# Variational Free Energy Minimization as a Mechanism for Active Inference and Curiosity

Modelling in neuroscience - and elsewhere (Master MVA)
Final Project

Amric Trudel
amric.trudel@ens-paris-saclay.fr

April 17, 2023

## 1 Introduction

Active inference is a theoretical framework proposed by Karl Friston that has gained significant attention in the field of neuroscience and cognitive science. It suggests that the brain is continuously making predictions about the world based on sensory input, and perception and action result from a process of minimizing prediction errors. In several papers published in the 2010's, Friston has shown that the active inference scheme can be used to model a variety of tasks and games, such as waiting games [5], two-step maze tasks [4] and evidence accumulation in the urn task [1], saccadic eye movements in scene construction and engineering benchmarks such as the mountain car problem [2].

In his 2017 paper titled "Active Inference, Curiosity and Insight,"[3], the paper that will be studied in the current report, Friston shows, with a simulation, that the emergence of curiosity and insight can be explained by a single principle: the minimization of free energy. A basic task that involves rule learning is learned by an artificial agent that has beliefs about hidden states of its environment and acts in order to reduce its uncertainty about the environment as well as to improve the accuracy of its prediction. Friston shows that the simple process of minimization of free energy is sufficient to foster the emergence of curiosity. In a second part, Friston simulates a bayesian model reduction process, akin to sleep for humans, in which the agent is able to reduce its model complexity by eliminating redundant parameters.

This report will go over the active inference principle and how it models the agent's beliefs in the task introduced in the paper. Then, a reinforcement-learning-based implementation of the task will be proposed in order to compare how the two afford to model a rule-learning task with various degrees of modelling expressiveness.

## 2 Context

We will start an overview of the historical and theoretical background that has led to the development of the active inference framework proposed by Karl Friston. This section highlights the key ideas and concepts that have shaped our understanding of perception and action in neuroscience and cognitive science.

### 2.1 Models of perception and cognition

For much of the 20th century, the dominant view in neuroscience was that the brain was a passive receiver of sensory input, and perception was largely determined by the properties of the external stimuli. However, this view was challenged by the emergence of the field of cognitive science in the 1950s and 1960s. Cognitive science highlighted the importance of internal mental processes in shaping perception and cognition, and proposed that perception and action were the result of active processes that involved top-down predictions and feedback mechanisms.

This perspective was further developed in the 1980s and 1990s with the emergence of the computational theory of mind and the Bayesian brain hypothesis. These theories proposed that the brain was constantly generating and updating probabilistic models of the world based on sensory input and prior knowledge, and that perception and action were the result of a process of probabilistic inference.

Karl Friston's active inference framework builds on these ideas, but takes them in a new direction. Friston suggests that the brain is not simply performing probabilistic inference but is actively seeking to minimize prediction errors. This means that the brain is constantly generating and testing hypotheses about the world and updating its predictions based on the discrepancy between its predictions and actual sensory input. The active inference framework has significant implications for our understanding of perception, decision-making, and cognition. It suggests that these processes are not just the result of passive reception of sensory input but are actively generated and shaped by the brain's internal models of the world. The key idea to remember here is that the sampling of the environment is not done randomly, is the result of the actions that the agent took to influence it.

## 2.2 Two different approaches to policy optimization

Friston's approach relies on the formalization of a Markov Decision Process, just like reinforcement learning, that has gained a large popularity recently. Reinforcement learning's optimization is *value-based*, which means that actions are sampled in order to maximize a value function that takes as input **actual states** of the world. Reinforcement learning defines a *policy* as a probability distribution over actions, at a given state:

$$\begin{aligned} a_t^* &= \arg\max V(s_{t+1}|a_t) \\ &= \pi(s_t) \end{aligned} \tag{1}$$

Friston's approach, on the other hand, relies solely on the agent's **beliefs about the states** of the world. It explicitly uses the agent's internal model for action selection, and the optimal action is expressed as an energy functional of a belief under a particular course of action:

$$a_t^* = \arg\min F(Q(s_{t+1})|a_t) \tag{2}$$

The goal in this second approach is not to optimize the individual actions themselves, but the policy, defined now as a sequence of actions through time. This approach stresses the importance of order within this course of actions.

