# Implementing Human Information-Seeking Behaviour with Action-Agnostic Bayesian Surprise

**Emmanuel Daucé**
Institut de Neurosciences de la Timone
Centrale Méditerranée
Marseille, France
emmanuel.dauce@univ-amu.fr

**Hamza Oueld**
Institut de Neurosciences de la Timone
Aix-Marseille Univ / CNRS
Marseille, France
hamza.oueld-kaddour-el-hallaloui@univ-amu.fr

**Andrea Brovelli**
Institut de Neurosciences de la Timone
CNRS
Marseille, France
andrea.brovelli@univ-amu.fr

## Abstract

In this paper, we aim to establish a link between model learning and the mechanism of curiosity. The main hypothesis developed is that exploration bonuses, as proposed in the reinforcement learning literature, are linked to Bayesian estimation principles through the construction of a parametric model of the causal relationships between actions and observations. At odd with the classic action-conditional Bayesian surprise widely used in the "curiosity" literature, action is here treated as an external variable, unknowingly of the agent's own control policy. It is thus called the "agnostic" Bayesian surprise (ABS), interpreted as an estimate of the information transfer between the observed data (including observations and motor commands) and the model parameters. We present here the general guidelines of this approach, and show results suggesting that action selection guided by information transfer can account for certain experimental, behavioral, and neurological data in humans.

## 1 Introduction

Our environment is full of unpredictable events, and the Bayesian brain hypothesis (Knill and Pouget, 2004) suggests that our brains build models to better predict these events. This assumption relies on the idea that the sensory environment acts as a source of random events, against which the brain would construct probabilistic models that would enable it to better predict and anticipate these events (Von Helmholtz, 1867). If model estimation and prediction seem to constitute the mechanism through which sensory data is processed (Rao and Ballard, 1999; L Griffiths et al., 2008; Doya, 2007; Fiser et al., 2010), it seems, in a more fundamental way, that this same principle could also be applied to action selection (Friston, 2010). While early works, such as Kalman (1960), established a link between estimation and behavior optimization, the specific issue of action selection focusing on knowledge acquisition (i.e. without considering rewards, penalties, or explicit goals) has only recently been explored under the broader framework of "curiosity" models (Baldassarre, 2011; Gottlieb et al., 2013). In short, this involves designing a policy whose objective is to best predict the environment's responses to the agent's actions. Within this framework, two main types of approaches can be identified: (i) perception-centric theories, such as the Bayesian surprise maximization (Itti and Baldi, 2009) and Variational Free Energy minimization principle (Friston, 2010), focusing on observation

Accepted at the Intrinsically Motivated Open-ended Learning workshop at NeurIPS 2024.

prediction mismatch to orient behaviour, and (ii) action-centric theories, inspired by Gibson (1979) and Varela et al. (1991), such as the "Empowerment" principle (Klyubin et al., 2005), rooted on Shannon's Information theory, and focusing on self-assessment, i.e. probing environment responses to our own action.

From the action-centered perspective, by executing motor commands, the brain itself becomes a source of sensory changes and, more broadly, a source of random events. Consequently, actions are not merely responses to sensory stimuli, but integral components of the variations observed in sensory data. There is curiously, as far as we know, little attempt to conceptually concile the generative view on action selection view with the Bayesian estimation framework. One possible reason could be the difficulty to formalize, in terms of of Bayesian optimization, the production of actions with no other purpose than the generation of information itself. For instance, the many conceptual frameworks used in neuroscience either emphasize (i) the role of surprise in modulating learning (i.e. the post-hoc discrepancy between prediction and current observation) (Gläscher et al., 2010; Liakoni et al., 2022), or (ii) assessing the prediction of uncertainty (Angela and Dayan, 2005; Kobayashi and Kable, 2024), in the form, e.g., of an "expected information gain" (Oaksford and Chater, 1994; Little and Sommer, 2013; Xu et al., 2021), without considering the *production* of uncertainty itself. Revisiting the fundamental concepts of Bayesian decision and information theory, we thus propose here a decision-making mechanism in which the consideration of action embedding in the generative model may conduct to maintaining surprise and diversity through actions, i.e. continually seeking for action-outcome relationships that depart from the baseline. The agent may ultimately implement a policy that approximates Shannon's capacity on the action/outcome channel (Klyubin et al., 2005), providing a consistent level of curiosity throughout the learning process.

