Ideal observers

Gonçalo Guiomar

June 8, 2025

1 Introduction

An ideal Bayesian observer is a hypothetical agent that satisfies four key properties. First, it possesses an exact generative model of the task environment. Second, it assigns a coherent prior distribution over every latent variable in that model. Third, it systematically updates its beliefs using Bayes' rule after each observation. Fourth, it makes decisions by optimizing a clearly stated utility function, typically involving either maximum-posterior estimation or minimum expected loss criteria, with random policies also being a possibility.

1.1 Bayesian Inference Framework

The mathematical foundation of Bayesian inference rests upon the systematic treatment of uncertainty through probability distributions. We denote the latent vector representing hidden states as $\mathbf{z}$, while observations are represented by $\mathbf{x}$, where each $\mathbf{x}_t$ captures the temporal structure of incoming data at time t. The ideal Bayesian observer operates with a complete generative model specified by the joint distribution $p(\mathbf{x}, \mathbf{z} \mid \boldsymbol{\theta})$ over observations and latent variables, where $\boldsymbol{\theta}$ represents the model parameters that govern the probabilistic relationships within the environment.

The cornerstone of Bayesian reasoning is Bayes' theorem, which provides the mathematical framework for updating beliefs in light of new evidence:

$$p(\mathbf{z} \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid \mathbf{z})p(\mathbf{z})}{p(\mathbf{x})} \tag{1}$$

where each component carries specific meaning within the inference process. The posterior distribution $p(\mathbf{z} \mid \mathbf{x})$ represents the updated beliefs about the hidden states after observing data $\mathbf{x}$. The likelihood function $p(\mathbf{x} \mid \mathbf{z})$ quantifies how probable the observed data is under each possible configuration of hidden states, effectively measuring the compatibility between observations and hypotheses. The prior distribution $p(\mathbf{z})$ encodes initial beliefs about the hidden states before observing any data, capturing background knowledge or assumptions about the problem structure. The evidence or marginal likelihood $p(\mathbf{x}) = \sum_{\mathbf{z}} p(\mathbf{x} \mid \mathbf{z})p(\mathbf{z})$ serves as a normalization constant ensuring that the posterior distribution integrates to unity.

1.2 Sequential Belief Updating

In temporal settings where observations arrive sequentially, the Bayesian framework naturally accommodates the accumulation of evidence over time. At each time step t, the agent receives a new observation X_t and must update its beliefs about the hidden state. The sequential nature of this process follows the recursive relationship:

$$p_t(z) = \frac{p(X_t \mid z, \boldsymbol{\theta})\, p_{t-1}(z)}{\sum_{z'} p(X_t \mid z', \boldsymbol{\theta})\, p_{t-1}(z')} \tag{2}$$

where $p_t(z)$ denotes the posterior distribution over the hidden state z after t observations, $p(X_t \mid z, \boldsymbol{\theta})$ represents the likelihood of the current observation given state z and parameters $\boldsymbol{\theta}$, and $p_{t-1}(z)$ captures the prior beliefs from the previous time step. The denominator ensures proper normalization across all possible states z'.

This recursive formulation demonstrates how each new observation systematically refines the agent's beliefs, with the posterior from one time step serving as the prior for the next. The initial prior $p_0(z) = p(z)$ establishes the starting point for this iterative process, while subsequent updates incorporate the cumulative evidence from all observed data.

1.3 Decision Theory and Utility Maximization

The ideal Bayesian observer extends beyond mere belief updating to encompass principled decision-making under uncertainty. Given a posterior distribution $p_t(z)$ over hidden states, the agent must select actions that optimize some criterion of performance. This decision-theoretic component introduces utility functions $U(a, z)$ that specify the value or reward associated with taking action a when the true state is z.

