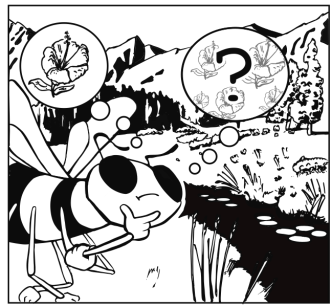
A way around the exploration-exploitation dilemma

Erik J Peterson

Research fellow | CoAxLab
Carnegie Mellon University
robotpuggle.com

The dilemma.



Should I exploit an available reward, or explore to try and find more rewards?

The dilemma.

- Exploration is the problem.

The dilemma.

- Exploration is an intractable problem.
    - There is no optimal solution.
    - Only average solutions. [4, 1, 2, 3].

The dilemma.

- An average life is full of regret.

The average country singer.

The dilemma.

- An average life is full of regret, G.
- Where $G = V^{*} - V_a$

Our goal.

- To lead a life of no regret.

Our goal.

- To lead a life of no regret, $G = 0$
- ...for all $s \in S$, $a \in A$ and $t \leq T$.

Why do animals explore?

- Find rewards
-
-

Why do animals explore?

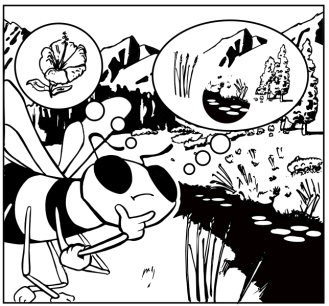
- ~~Find rewards~~
-
-

Why do animals explore?

- ~~Find rewards~~
- Learn about their niche
- Learn about other animals

Why do animals explore?

- ~~Find rewards~~
- They are curious



Curiosity is not a luxury.

- If you do not learn about your niche, you die.

Big conjecture #1.

Information is fundamentally valuable.

Big conjecture #2.

Exploration for reward is never needed. The only exploratory behavior an animal needs is that which builds its world model.

Ad hoc curiosity.

- Novelty
- Counts/Successors
- Information gain
- Entropy
- State prediction

Reinforcement learning

- Reward learning has made progress because it has a clear objective:
  - $\max \sum_T R$

Goal.

- Curiosity learning needs a clear and common objective:
  - $\max \sum_T E$

What is E?

- Novelty?
- Counts/Successors?
- Information gain?
- Mutual information?
- State prediction?

What is E?

- Novelty?
- Counts/Successors?
- Information gain?
- Mutual information?
- State prediction?
- Free energy?

A minimum definition.

- At an *absolute minimum* what do we need to value information ?

A minimum definition.

What do we need to value information?

1. A world

A minimum definition.

What do we need to value information?

1. A world
2. Actions

A minimum definition.

What do we need to value information?

1. A world
2. Actions
3. A memory

A minimum definition.

What do we need to value information?

1. A world
2. Actions
3. A memory $\Leftrightarrow$ *world model*

A minimum definition.

What do we need to value information?

1. A world, S^n
2. Actions A^m
3. A memory, M^k

A minimum memory.

What does a memory need to be a memory?

1. Finite

A minimum memory.

What does a memory need to be a memory?

1. Finite
2. It must remember

A minimum memory.

What does a memory need to be a memory?

1. Finite
2. It must remember
3. It must recall

A minimum memory.

What does a memory need to be a memory?

1. Finite
2. It must remember
3. It must recall
4. It must forget

A minimum memory.

What does a memory need to be a memory?

1. Finite, M^k
2. It must remember, $f : s, M \rightarrow M'$ where $s \in S$
3. It must recall, $g : s, M \rightarrow \hat{s}$
4. It must forget, $f^{-1} : s, M' \rightarrow M$

A minimum memory.

What does a memory need to be a memory?

1. Finite, M^k
2. It must remember, $f : s, M \to M'$ where $s \in S$
3. It must recall, $g : s, M \to \hat{s}$
4. It must forget, $f^{-1} : s, M' \to M$
5. Made of real numbers, $M \in \mathbb{R}$

Commonalities?