## 2 Formal model

The approach adopted in this article thus involves a detailed analysis of a minimal estimation problem where the agent is motivated by providing an accurate estimate of action outcomes rather than maximizing the expected rewards, in the presence of noise. Assume an agent having to learn the effect of its action by interacting with the environment. During learning, the agent will do two things : (i) build a statistical model of its environment and (ii) update its behavior (selection of actions). The first element we need to consider is thus describing a mechanism of model estimation. The Bayesian estimation framework allows to formalize the principles of an "ideal observer" (Geisler, 1989). Consider a series of actions and outcomes, any action being a cause of sensory change, with a data model $\theta$ that can be refined from observing both the actions $a$'s and observations $o$'s, using Bayes rule. In a temporal sequence, the generative model (set of) parameters $\theta_t$ is conditionally dependent on $a_t$ (the current action), on $o_t$ (the current observation), and on the history of past observations $h_{t-1} = \{a_1, o_1, ..., a_{t-1}, o_{t-1}\}$ (through the chain rule), i.e. $\theta_t \sim q(\Theta|a_t, o_t, h_{t-1})$ (fig. 1A). Importantly, by construction, an observer does not intervene onto the data it observes. It is a pure passive method of estimation.

Then, the objective of the controller is to provide relevant sensory data to the model, in order to have effective model updates. Before action selection, one can thus use the generative model to make a prediction about the effect of each action, and choose the one that contributes to minimize the error (or the surprise). It becomes obvious however (a well known caveat) that minimizing the prediction error alone, as an ideal observer would do, *should not be* the objective of the controller, because minimizing the prediction error generally implies minimizing the action diversity (dark room problem, see Friston et al. (2012)). A general solution consists in taking into account duality structure of estimation and control (Todorov, 2008), that allows to formulate action selection as a min-max problem, that is minimizing the prediction error in the "worst case" of action intervention, which conducts to *maximize* the "Bayesian surprise" (Itti and Baldi, 2009), classically written as a conditional *Information Gain*:

$$\max_{a_t; o_t \sim p(o_t|a_t, h_{t-1})} \mathrm{I}(\Theta_t; o_t|a_t, h_{t-1}) = \mathbb{E}_{\theta_t \sim q(\theta|a_t, o_t, h_{t-1})} \log q(\theta_t|a_t, o_t, h_{t-1}) - \log q(\theta_t|a_t, h_{t-1}) \quad (1)$$

often referred as the "Learning Progress", pivotal in the curiosity literature (Oudeyer et al., 2007; Schmidhuber, 2010; Barto et al., 2013; Houthooft et al., 2016; Achiam and Sastry, 2017; Pathak et al., 2017; Mazzaglia et al., 2022). Importantly, consistently with the "Thompson sampling" approach (Thompson, 1933), the parametric model itself is a random variable, whose distribution changes over time. On the predictive side, the information gain can be considered as an estimator of the *mutual information* $I(\Theta_t; O_t|a_t, h_{t-1})$, between a random variable $\Theta_t$ and a random variable $O_t$,

conditionally on $a_t$ and $h_{t-1}$. Interestingly, this quantity tends to decrease over time as the model improves.

Taking now the action-centered perspective, the action $a_t$ should also be considered as a random variable, generated from the (stochastic) controller $\pi_t$. The optimization problem should then be expressed the following way :

$$\max_{\pi_t; a_t, o_t \sim p(a_t, o_t | h_{t-1})} \mathrm{I}(\Theta_t; a_t, o_t | h_{t-1}, \pi_t) = \mathbb{E}_{\theta_t \sim q(\theta | a_t, o_t, h_{t-1})} \log q(\theta_t | a_t, o_t, h_{t-1}) - \log q(\theta_t | h_{t-1}, \pi_t) \tag{2}$$

