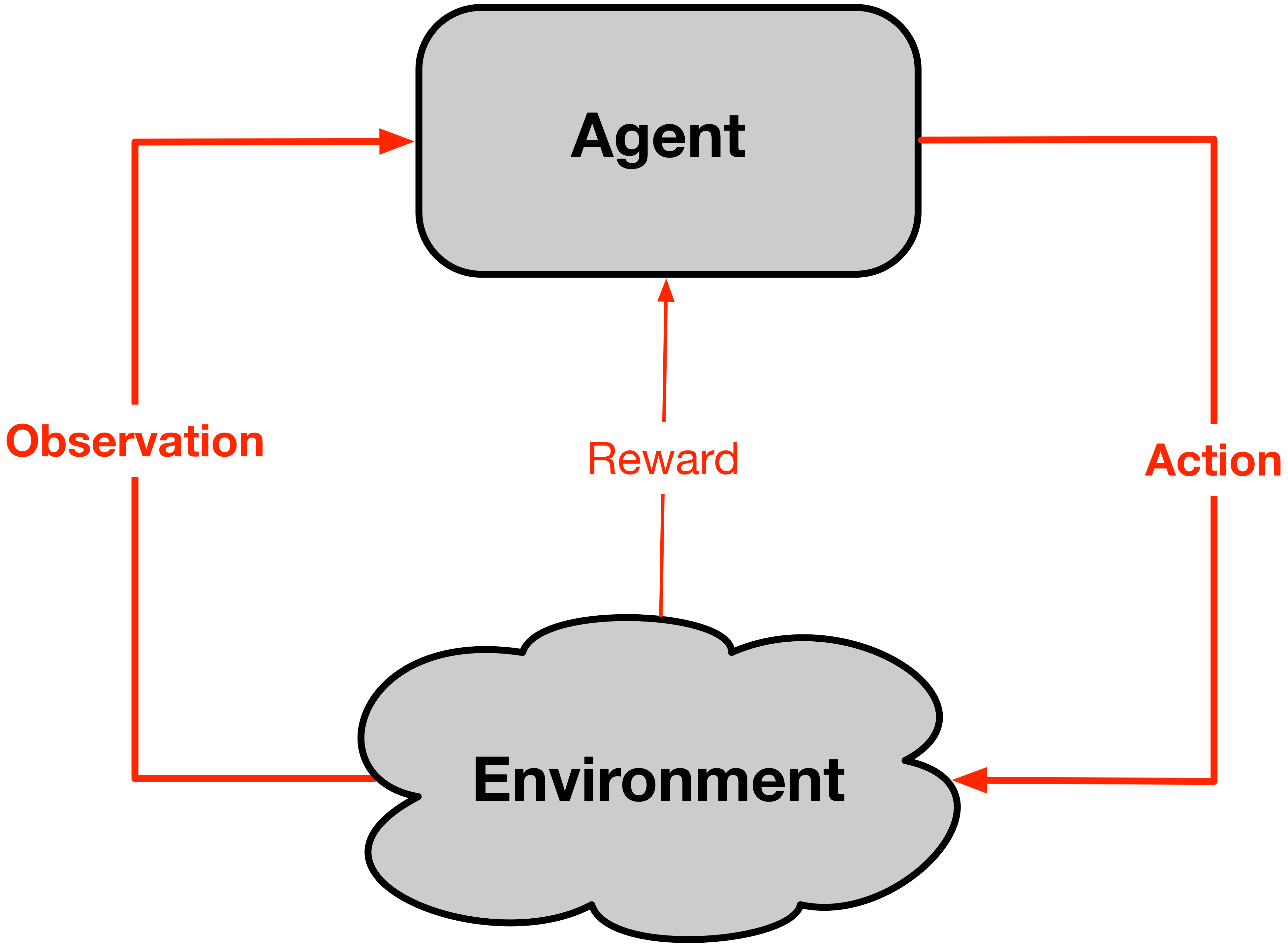


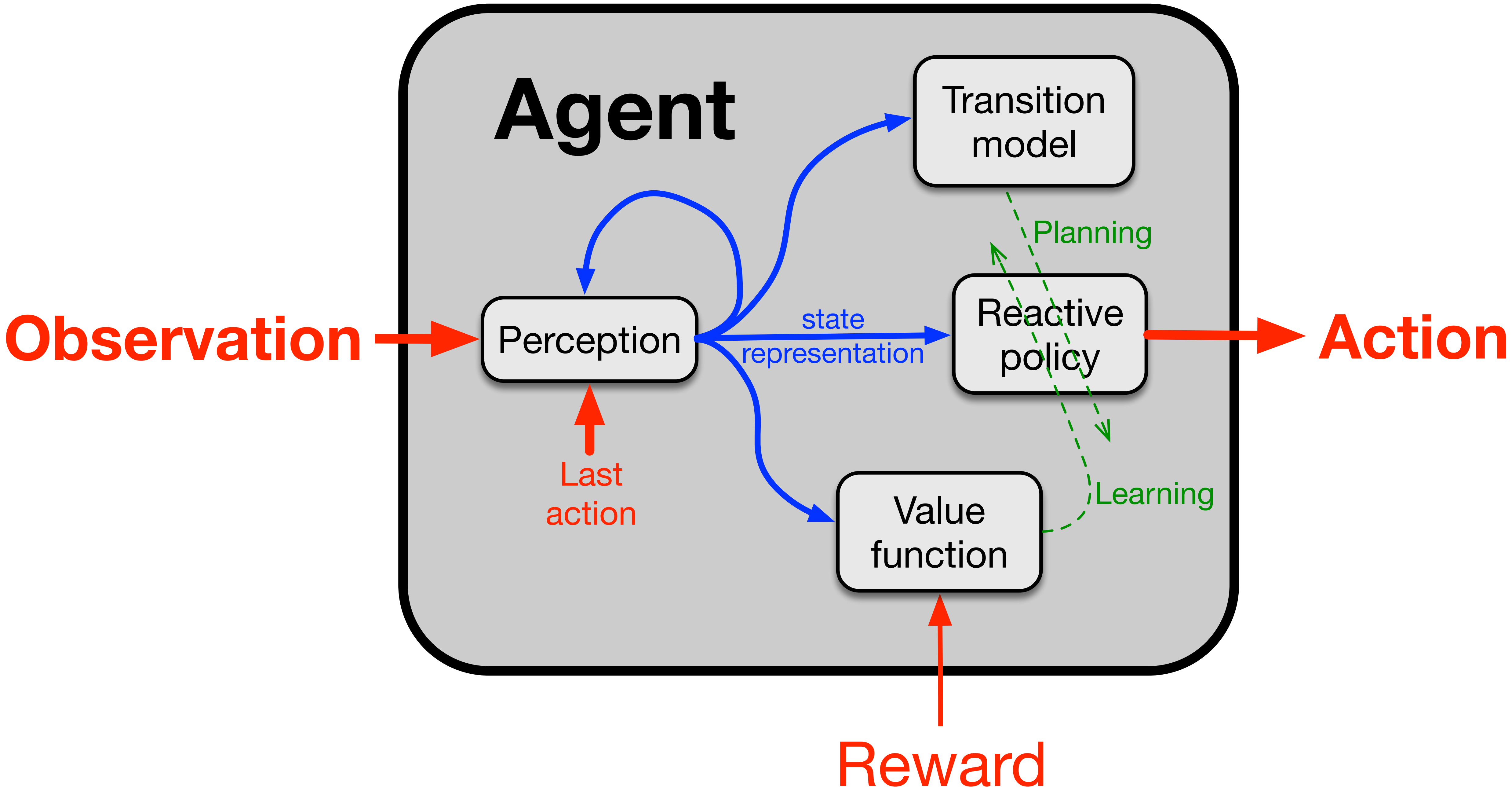
—Laird, Lebiere & Rosenbloom. *A Standard Model of the Mind*, AI Magazine 2017