The expected utility of an action a given the current beliefs is computed as:

$$\mathbb{E}[U(a, z) \mid X_{1:t}] = \sum_{z} U(a, z)\, p_t(z) \tag{3}$$

where $X_{1:t} = \{X_1, X_2, \ldots, X_t\}$ represents the sequence of observations up to time t. The optimal action under this framework is the one that maximizes expected utility:

$$a^* = \arg\max_{a} \mathbb{E}[U(a, z) \mid X_{1:t}] \tag{4}$$

Alternative decision criteria include maximum a posteriori (MAP) estimation, where the agent selects the action corresponding to the most probable state:

$$a^*_{\text{MAP}} = \arg\max_{z} p_t(z) \tag{5}$$

and minimum expected loss, where the agent minimizes expected cost rather than maximizing expected utility. The choice of decision criterion depends on the specific task requirements and the agent's risk preferences.

1.4 Predictive Distributions and Information Theory

Beyond state estimation and decision-making, the Bayesian framework provides tools for predicting future observations and quantifying information content. The predictive distribution for the next observation X_{t+1} given the history $X_{1:t}$ is obtained by marginalizing over the current posterior:

$$p(X_{t+1} \mid X_{1:t}) = \sum_z p(X_{t+1} \mid z, \boldsymbol{\theta})\, p_t(z) \tag{6}$$

This predictive distribution captures the agent's uncertainty about future observations and forms the basis for information-theoretic measures such as entropy and mutual information. The entropy of the predictive distribution quantifies the expected surprise or information content of the next observation:

$$H[X_{t+1} \mid X_{1:t}] = -\sum_x p(X_{t+1} = x \mid X_{1:t}) \log p(X_{t+1} = x \mid X_{1:t}) \tag{7}$$

These information-theoretic quantities become particularly relevant in active inference settings, where agents must choose actions to maximize information gain or minimize predictive uncertainty.

The ideal Bayesian observer framework thus provides a comprehensive mathematical foundation for reasoning under uncertainty, encompassing belief updating, decision-making, and information processing within a unified probabilistic framework. This foundation underlies all subsequent developments in this work, from the specific bias detection tasks to the active inference and reinforcement learning extensions.

2 Bias Detection Task

The bias detection task is based on K possible sampling locations out of which one can be sampled at each round t. At each sample of a cue the agent can observe one of two possible outcomes, denoted as $X_t \in \{0, 1\}$. The environment contains a bias such that one specific location z^* (the target location) exhibits a higher probability of producing an outcome out of the possible samples in X_t compared to all other locations. Specifically, when probing the target location z^*, the probability of observing outcome 1 is p_T, while probing any non-target location yields 1 with probability p_F, where typically $p_T > p_F$. The remaining $K - 1$ locations are statistically equivalent, each producing outcomes according to the same uniform probability p_F.

The initial prior distribution over the hidden state z assigns equal probability to each location, reflecting the absence of preferential information:

$$p_0(z) = \frac{1}{K} \quad \text{for every location index } z = 0, \ldots, K-1 \tag{8}$$

where z represents the true target location and K denotes the total number of possible locations.

2.1 Task Variants: Standard and Hidden Cues

The bias detection framework encompasses two distinct variants that differ in the availability of cues at each decision point. In the standard variant, all K locations remain accessible throughout the trial, providing complete observability of the environment. Conversely, the hidden cues variant introduces partial observability by restricting the agent's choices to a randomly selected subset of locations at each round.

For the standard bias detection task, the set of available cues $\mathbf{s}_t$ encompasses all possible locations at every time step:

$$\mathbf{s}_t = \{0, 1, \ldots, K-1\} \quad \text{for all } t \tag{9}$$

where $\mathbf{s}_t$ represents the set of cues available for selection at round t. The cue selection policy follows a uniform distribution over all available options:

$$p(C_t = c \mid \mathbf{s}_t) = \frac{1}{|\mathbf{s}_t|} \tag{10}$$

where C_t denotes the selected cue at round t, c represents a specific cue location, and $|\mathbf{s}_t|$ indicates the cardinality of the available cue set.