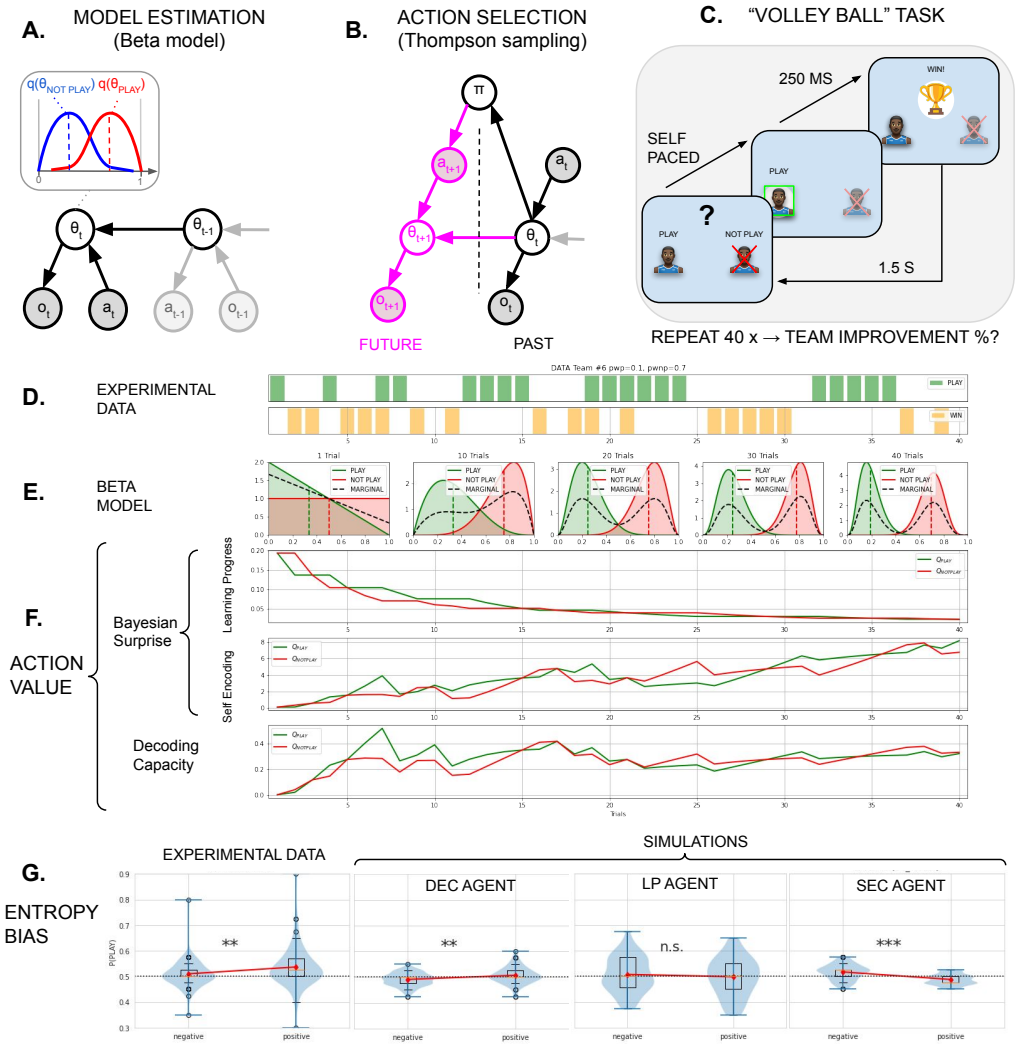


Figure 1: **A. Ideal Observer (parameter estimation)**. Action $a$ and outcome $o$ are the observed variables. A belief on the Bernoulli parameter $\theta$ is inferred from the past estimation and the current observations. **B. Ideal Participant (predictive model of action selection)**. At step $t$, the generative model predicts the next observations from Thompson sampling, allowing to select the next action according to an MLE objective. **C. Volley Ball task** (see text). **D.** A sequence of actions and outcomes observed in experiment (here $p(W|P) = 0.1, p(W|nP) = 0.7$). **E.** Bernoulli parameter inference using a Beta-model, after trials 1, 10, 20, 30 and 40. **F. Example action-value functions** Top : action conditional Bayesian Surprise, aka "Learning Progress" (LP). Middle: Self-encoding Capacity (SEC). Bottom: Outcome divergence, interpreted as Decoding Capacity (DEC). **G. Action selection biases in models and experiments**. Left : A significant entropy bias (preference for higher entropy) is observed in experiments (14 subjects). Right : in simulation, this positive bias is only reproduced with the DEC action value. LP shows no significant bias while while the SEC action value provides a negative selection bias.

with $\pi_t$ a distribution over actions (a policy) and the couple $(a_t, o_t)$ taking the role of a new observational data. The information gain should now be considered as an estimator of the mutual information $I(\Theta_t; A_t, O_t|h_{t-1})$, quantifying the *transfer of information* between the data and the model parameters, i.e. telling how much knowledge about the statistics of the data is obtained from choosing $a_t$ and reading $o_t$.