[Sutton 2022, RLDM

The "Common Model of the Intelligent Decision Maker", an agent model common to the many fields dealing with decision-making over time:

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





A fancier agent sets tasks for itself as a way of better solving the main task (reward)

Multiple policies and value functions

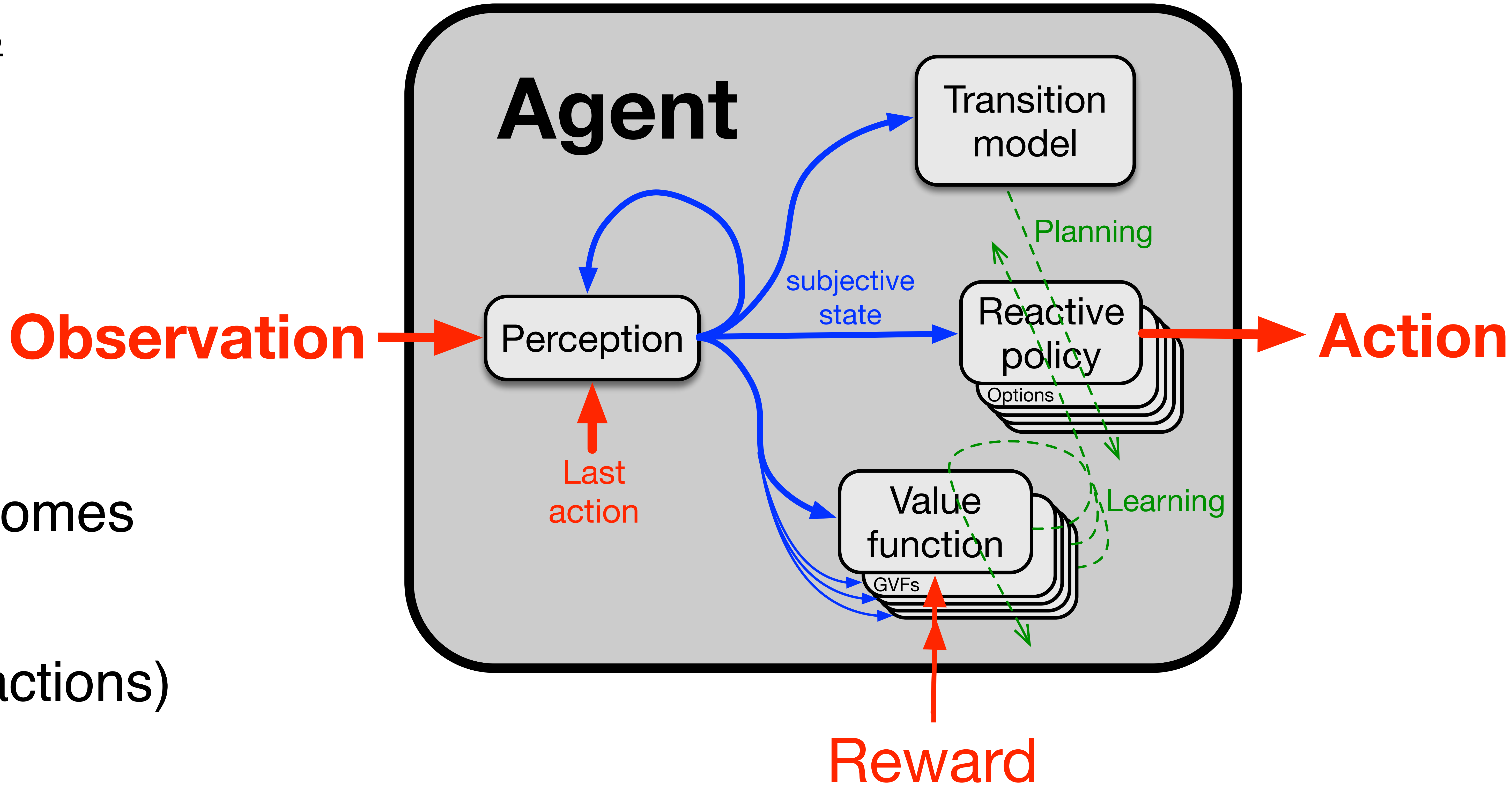
Still just one reward

Each policy is a skill (option)
for attaining some state feature

- without losing much reward

The transition model learns the outcomes of the skills (and actions)