In the hidden cues variant, the available cue set $\mathbf{s}_t$ constitutes a proper subset of all possible locations, sampled uniformly at each round. The number of available cues typically ranges from a minimum value $n_{\min}$ to a maximum value $n_{\max}$, where $1 \leq n_{\min} \leq n_{\max} \leq K$. At each round t, the environment first determines the number of available cues n_t according to:

$$p(n_t = n) = \frac{1}{n_{\max} - n_{\min} + 1} \quad \text{for } n \in \{n_{\min}, \ldots, n_{\max}\} \tag{11}$$

Subsequently, the specific cues comprising $\mathbf{s}_t$ are selected uniformly from all possible $\binom{K}{n_t}$ combinations of n_t locations. The cue selection policy remains uniform over the restricted set:

$$p(C_t = c \mid \mathbf{s}_t) = \frac{1}{|\mathbf{s}_t|} \quad \text{where } c \in \mathbf{s}_t \tag{12}$$

The color likelihood function captures the probabilistic relationship between cue location, true target location, and observed color outcome for both variants:

$$p(X_t = 1 \mid C_t = c, z = z^*, \boldsymbol{\theta}) = \begin{cases} p_T & \text{if } z^* = c \\ p_F & \text{if } z^* \neq c \end{cases} \tag{13}$$

where X_t represents the binary outcome at round t, C_t denotes the selected cue, z^* indicates the true target location, and $\boldsymbol{\theta} = \{p_T, p_F\}$ encompasses the environment parameters.

2.2 Sequential Bayesian Update

For a single observation (c, x), the per-hypothesis likelihood function determines how well each possible location explains the observed data:

$$\ell(z) = p(X_t = x \mid C_t = c, z = z^*, \boldsymbol{\theta}) = \begin{cases} p_T & \text{if } z = c, x = 1 \\ 1 - p_T & \text{if } z = c, x = 0 \\ p_F & \text{if } z \neq c, x = 1 \\ 1 - p_F & \text{if } z \neq c, x = 0 \end{cases} \tag{14}$$

where $\ell(z)$ represents the likelihood of observing outcome x when selecting cue c, given that the true target location is z.

The temporal evolution of beliefs as a consequence of each sample follows the recursive Bayesian update rule. After t observations, the posterior distribution is computed as:

$$p_t(z) = \frac{\ell(z)\, p_{t-1}(z)}{\sum_{z'} \ell(z')\, p_{t-1}(z')} \tag{15}$$

where $p_0(z) = p(z)$ represents the initial prior distribution and z' denotes the summation variable over all possible target locations. This update mechanism ensures that each new observation systematically refines the agent's beliefs about the target location.

2.3 MAP Agent

A Maximum A Posteriori (MAP) agent represents a specific instantiation of the ideal Bayesian observer framework that implements a particular decision strategy. Rather than minimizing expected loss across all possible outcomes, the MAP agent adopts a simplified approach by selecting the action corresponding to the hypothesis with maximum posterior probability.

For the bias detection task described earlier, the MAP agent implements its decision rule by selecting:

$$\pi(a|s_t) = \arg\max_z p_t(z) \tag{16}$$

after accumulating t observations, where $\pi(a|s_t)$ represents the policy function that maps the current state to an action, and a denotes the selected action. This choice corresponds to the location with the highest posterior probability, representing the agent's best single estimate of the target's true position.

3 Active Inference in Bias Detection

Having established the bias detection framework with both standard and hidden cues variants, we now examine how agents can actively select cues to optimize their information gathering. The transition from passive observation to active cue selection introduces strategic considerations that extend beyond simple

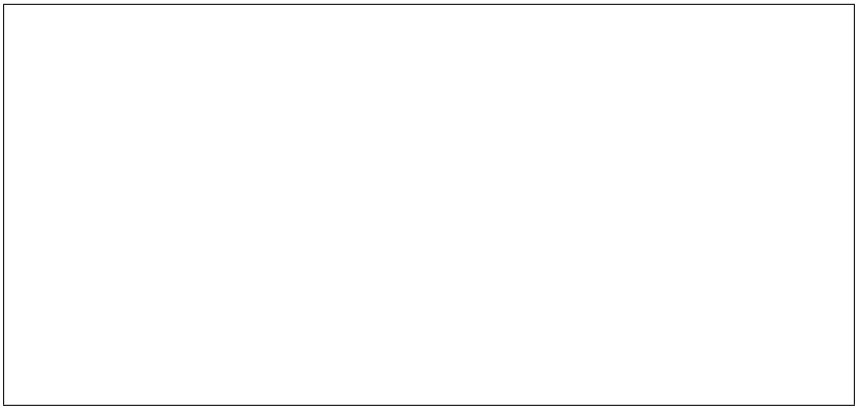

Figure 1: Placeholder Figure

Bayesian updating. In this active inference setting, agents must balance exploration of uncertain hypotheses against exploitation of current beliefs, particularly when operating under the constraints of partial observability introduced by the hidden cues variant.