- Novelty
- Counts/Successors
- Information gain
- Mutual information
- State prediction
- Value comes from how memory changes?

Axiom 1.

Axiom (Axiom of Change)
The value of information E depends only on the total distance M moves by making observation s.

Minimal distance d.

- Let $\delta = d(m, m')$, where $m \in M$ and $m' \in M'$
- $\delta \geq 0$
- $\sum \delta = 0$ only if $M = M'$
- (d is a pre-metric)

Total distance $||\Delta||$.

- The total distance is the norm of Δ, $||\Delta||$
- Where $\Delta = \{\delta_1, \delta_2, ..., \delta_K\}$

Total distance $||\Delta||$.

- The total distance is the norm of Δ, $||\Delta||$
- Where $\Delta = \{\delta_1, \delta_2, ..., \delta_K\}$
- **Let** $E \equiv ||\Delta||$

What is E.

Axiom 1 $\rightarrow ||\Delta||$.

Examples.

Choose $\{M, f, g, d\}$.

-

Examples.

Choose $\{M, f, g, d\}$.

- A novelty world model:
  - $M \in \mathbb{R}^n$
  - f, g : add s, returns 1 if no s
  - d : $|m - m'|^1$ (the *l1 norm* or *Manhattan distance*)

Examples.

Choose $\{M, f, g, d\}$.

- A count-based world model:
  - $M \in \mathbb{Z}^n$
  - f, g : counts s, returns counts of s
  - d : $|m - m'|^1$

Examples.

Choose $\{M, f, g, d\}$.

- A compressed world model:
  - $W^c < \mathbb{R}^n$
  - f, g : lossy encoder, returns $\hat{s}$
  - d : $\sqrt{|w - w'|^2}$ euclidean distance on weight parameters, W

Examples.

Choose $\{M, f, g, d\}$.

- A compressed world model:
  - Linear regression (regularized)
  - GAN
  - VAE

Examples.

Choose $\{M, f, g, d\}$.

- An episodic world model:
  - $M \in \mathbb{R}^n$
  - f, g : store s, recall s
  - d : $\sqrt{|m - m'|^2}$

Examples.

Choose $\{M, f, g, d\}$.

- An categorization world model:
  - $W^c \leq \mathbb{R}^n$
  - f, g : categorize s, recall category of s
  - d : $|m - m'|^1$ or
  - d : $\sqrt{|w - w'|^2}$ euclidean distance on weight parameters, W

Examples.

Choose $\{M, f, g, d\}$.

- An categorization world model:
  - Clustering (*e.g.*, K-means)
  - Classification (*e.g.*, SVM, Random Forrest)

Examples.

Choose $\{M, f, g, d\}$.

- A Bayesian world model:
  - $M \in \mathbb{R}^{nm}$
  - f, g : Bayes rule
  - $d \approx KL - divergence$

Axiom 2.

Axiom (Axiom of Equilibrium)
To be valuable an observation s must be learnable by M.

Axiom 2.

Learnable:

1. With every (re)observation of s, M should change.
2. The change in M must eventually reach a learned equilibrium.

Axiom 2.

- Learnable $\approx$ gradient
- $\mathbb{E}\left[\nabla^2 M\right] \leq 0$

How to maximize?

Recall: $E \equiv ||\Delta||$

Goal: $\max \sum_T E$

Is it enough?

Recall:

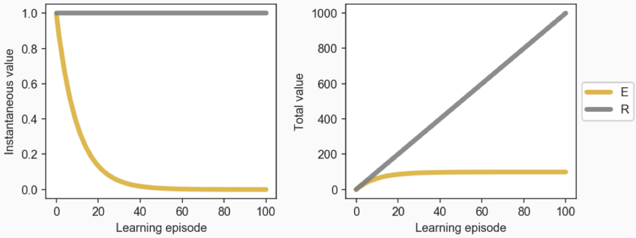
- A minimal memory
- Axiom of change
- Axiom of equilibrium

It is.