In our setup, $\theta_t$ plays the role of an auxiliary variable that conveys information from $a_t$ toward $o_t$. In information theory, the passing from $a_t$ toward $\theta_t$ corresponds to an *encoding* while the passing from $\theta_t$ toward $o_t$ corresponds to a *decoding*. This allows to decompose the Bayesian surprise into $I(\Theta_t; o_t|a_t, h_{t-1})$ (the learning improvement) and $I(\Theta_t; a_t|h_{t-1})$, that appears to be the "self-encoding" information, i.e. the encoding of the agent own actions into the parameters of the generative model. That means, in short, considering actions as external data, i.e. ignoring they were internally generated, thus the term "Agnostic" Bayesian Surprise.This self-encoding estimate appears to play a pivotal role promoting a form of self-sustained information seeking found in experiments.

Going one step further, our approach also implies using Thompson Sampling at the decision step, which means selecting an action based on a sample of the model parameters, i.e. $\tilde{\theta}_t \sim q(\theta_t|a_t, h_{t-1})$, instead of computing the integral, while the outcome (inferred from the generative model) is yet to be determined. This allows to consider, in turn, a complementary "decoding capacity" estimator, putting the focus on the outcome distribution rather than the model parameters, i.e.:

$$\max_{a_t;\tilde{\theta}_t\sim q(\theta|a_t,h_{t-1})} \mathrm{I}(O_t;\tilde{\theta}_t|h_{t-1}) = \mathbb{E}_{o_t\sim p(o_t|\tilde{\theta}_t,h_{t-1})} \log p(o_t|\tilde{\theta}_t, h_{t-1}) - \log p(o_t|h_{t-1}) \tag{3}$$

Both formulas (2) and (3) contain a positive term reflecting a *precision* objective (that is having low prediction errors), and a negative term being an estimator of the entropy of a *marginal* distribution, *before* action encoding.

For instance, on the decoding side, $I(O_t; \theta_t|h_{t-1})$ can be rewritten as $\mathbb{E}_{\tilde{\theta}_t\sim q(\theta|a_t,h_t)}\mathrm{KL}(p(O_t|\tilde{\theta}_t, h_{t-1})||p(O_t|h_{t-1}))$, that is the Kullback-Leibler divergence between the outcome distribution that is conditioned on the current action, and that of the marginal distribution. The action providing the outcomes that are *the most distant* from the marginal distribution is thus supposed to be selected more often. This "maximal decoding capacity", which is reminiscent of the "Information bottleneck" principle (Tishby et al., 2000), conducts in fact to an equilibrium state in which both action-outcome distributions should remain at equal distance from the marginal distribution, favoring alternate action selection in the long run.

## 3 Results

Several experiments were conducted in the lab in order to specifically assess sampling preference in undirected tasks in which subjects are asked to improve their understanding of their action/outcome causal relationships. We take here as an example a task called the "volleyball" task (Basanisi, 2021), in which 14 participants act as trainers tasked with hiring players for different teams. They can simulate the outcome of 40 matches with or without a particular player. They have to decide before each match whether to include the player or not ("PLAY" or "NOT PLAY"). Then, they observe the match outcome, and repeat the process until all 40 trials are completed. They are finally asked to assess the player causal effect on the team's success rate. Importantly here, and contrarily to the classic "bandit" setup, they are not asked to maximize success, but instead select which condition to sample in order to form a clear quantitative view of the player causal influence on the result. Different settings are considered in which $p_1$ (the Bernoulli probability to win when the player is present) is greater than $p_0$, and vice versa. A total of 15 different settings where considered, and for each of them the participant had to quantify how better or worst was the selection of the player.

We are interested here in the action selection strategies developed by the subjects, and by the action selection *biases*. At odd with a bandit setup, a correct strategy here would be to choose the actions 'PLAY' and 'NOT PLAY' in equal proportion. This is mostly what is observed in humans, which exhibit mostly balanced but irregular action selection, resembling that of a Bernoulli draw, with some form of periodic alternation (see fig. 1D). We estimate here three variant of "information seeking" objective functions, i.e. the Learning Progress eq. (1), the Self encoding capacity – eq. (2)), and the "decoding capacity" (eq. (3)). As shown in fig. 1F, despite a difference in monotony, both action values show concurrently evolving trend, i.e. implement a dynamic equilibrium in which concurrent action values alternately take over in inverse proportion to the choice frequency. In detail, while the

learning progress decreases over time, the self-encoding capacity shows an increasing trend, while the decoding capacity seems to reach a plateau, approaching the "Channel capacity".