$$\pi^* = \arg\min \sum_{\tau} F(Q(s_\tau)|\pi) \tag{3}$$

Equation 3 shows that the optimal policy minimizes a time average (or path integral) of an energy function, which is also known as *action* in physics. It shouldn't be confused with the actions performed by agents in our context, but the interesting point is that the parallel with the domain of physics allows us to interpret good behavior through the lens of Hamilton's principle of least action. Table 1 shows equivalencies between Bellman Optimality principle and Hamilton's principle of least action.

| Bellman's Optimality Principle (RL) | Hamilton's Principle of least Action |
|---|---|
| Optimal control theory | Free energy principle |
| Dynamic programming | Active inference |
| Reinforcement learning and expected utility | Artificial curiosity and intrinsic motivaiton |
| Backwards induction | Bayesian theory |
| State-action policy iteration | Bayesian sequential policy optimisation |
| Markov Decision Process (MDP) | Partially-Observable MDP (POMDP) |

Table 1: Comparative association of approaches that appeal to Bellman's Optimality Principle, used in reinforcement learning, as opposed to Hamilton's Principle of least Action, used in active inference. (source [6])

# 3 Active Inference and Free Energy

Active inference is a theoretical framework for understanding perception, planning and learning that emphasizes the importance of prediction and inference. As explained earlier, it portrays an agent who is constantly generating predictions about the sensory input it expects to receive, based on its internal models of the world. These predictions are then compared to the actual sensory input, and any discrepancies are used to update the internal models and improve future predictions. Active inference can be seen as a form of Bayesian inference, where beliefs about the world are being upadated based on new sensory evidence. However, active inference also emphasizes the importance of action, and proposes that the agent selects actions that are likely to minimize prediction errors and maximize the accuracy of its internal models. The sampling of the environment by the agent can have to main objectives: to better understand how the environment works or to satisfy its prior preferences in terms of the sensory inputs it expects. When the context is ambiguous, the free energy principle pushes the agent to engage in exploratory behavior. When there is no further uncertainty to resolve, the agent becomes rather inclined to exploitative behavior.

The generic process that underpins perception, planning and learning under this framework is the following:

1. The agent **infers** the hidden states that will result from each policy it considers.
2. The agent **evaluates** the evidence for each policy based on observed outcomes and beliefs about future states.
3. The agent **selects** the next action based on posterior beliefs about each policy.
4. The outcomes of the action is added as experience to the contingency table of outcomes with regard to expected hidden states (in other words, model parameters are learned.)

## 3.1 Generative Model

Since active inference is all about minimizing expected outcomes and actual perceived outcomes, the agent needs a generative model that it can use to predict future outcomes. The generative model works as follows:

1. A policy is selected using a softmax function of the expected free energy of each policy
2. Sequences of hidden states are generated using the probability transitions specified by the selected policy
3. Hidden states generate outcomes

Figure 1 illustrates the probabilistic dependencies involved in the generative model.

## 3.2 Free Energy Minimization

In the previous model, we saw that policy selection favors policy that have the lowest free energy:

$$\begin{aligned} P(\pi) &= \sigma(-G(\pi)) \\ \text{where} \quad G(\pi) &= \sum_{\tau} G(\pi, \tau) \end{aligned}$$

The negative free energy of a policy at a given time can be expressed as such:

$$\begin{aligned} -G(\pi, \tau) &= E_{Q(o_\tau, s_\tau \tau \pi)} \left[\ln P\left(o_\tau, s_\tau \mid \pi\right)\right] + H\left[Q\left(s_\tau \mid \pi\right)\right] \\ &= E_{Q(o_z, s_\tau, \pi)} \left[\ln Q\left(s_\tau \mid o_\tau, \pi\right) + \ln P\left(o_\tau \mid m\right) - \ln Q\left(s_\tau \mid \pi\right)\right] \\ &= \underbrace{E_{Q(o_\tau \mid \pi)} \left[\ln P\left(o_\tau \mid m\right)\right]}_{\text{Extrinsic value}} + \underbrace{E_{Q(o_z \mid \pi)} \left[D\left[Q\left(s_\tau \mid o_\tau, \pi\right) \| Q\left(s_\tau \mid \pi\right)\right]\right]}_{\text{Epistemic value or information gain}} \end{aligned} \tag{4}$$

The above decomposition illustrates that the expected free energy of a policy can be rewritten in a form that illustrates its two parts: its extrinsic value and its epistemic values. This connects to the balance between exploration and exploitation mentioned earlier. One the one hand, maximizing the intrinsic value means expecting that the actions specified by the policy minimize the agent's surprise about future outcomes (exploitation). On the other hand, maximizing the epistemic value will reduce the uncertainty of the agent about the states of the world.