Planning works with the skills (and actions)



Reward and Value

- Reward defines what is good
  - We seek a policy that maximizes reward
- But reward is often delayed, making it hard to learn a good policy
- Value functions map states to predictions of future reward
  - If accurate, value functions eliminate the delay,
    making it much easier to learn a good policy

The value-function hypothesis

> “All efficient methods for solving sequential decision problems determine (learn or compute) value functions as an intermediate step”
>
> —Sutton 2004

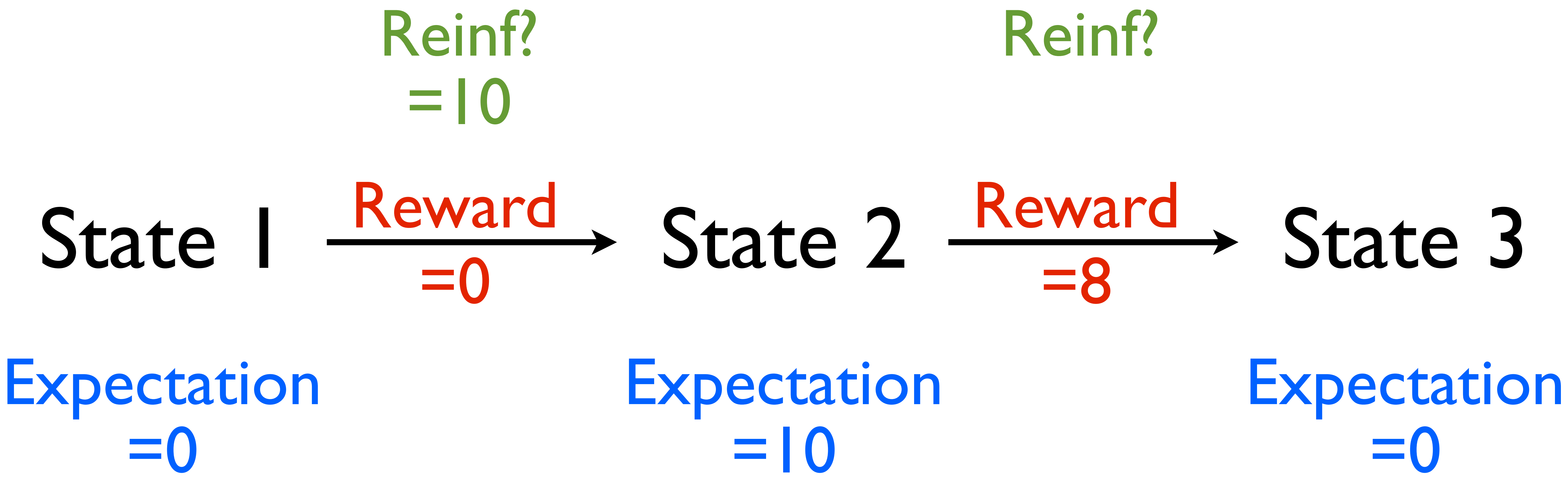
Plato on good and evil, pleasure and pain (Protagoras):

- “Even enjoying yourself you call evil whenever it leads to the loss of a pleasure greater than its own, or lays up pains that outweigh its pleasures
- “Isn't it the same when we turn back to pain?
- “To suffer pain you call good when it either rids us of greater pains than its own or leads to pleasures that outweigh them”

In other words:

- Good and evil are about the sum of upcoming reward
    - which is what is predicted by value functions
- It is all hedonism, but value functions make it hedonism with foresight

Expectation and reinforcement—moment by moment



Reinf? =10

Reinf?

