Active Inference: Towards A Theory of Agency

Neuro x Statistical Physics framework for Agent Foundations with empirical experiments

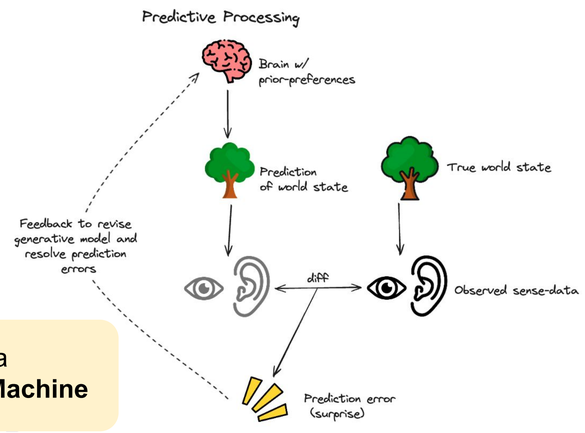
MATS

Kola Ayonrinde

Summary

- We use the **Active Inference** framework as a foundation for **understanding agency** in AI systems.
- We show empirically that Active Inference - perceiving and acting in order to **minimise prediction error** - leads to learning **better world models** than perception-only settings like supervised learning.
- We focus on **multi-agent interactions** for bounded-rational agents and learning equilibria in mixed-motive games.

Predictive Processing



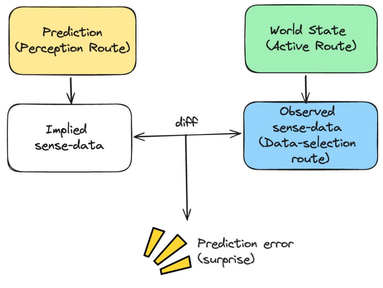
The brain is a **Prediction Machine**

Active Inference

Active Inference = Predictive Processing + Actions As Inference

There are 3 ways for agents to resolve prediction error:

1. **Prediction**: Change the model to match the world
2. **Active**: Act to change the world
3. **Data Selection**: Act to change your future observations

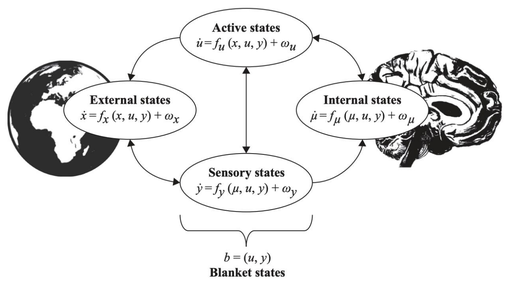


The Free Energy Principle

Theorem: Free Energy Principle
Any "thing" that (1) has a **statistical boundary** with its environment and (2) attains a **nonequilibrium steady state** (C) can be construed as performing **approximate Bayesian inference**.

Active Inference agents **minimise Free Energy**, **F** (right)

$$
\begin{aligned}
F[Q,y] &= \underbrace{-\mathbb{E}_{Q(x)}[\ln P(y,x)]}_{Energy} - \underbrace{H[Q(x)]}_{Entropy} \\
&= \underbrace{D_{KL}[Q(x)\,||\,P(x)]}_{Complexity} - \underbrace{\mathbb{E}_{Q(x)}[\ln P(y|x)]}_{Accuracy} \\
&= \underbrace{D_{KL}[Q(x)\,||\,P(x|y)]}_{Divergence} - \underbrace{\ln P(y)}_{Evidence}
\end{aligned}
$$



Agents are statistically separated from the environment through an information boundary known as a **Markov boundary** (left).

Why Active Inference for Agent Foundations?

Active Inference offers a framework which is complementary to other Agent Foundations work.

We would like formal definitions for concepts important for agency:

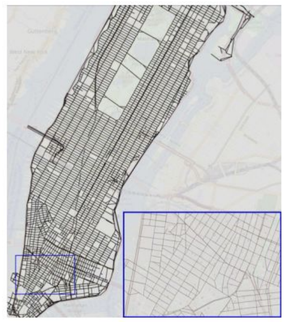
- **Boundaries**: Where does the agent end and the environment start?
- **Agency**: What are the foundational parts of an agent?
- **Optimisation**: What does it mean to optimise towards a goal?
- **Beliefs & Preferences**: How do minds emerge from matter?
- **Abstractions**: What does an agent's world model look like? What must it learn?

Active Inference offers both a **normative** and **descriptive** theory of agency:

- We can describe current systems like LLMs in the natural world (see experiments).
- We can make useful predictions about future systems.

LLMs as Active Inference Agents

LLMs are structurally similar to Active Inference agents.
On the task of generating a world model of the MDP representing the streets of NYC, we observe **more robust world models**.



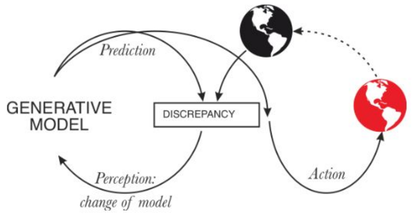
(a) World model

Learning Method	World Model Metric
Shortest paths	0.19 (0.00)
Noisy shortest paths	0.07 (0.01)
Random walks	0.68 (0.02)
Active Inference (ours)	0.92 (0.03)
True world model	1.0 (0.0)

We use Vafa et al's behavioural metrics for eliciting world models.

Multi-Agent Experiments: early results

In our custom environments of mixed-motive games we find that Active Inference agents typically learn cooperative equilibria similar to RL agents *without explicit regularisation and auxiliary exploration losses*. We hypothesise that better models of other agents improves cooperative capability.



Formal results incoming 👀

Future Work

- Releasing our library for efficient training of Active Inference multi-agent systems.
- Sharing an Alignment Forum sequence detailing Active Inference as a complement to other Agent Foundations work.

References

Hyland et al. (2024). "Free Energy Equilibria: Toward a Theory of Interactions Between Boundedly-Rational Agents"
Kulveit et al. (2023). "Predictive Minds: LLMs as Atypical Active Inference Agents"
Parr et al. (2022). "Active Inference: The Free Energy Principle in Mind, Brain and Behaviour"
Vafa et al (2024). "Evaluating the World Model Implicit in a Generative Model"