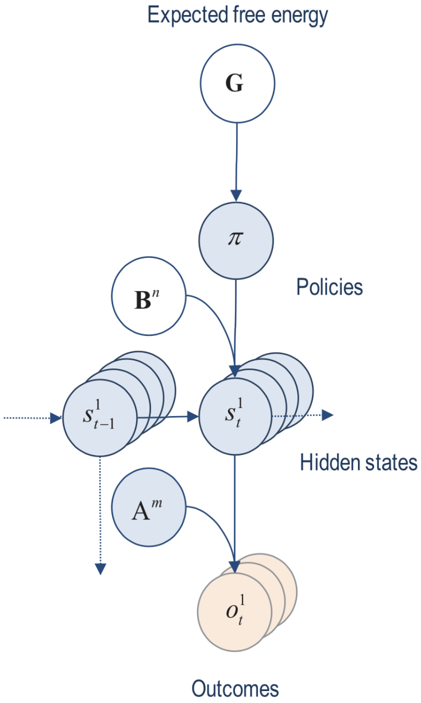


Figure 1: Bayesian network with a graphical illustration of the dependencies involved in the generative model (source [3]).

## 4 Task description

This section will explain the task introduced by Friston in [3] to provide a paradigm that is simple enough to describe formally while at the same time illustrate curiosity in terms of pursuing policies that allow novelty and epistemic learning. The task takes place as a simulation with an artificial agent, whose job is to report the correct color choice, between red, green and blue. The correct color changes from episode to episode, but it always follows the same rule. Three colored circles are arranged around a central fixation point and The agent can only look at one at a time and can direct its gaze in an exploratory manner over multiple trials (episodes) in order to figure out what the rule and increase its accuracy over episodes. All that is told to the agent at the beginning is that the color of the upper middle cue holds the necessary information.

Here is a more specific example to understand what is at stake. Figure 2 represents nine different episodes. At the very beginning, all the agent knows is that they must choose the correct color among the colors indicated by the large circle. The other thing it knows is that the location of the correct color depends on the color of the upper circle. Here, the rule that has to be inferred is the following:

- If the upper circle is **red**, then the correct color is the color of the circle on its *left* (2nd, 3rd, 8th examples);
- If the upper circle is **green**, then the correct color is the color of the upper circle (itself) (4th, 5th, 6th, 9th examples);
- If the upper circle is **blue**, then the correct color is the color of the circle on the *right* (1st, 7th examples)

An episode unfolds as follows: at the beginning, the gaze of the agent is directed towards the (empty) middle point. At each timestep, the agent can orient its gaze towards a colored circle. At the same time, it must report a color choice with a (virtual) button press. There is an "undecided" option that can be selected before they have gathered sufficient information. The agent should report the color as accurately as possible after looking at a maximum of three cues. The episode finishes after 6 time steps.

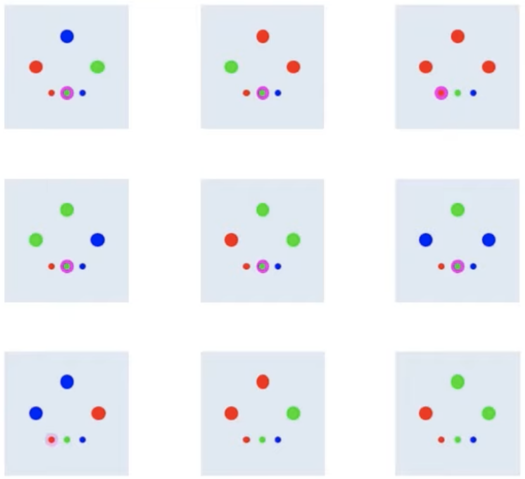


Figure 2: Illustration of the rule to learn for nine episodes. In each square, the three large circles indicate the cues that are presented to the agent and the three small circles indicate the color choices that are offered. (source [6])