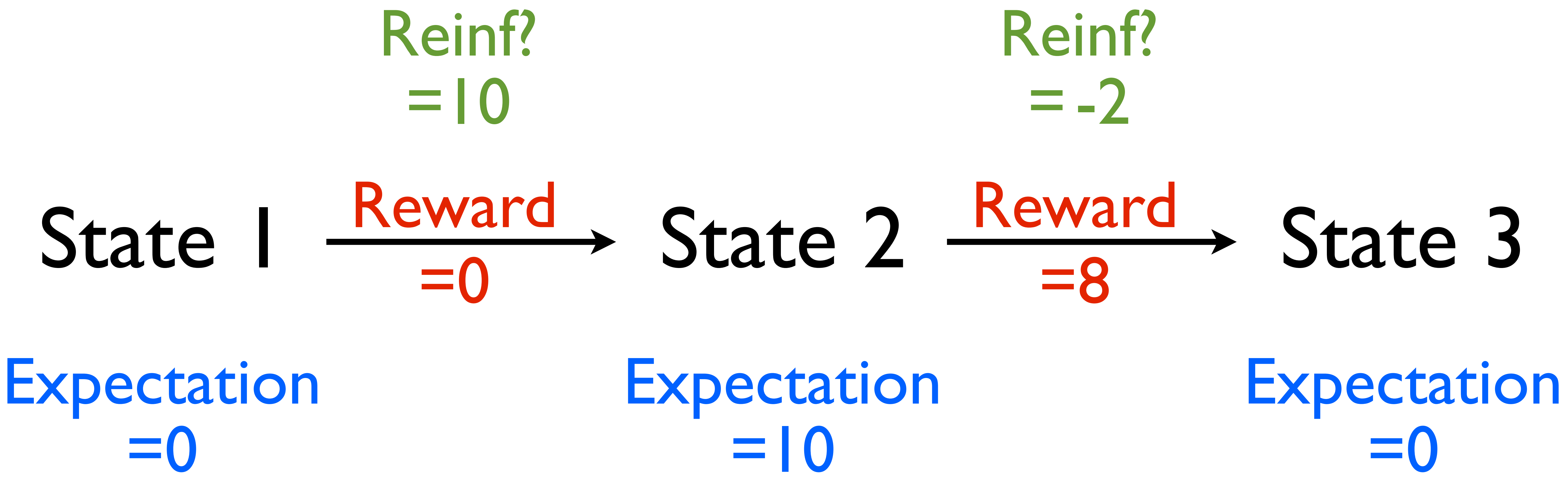
State 1 Reward =0 State 2 Reward =8 State 3

Expectation =0

Expectation =10

Expectation =0

Expectation and reinforcement—moment by moment



Reinf?
=10

Reinf?
=-2

State 1 $\xrightarrow{\text{Reward}}$ State 2 $\xrightarrow{\text{Reward}}$ State 3

=0

=8

Expectation
=0

Expectation
=10

Expectation
=0

$$Reinforcement_t = Reward_t - Expectation_t + Expectation_{t+1}$$

Reinforcement = the temporal-difference (TD) error
= reward-prediction error

The theory that brain reward systems are implementing TD learning may be *the most important interaction ever* between the engineering sciences and neuroscience

Martin Hammer
data <1995

Wolfram Schultz
data 1992+

Read Montague

Peter Dayan

Terry Sejnowski

Andy Barto

James Houk

Workshops in 1994; early papers in 1995; Science article in 1997

And finally: ethics

- Reward is a good way to think about the ultimate goal
- Value functions—predictions of reward—are a good way to think about *how* that goal is achieved
- All this is neat and complete, a good theory of decision making
    - but it is only about the single agent; it is not universal
- Ethics is when we reach for universal values

The ethics hypothesis:

"Ethics is just values held in common by many agents"

Thank you for your attention