The Bellman equation is the optimal learning rule for E.

- $V^*_{\pi_E} = \left[E + \ \text{argmax}_{a \in A} E' \,\middle|\, S, M, f, g, d \right]$
- (Proof by induction on M; M has optimal substructure.)

Value during optimal play.



Reward and information value.

Progress.

1. Information is fundamental
2. An axiomatic definition for information value, E
   - A minimal memory
   - A minimal distance
   - Value is the total distance the memory moves
3. Proven how to maximize E, optimally
   - Every time step t is greedy; always pick the biggest distance.

~~Goal.~~

- Curiosity learning *has* a clear and common objective:
  - $\max \sum_T E$

The dilemma.

Should I exploit an available reward, or explore to try find more rewards?

The dilemma.

- Reward is fundamentally valuable
- Information is fundamentally valuable

The dilemma.

- Reward is fundamentally valuable
- Information is fundamentally valuable
- Reward $\Leftrightarrow$ Information?

Information is not a reward?

- Rewards are a conserved resource.
- Information is not.

Information is not a reward?

- Reward value is fixed(-ish).
- Information value is *never* fixed when learning.

Big conjecture #3.

Information is not a reward.

Information is not a reward.

No:

- $R + E$
- $R + \beta I$
- $R + novelty$

Now we have two problems....

Learn:

1. An action policy π_R to $\max \sum_T R$
2. An action policy π_E to $\max \sum_T E$

Dual value learning

Solution:

$$\pi^{\pi} = \begin{cases} \pi_E^* & : E - \eta > R \\ \pi_R & : E - \eta \leq R \end{cases}$$

$$\text{subject to the constraints}$$

$$R \in \{0, 1\}$$

$$p(\mathbb{E}[R]) < 1$$

$$E - \eta \geq 0$$

$$\text{choose } E_0 > 0$$

- (Proof by induction; it's a classic scheduling problem)

Optimality of π^π.

Search:

- *If* M is complete in *S*,
- *then* *S* is searched completely and exhaustively

Optimality of π^π.

Reward:

- *If* π^*_R and $M \leftarrow R$,
- *then* $\max \sum_T R$ (or $\max \sum_T \gamma^t R$)

Optimality of π^π.

No regret:

- $G = 0$ for all $a \in A$ and $s \in S$
- (Every choice is greedy in π^π)

Progress.

- A no regret way to solve exploration-exploitation problems
- Learns a/the necessary world model along the way
    - (Works for *nearly* any world model)
- Straight forward to ensure the best reward policy is found
    - (Works for *nearly* on reinforcement learning rule)
    - (...Just pick a *good one*)

In practice.

- Theoretical promises often fail in practice.
- (But initial results look promising.)

n-armed bandits.

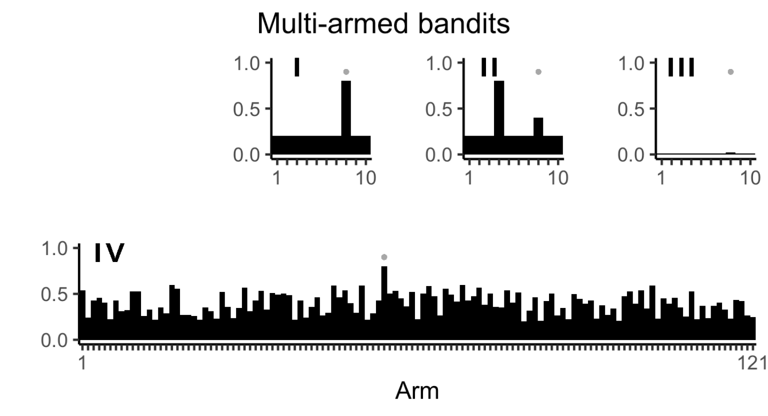
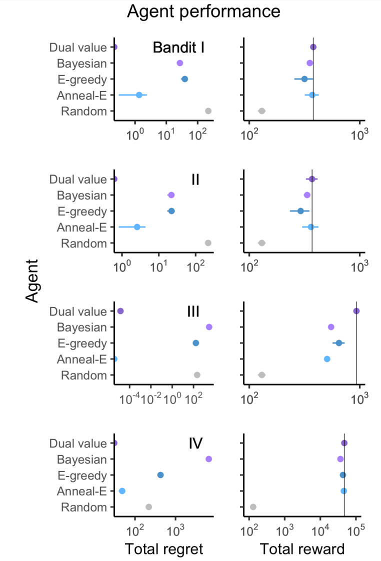