The active inference model used for the task comprises several elements. There are the environment's hidden states, the outcomes that can be observed by the agent and action possibilities. Table 2 summarizes the various components of the task and how they are mapped in the model. Figure 3 represents more in details the task at hand.

| Model Elements | Task Component |
|---|---|
| Hidden states | Rule |
| | Color |
| | where |
| | choice |
| Outcomes | Colored visual cue (what) |
| | Proprioceptive cues indicating direction of gaze (where) |
| | Auditory cues providing feedback (feedback) |
| Action | gaze direction or color selection |
| Policy | sequence of gaze directions |

Table 2: Table of correspondence between the basic elements of an active inference model with the task

# 5 Implementation in Reinforcement Learning

This section contains the contribution done during this project. It consists in implementing the task with Deep Q-learning, a very well-known reinforcement learning technique that works well for discrete-action environment. The code for this implementation is available on GitHub [1]. Changing frameworks required some adaptations that will be exposed in this section. This work allowed to highlight key differences between the two approaches.

## 5.1 State encoding

The main adaptation that is required is to modify the state representation is twofold. First, the "outcomes" that are observed by the agent need to contain much more information in order to encode the markov property required by Q-learning. In reinforcement learning, a state representation must include all the information that is necessary to the agent at any time step to make a decision about the next action. Therefore, when an agent chooses to look at a location, instead of only perceiving the color of that location, as would be the case under the original task, the environment must return to the RL agent a state representation that contains all the colors of the cues that have been seen up to the current time, the current position of the gaze and the

[1] https://github.com/atrudel/rl_rule_learning

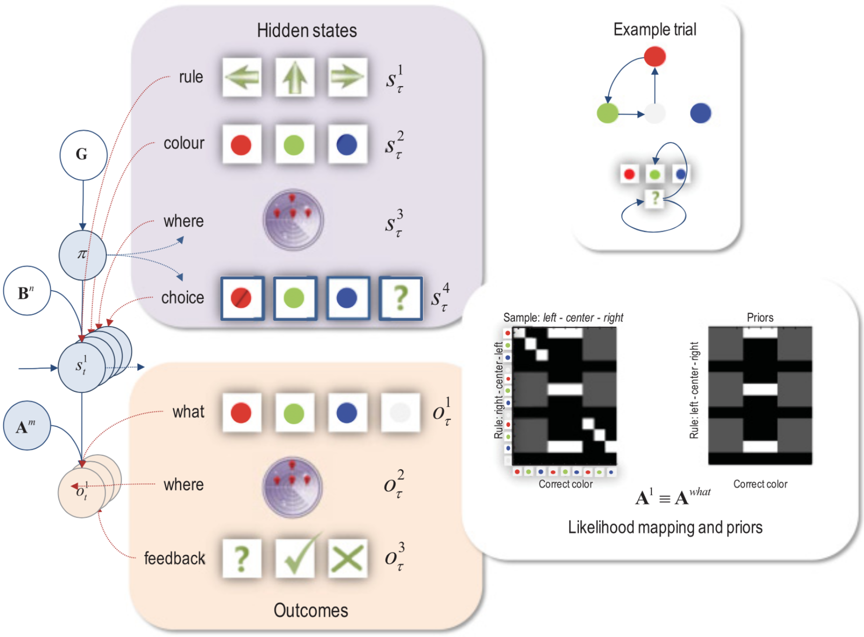
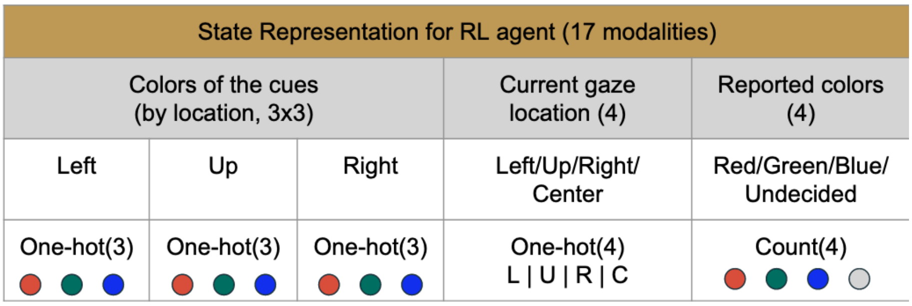