The active inference framework builds upon the hidden cues variant of the bias detection task, where at each round t the agent observes the available cue set $\mathbf{s}_t \subseteq \{0, \ldots, K-1\}$ and must select an action $a_t \in \mathbf{s}_t$ that corresponds to the chosen cue C_t. The hidden state z^* represents the true target location, while the restricted cue set $\mathbf{s}_t$ captures the realistic constraint of partial observability at each decision point.

3.1 Adding Surprise to the Bayes Agent

The incorporation of surprise transforms classical Bayesian decision-making into an active inference framework. The surprise term $S_t = -\log Z_t$ quantifies the unexpectedness of observations under the current belief state, providing a natural information-theoretic objective for bias prediction tasks.

In standard Bayesian settings, agents typically maximize expected utility or minimize prediction error. However, systems operating under the free-energy principle minimize surprise to maintain coherent internal models and avoid potentially misleading deviations from expected patterns.

The *expected raw-surprise objective* is:

$$\mathcal{G}(a_t) = \mathbb{E}_{X_t \mid a_t, \mathbf{s}_t}\big[S_t\big] = -\sum_{x \in \{0,1\}} p(x \mid a_t)\, \log p(x \mid a_t), \tag{17}$$

where

$$p(x \mid a_t) = \sum_{z} p(x \mid a_t, z)\, p_{t-1}(z). \tag{18}$$

This objective represents the entropy of the predictive distribution $p(x \mid a_t)$, weighted by current beliefs $p_{t-1}(z)$. Actions that lead to high-confidence

predictions (low entropy) are preferred, as they minimize expected surprise. This creates a natural preference for confirmatory evidence while still allowing for belief updates when surprising observations occur.

The mathematical structure reveals key properties: it is invariant to belief rescaling, exhibits diminishing returns as beliefs concentrate, and naturally balances confirmation of existing beliefs with openness to revision. When beliefs are uniform, all actions yield identical expected surprise; when beliefs are concentrated, the policy strongly prefers the most likely state if available.

The surprise-minimizing policy implements the core principle of active inference: agents act to maintain their predictions about the world. This policy emerges from variational free-energy minimization, where the agent seeks to minimize divergence between its internal model and environmental observations.

Policy (soft-min over available cues):

$$\pi_A(a_t \mid \mathbf{s}_t) = \frac{\exp\big[-\beta\, \mathcal{G}(a_t)\big]}{\sum_{a' \in \mathbf{s}_t} \exp\big[-\beta\, \mathcal{G}(a')\big]} \tag{A-1}$$

with inverse-temperature $\beta > 0$. The policy chooses cues whose outcome distributions are *most predictable* under current beliefs.

The policy exhibits several important limiting behaviors:

- When beliefs are uniform ($p_{t-1}(z) = 1/K$ for all z), all available actions yield identical expected surprise, leading to uniform selection over $\mathbf{s}_t$.
- When beliefs are highly concentrated on state z^*, the policy strongly prefers $a_t = z^*$ if available in $\mathbf{s}_t$.
- The temperature parameter β modulates exploration: low β encourages broad sampling, while high β enforces focused, exploitation-like behavior.

As $\beta \to \infty$, the agent becomes deterministically surprise-minimizing, while $\beta \to 0$ yields uniform random selection. This provides a principled approach to active learning under uncertainty, naturally balancing confirmation with exploration.

3.2 Entropy-Minimizing Decision Policy

While the surprise-minimizing policy governs cue selection during the exploration phase, the active inference agent must also implement a principled decision strategy for the final choice of target location. Rather than employing the standard Maximum A Posteriori (MAP) criterion, we introduce an entropy-minimizing decision policy that remains consistent with the active inference framework's emphasis on uncertainty reduction.

The entropy-minimizing decision policy extends the surprise minimization principle from action selection to final decision-making. Instead of simply choosing the location with highest posterior probability, the agent evaluates each potential decision based on its expected impact on remaining uncertainty.