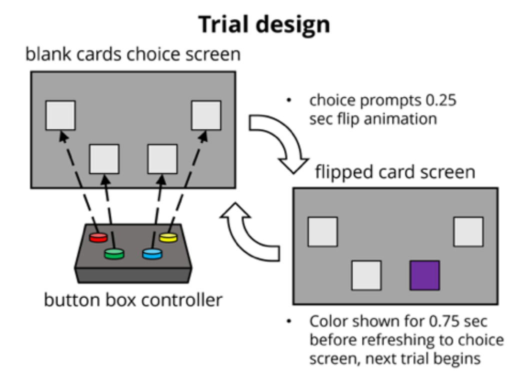
Table 1. Artificial agents.

Agent	Exploration mechanism
Dual value	Our algorithm (Eq 5).
E-greedy	With probability $1-\epsilon$ follow a greedy policy. With probability ϵ follow a random policy.
Annealed e-greedy	Identical to E-greedy, but ϵ is decayed at fixed rate.
Bayesian reward	Use the KL divergence as a weighted intrinsic reward, sampling actions by a soft-max policy. $\sum_T R_t + \beta E_t$
Random	Action are selected with a random policy (no learning)

n-armed bandits.

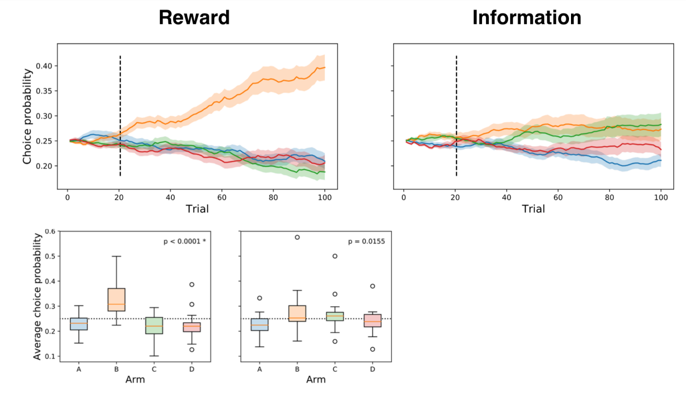


Do humans value information equally?



The *bitjoy* bandit task.

Bitjoy.



Human behavioral performance ($n = 24$).

Conclusions.

- Don’t explore to get rewards.
- Explore to learn.
- You’ll have no regrets, and the reward will come.

Future work.

Explore deterministic (optimal) exploration in:

- Artificial intelligence
- Animal behavior

Open science.

Paper `biorxiv.org/content/10.1101/671362`

Code `github.com/CoAxLab/infomercial`

Talk `github.com/parenthetical-e/dilemma-talk-irvine-2019`

Thank you!

Hire me!

`robotpuggle.com`

References i

- P. Dayan and T. J. Sejnowski.
  Exploration bonuses and dual control.
  Machine Learning, 25(1):5–22, Oct. 1996.
- C. Findling, V. Skvortsova, R. Dromnelle, S. Palminteri, and V. Wyart.
  Computational noise in reward-guided learning drives behavioral variability in volatile environments.
  Preprint, Neuroscience, Oct. 2018.
- S. J. Gershman.
  Deconstructing the human algorithms for exploration.
  Cognition, 173:34–42, Apr. 2018.

S. Thrun and K. Möller.
Active Exploration in Dynamic Environments.
Advances in neural information processing systems, pages 531–538, 1992.