Figure 3: ***Left:*** Illustration of the bayesian model along with the associated elements. ***Upper right:*** Example of an episode (trial). The gaze is first directed to the upper circle in order to get the information that the left circle must be looked at next. The choice is undecided for the first timestep. Then, the gaze is directed to the left circle, which allows the agent to correctly select the green color. (source [3])

count of all the answers that it has given so far in the episode. Table 3 shows how these pieces of information are put together in a 17-dimensional array. We can really understand here what Friston means when he says that active inference allows to optimize policies as entire trajectories. Under the Deep Q-learning framework, it is quite hard to craft a state representation that allows the agent to plan its actions while taking into account its past attempts, along with whether it was right or wrong, without making it excruciatingly complex. Reinforcement learning conveys the reward to the agent separately from the state observation. It is useful to reinforce the probability of choosing an action given a state, but it makes it quite difficult to give the agent all the information necessary to choose an action based on past decisions and whether they were right or wrong.



<table>
<tr><th colspan="5">State Representation for RL agent (17 modalities)</th></tr>
<tr><th colspan="3">Colors of the cues<br>(by location, 3x3)</th><th>Current gaze<br>location (4)</th><th>Reported colors<br>(4)</th></tr>
<tr><td>Left</td><td>Up</td><td>Right</td><td>Left/Up/Right/<br>Center</td><td>Red/Green/Blue/<br>Undecided</td></tr>
<tr><td>One-hot(3)<br>● ● ●</td><td>One-hot(3)<br>● ● ●</td><td>One-hot(3)<br>● ● ●</td><td>One-hot(4)<br>L | U | R | C</td><td>Count(4)<br>● ● ● ○</td></tr>
</table>

Table 3: State representation for the Deep Q-learning agent as a 17-dimensional vector. The color of each cue is one-hot encoded in three values that are equal to zero, except the one corresponding to the cue's color once it has been seen. The current gaze location is also one-hot encoded. The reported colors are encoded as separate counts, each of which can have a value between 0 and 6 (maximum number of time steps of an episode).

The second main difference in state representation is that all explicit beliefs about hidden states within the agent completely disappear when we switch to reinforcement learning. Everything is encoded in the neural network that approximates the q-function. We only have to make sure that the neural network is large enough to handle the complexity of the function. The architecture chosen is a 3-layered perceptron with 20 hidden units on each hidden layer. The q-function is defined as the expected sum of cumulative discounted reward when choosing a specific action at a given state:

$$Q(s_t, a_t) = \arg\max_a \mathbf{E}\left[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \ldots | s_t, a_t\right]$$

Where $t$ is the current time step, $s_t$ the current state, $a_t$ the selected action, $R_t$ the reward received at each timestep and $\gamma \in [0, 1]$ the discounting factor. The next subsection will develop the rewards.

## 5.2 Reward

The reward is specific to reinforcement learning, and its equivalent in the original task is the *feedback* perceived outcome, along with the surprise associated with its value. The agent simply tries to predict the feedback accurately, the same way as its other perceptions. The surprise that was associated with a good answer was 0, a wrong answer triggered a surprise of 4, and an undecided choice past the 3rd timestep got a surprise of 8. We propose the following mapping for rewards:

| Color repored | Timestep | Reward |
|---|---|---|
| Correct | >0 | 0 |
| Wrong | >0 | -1 |
| Undecided | <=3 | 0 |
| Undecided | >3 | -2 |

Table 4: Reward received depending on the color reported and the timestep.

## 5.3 Exploration-exploitation trade-off

Active inference automatically regulates the curiosity of its agent by the mere process of trying to decrease uncertainty about its perceptions. So long as the environment is uncertain, it will gather information to diminish its uncertainty, and when outcomes get more certain, the desire to reduce the surprise over *feedback* outcomes kick in and the agent chooses the correct moves to figure out the rule and report the right answer every time, as shown in equation 4.