For each possible decision $d \in \{0, 1, \ldots, K-1\}$, the agent computes the expected entropy that would remain after making that decision:

$$\mathcal{H}(d) = H[p_t(z)] \cdot (1 - p_t(d)) \tag{19}$$

where $H[p_t(z)] = -\sum_z p_t(z) \log p_t(z)$ represents the current posterior entropy and $p_t(d)$ denotes the posterior probability assigned to location d. This formulation captures the intuition that decisions with higher confidence (larger $p_t(d)$) should result in lower expected remaining entropy.

The entropy-minimizing decision rule selects the action that minimizes this expected entropy:

$$d^*_{\text{entropy}} = \arg\min_d \mathcal{H}(d) = \arg\min_d \left[H[p_t(z)] \cdot (1 - p_t(d))\right] \tag{20}$$

This is equivalent to maximizing the decision confidence $p_t(d)$, but the entropy-based formulation provides a more principled information-theoretic foundation that aligns with the active inference framework.

4 Reinforcement Learning Agent

We retain the original delayed-reward setting with episode length T and sampling rounds ($t = 1, \ldots, T-1$) where the agent selects a cue but receives no reward i.e. $r_{t+1} = 0, \quad t \in \{1, T-1\}$ and final decision round ($t = T$) where the agent selects a final guess and receives:

$$r_{T+1} = R \cdot \mathbb{I}[a_t = z^\star] \tag{21}$$

where $z^\star$ is the true hidden target location (the biased cue).

We define a Q-function that estimates cumulative discounted reward $Q^{\mathrm{R}}(s, a)$ and update it by the usual temporal-difference (TD) update

$$\delta^{\mathrm{R}}_t = r_{t+1} + \gamma \max_{a'} Q^{\mathrm{R}}(s_{t+1}, a') - Q^{\mathrm{R}}(s_t, a_t) \tag{22}$$

$$Q^{\mathrm{R}}(s_t, a_t) \leftarrow Q^{\mathrm{R}}(s_t, a_t) + \alpha \cdot \delta^{\mathrm{R}}_t \tag{23}$$

4.1 Defining a Separate Surprise Signal

During each sensing step ($t < T$), compute the Bayesian surprise as the KL divergence between posterior and prior:

$$B_t = \mathrm{KL}\left[p_t(z) \,\|\, p_{t-1}(z)\right] = \sum_z p_t(z) \log \frac{p_t(z)}{p_{t-1}(z)} \tag{24}$$

To encourage belief stability, define a shaped auxiliary reward:

$$c_t = -B_t \qquad \text{(only for } t < T\text{)} \tag{25}$$

No intrinsic reward is given at the decision step:

$$c_T = 0 \tag{26}$$

We introduce a second value function to track this signal:

$$Q^{\mathrm{S}}(s, a) \quad \text{(S for "surprise")} \tag{27}$$

Its TD update is:

$$\delta_t^{\mathrm{S}} = c_t + \gamma \max_{a'} Q^{\mathrm{S}}(s_{t+1}, a') - Q^{\mathrm{S}}(s_t, a_t) \tag{28}$$

$$Q^{\mathrm{S}}(s_t, a_t) \leftarrow Q^{\mathrm{S}}(s_t, a_t) + \alpha \cdot \delta_t^{\mathrm{S}} \tag{29}$$

The KL term is thus entirely outside the external-reward channel, yet fully learnable through a standard TD rule.

We compute a combined scalar score:

$$F_\lambda(s, a) = Q^{\mathrm{R}}(s, a) + \lambda \cdot Q^{\mathrm{S}}(s, a) \tag{30}$$

Here $\lambda \geq 0$ controls how much weight the agent places on surprise minimisation.

A softmax policy over the available action set $\mathbf{s}$ becomes:

$$\pi_{\beta,\lambda}(a \mid s) = \frac{\exp\left[\beta F_\lambda(s, a)\right]}{\sum_{a' \in \mathbf{s}} \exp\left[\beta F_\lambda(s, a')\right]} \tag{31}$$

where β controls the exploration–exploitation trade-off.