Reinforcement learning, on the other hand, does not handle this problem automatically, and this is known as the exploration-exploitation trade-off. The issue arises from the very definition of a policy, as it was exposed in the first section. Let us the definition of the policy as the arg max of the q-function:

$$\pi(s_t) = \arg\max_a Q(s_t | a_t)$$

Under this optimal policy, given parametrization of the q-function, we can expect the agent to always choose the same action at a given state. This means that it is not prone to explore other possibilities than the choice it currently favors. This poses a problem if we want the agent to find the best policy, because it entails choosing novel actions to see what reward they generate. The solution that is found is a quite artificial trick, called *epsilon-greedy policy.* It requires a hyperparameter $\epsilon \in [0, 1]$ which defines a probability that the agent will select a random action instead of following its optimal policy, at any timestep. Many decay techniques exist to start with a high random rate (epsilon) and diminish it over the episodes. Exponential decay was implemented for this task. This way, we start the training with high exploration when the policy is unreliable, and gradually move to more exploitation when sufficient experiences have given a good policy. It is useful to note that calibrating the epsilon value was relatively tideous, as it makes a big difference on the learning curve, and it is one more hyperparameter to tune on top of the other ones.

## 5.4 Results

Friston reported that its agent could act optimally after the 14th episode, at which point the uncertainty about the task is resolved. The idea of the current experiment is to verify how many episodes are necessary for a deep q-learning agent to reach perfect performance, if possible. The RL agent was expected to take longer to learn the task because its encoding of the environment is less sturctured. It turned out that, even while trying many hyperparameter settings, the agent couldn't reach perfect peformance. Figure 4 shows the results of a training over 100 episodes. We notice a slight improvement in reward and accuracy of prediction starting at episode 30, but it is still very unstable and nowhere near the level of accuracy of the active inference agent. Furthermore, due to the stochasticity of the reinforcement learning technique, the training curves could be radically different between trainings.

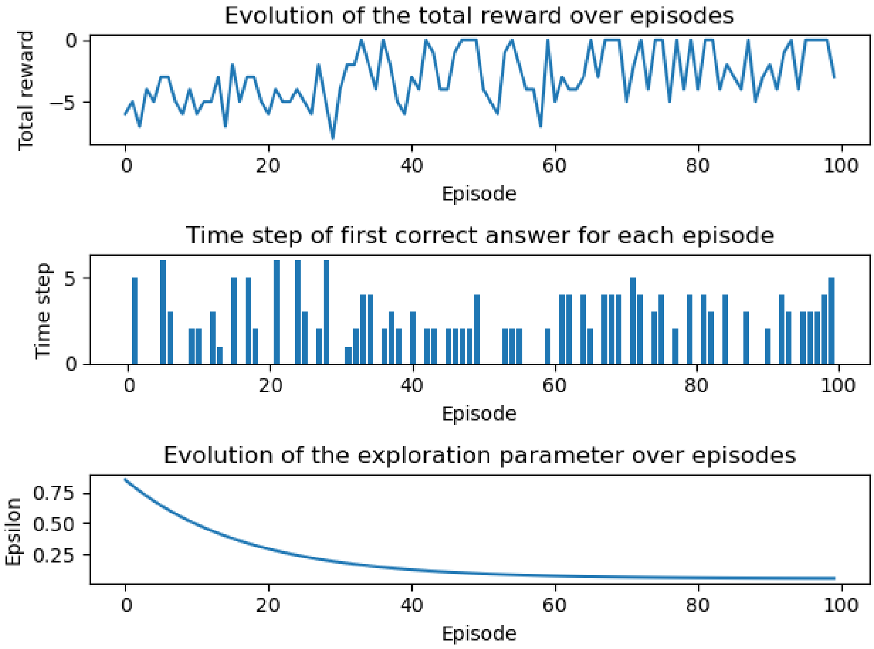


Figure 4: Results of the trainig of the RL agent over 100 episodes. The top graphs shows the total reward per episode. The second one shows the first time step of the episode where the agent reported the correct color. The third graph shows the decay of the exploration parameter of the epsilon-greedy policy.

These results may be due to an insufficient expressivity of the state representation chosen for the RL implementation. Maybe the neural network that approximates the q-function should be more powerful, and it is also possible that reinforcement learning is not suited for this task. In any case, active learning seems to be the best approach for this kind of rule-learning task.

## 6 Discussion

The great promise of Friston's theory is its ability to explain a wide range of phenomena, from perception to decision making. The idea that the brain is constantly generating and testing hypotheses about the world is intuitive and resonates with our experience of the world. Furthermore, the active inference framework provides a unifying theory that can integrate disparate findings from neuroscience and psychology. It has the potential to provide a powerful new tool for understanding the complex interplay between perception, action, and curiosity.

The major promise of the theory is its ability to encode complex preferences into the priors of the agent. This has potential applications at the moment, in the rush towards general artificial intelligence. Many researchers in AI safety [8] worry about misalignment, a problem in which agent trained with reinforcement learning develop interests that are misaligned with the intent of the humans who developed them. A common example is reward hacking, in which agents find a way to accumulate reward in a manner that was unforeseen by the developers. This is a little bit analogous to what we can observe with human behavior. If a strict rule is enforced to prohibit a behavior, people often bypass the rule, and it is a challenge to craft a system of incentives that shapes behavior at scale. The same way that psychologists talk about intrinsic motivation [7] as a better approach, it is possible that active inference could be an alternative to reward and value-based learning and a pathway to communicate better the expected behavior to the agent by finding a way to encode our instructions with their priors.

The theory, however, remains hard difficult to implement because of the complexity of the modelling of all the modalities of the hidden states of the world. For the simple task used in Friston's paper, it is doable, but it can easily become intractable as soon as we move to realistic environments in which robots may be immersed for example. The problem is twofold: the complexity of the model, especially when considering many hidden state modalities, makes it hard to build, and even if we could model it, bayesian statistics are generally very expensive to compute. Finally, prior knowledge is both a blessing and a curse, in the sense that it can be used to accurately shape an agent's behavior, but at the same time can be problematic in cases where it is not available. This is why Machine Learning tends to be used on large datasets without needing to encode any prior knowledge.

## 7 Conclusion

We have exposed Karl Friston's active inference theory and how it can efficiently model a relatively complex rule learning task in an environment with few varying components. The structure that bayesian learning approach gives to the search space allows an incredibly efficient sampling of the environment in order to reduce the uncertainty while at the same time eliciting the expected behavior that has been encoded in the agent's prior preferences. We have shown that deep q-learning, an alternative approach based on reinforcement learning, is incapable of picking up the subtleties of the task in a reasonable amount of episodes. Friston's approach seems much closer to human cognitive processes. We shall see in the future if we manage to use it in problems where it is currently intractable.

## References

[1] Thomas H. B. FitzGerald, Philipp Schwartenbeck, Michael Moutoussis, Raymond J. Dolan, and Karl Friston. Active inference, evidence accumulation, and the urn task. *Neural Computation*, 27(2):306–328, February 2015. Citation information: 'Active Inference, Evidence Accumulation, and the Urn Task No Access', Thomas H. B. FitzGerald, Philipp Schwartenbeck, Michael Moutoussis, Raymond J. Dolan, Karl Friston, Neural Computation February 2015, Vol. 27, No. 2: 306–328.

[2] Karl J. Friston, Rick A Adams, and P. Read Montague. What is value—accumulated reward or evidence? *Frontiers in Neurorobotics*, 6, 2012.

[3] Karl J. Friston, Marco Lin, Chris D. Frith, Giovanni Pezzulo, J. Allan Hobson, and Sasha Ondobaka. Active inference, curiosity and insight. *Neural Computation*, 29:2633–2683, 2017.

[4] Karl J. Friston, Francesco Rigoli, Dimitri Ognibene, Christoph D. Mathys, Thomas H. B. FitzGerald, and Giovanni Pezzulo. Active inference and epistemic value. *Cognitive Neuroscience*, 6:187 – 214, 2015.

[5] Karl J. Friston, Philipp Schwartenbeck, Thomas FitzGerald, Michael Moutoussis, Timothy Behrens, and Raymond J. Dolan. The anatomy of choice: active inference and agency. *Frontiers in Human Neuroscience*, 7, September 2013.

[6] Karl Friston. Karl Friston: Active inference and artificial curiosity. Online video, 2017.

[7] Deci E. L. Ryan, R. M. elf-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist*, 55:68–78, 2000.

[8] Vladimir Mikulik Matthew Rahtz Tom Everitt Ramana Kumar Zac Kenton Jan Leike Shane Legg Victoria Krakovna, Jonathan Uesato. Specification gaming: the flip side of ai ingenuity. Deepmind Research Blog, 2020.