At a more subtle level however, beyond the apparent balance between "PLAY" and "NOT PLAY", significant action selection biases, depending on the task setting, can be identified in behavioral data. Of interest here, a bias toward entropic (action, outcome) relationships was observed in experiments. Indeed, the entropy difference $\Delta H = H(p_1) - H(p_0)$, where $H(p) = -\sum_{s \in S} p(s) \log(p(s))$, indicates whether $p_1$ is more (or less) irregular than $p_0$. Such a bias was shown significant on our group of subjects (fig. 1G (left)). We thus assessed the presence of this entropy bias on a set of 14 information seeking agents (with different initialization and seeds) resolving the task in the same 15 different settings than the subjects.Then, regrouping the task settings in "positive" vs "negative" entropy bias, we compared action selection biases at the group level, using a non-parametric 2-sample test (Mann-Whitney U-test) (fig. 1G (right)). Interestingly, only the agents following the DEC objective were found to reproduce this entropy bias, while no significance was shown under the LP objective, and even a reverse effect was observed for the SEC agents.

## 4 Discussion

These results finally confirm our assumptions, suggesting that human subjects may develop information-guided action selection strategies, by combining principles of Bayesian estimation (Knill and Pouget, 2004) and action read-out information maximization (Klyubin et al., 2005), providing an interesting insight into the likely role of action decoding in information-seeking behavior. Indeed, while it is still premature to draw firm conclusions, the mathematical properties, behavior, and selection biases of the action selection driven by the action-decoding capacity of the generative model appear to be the most likely explanation for the behavioral data. The information-seeking action-value objectives suggested here are currently being investigated in ongoing experiments, where they serve as regressors in a model-based analysis of electrophysiological signals. Our analysis should also be extended to the more general case of curiosity-driven reinforcement learning, by serving as a foundation for exploration strategies based on the entropy of the *marginal* (i.e. action-agnostic) distribution of observations (Lee et al., 2019; Daucé, 2022). Finally, the use of generative models for encoding/decoding the dynamics of the controlled system, as suggested by (Janner et al., 2022), would allow for the combination of pursuing an external objective with efficient sampling of environmental data.

#### Acknowledgments

We thank Ruggero Basanisi who contributed to collect the behavioral data, and the reviewers for their careful reading and helpful suggestions, which have significantly enriched this work.

## References

Achiam, J. and Sastry, S. (2017). Surprise-based intrinsic motivation for deep reinforcement learning. *arXiv preprint arXiv:1703.01732*.

Angela, J. Y. and Dayan, P. (2005). Uncertainty, neuromodulation, and attention. *Neuron*, 46(4):681–692.

Baldassarre, G. (2011). What are intrinsic motivations? a biological perspective. In *2011 IEEE international conference on development and learning (ICDL)*, volume 2, pages 1–8. IEEE.

Barto, A., Mirolli, M., and Baldassarre, G. (2013). Novelty or surprise? *Frontiers in psychology*, 4:907.

Basanisi, R. (2021). *Neurophysiological and computational bases of goal-directed behavior*. PhD thesis, Aix-Marseille.

Daucé, E. (2022). Concurrent credit assignment for data-efficient reinforcement learning. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Doya, K. (2007). *Bayesian brain: Probabilistic approaches to neural coding*. MIT press.

Fiser, J., Berkes, P., Orbán, G., and Lengyel, M. (2010). Statistically optimal perception and learning: from behavior to neural representations. *Trends in cognitive sciences*, 14(3):119–130.

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nature reviews neuroscience*, 11(2):127–138.

Friston, K., Thornton, C., and Clark, A. (2012). Free-energy minimization and the dark-room problem. *Frontiers in psychology*, 3:130.

Geisler, W. S. (1989). Sequential ideal-observer analysis of visual discriminations. *Psychological review*, 96(2):267.

Gibson, J. J. (1979). *The Ecological Approach to Visual Perception: Classic Edition*.

Gläscher, J., Daw, N., Dayan, P., and O'Doherty, J. P. (2010). States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron*, 66(4):585–595.

Gottlieb, J., Oudeyer, P.-Y., Lopes, M., and Baranes, A. (2013). Information-seeking, curiosity, and attention: computational and neural mechanisms. *Trends in cognitive sciences*, 17(11):585–593.

Houthooft, R., Chen, X., Duan, Y., Schulman, J., De Turck, F., and Abbeel, P. (2016). Vime: Variational information maximizing exploration. *arXiv preprint arXiv:1605.09674*.

Itti, L. and Baldi, P. (2009). Bayesian surprise attracts human attention. *Vision Research*, 49(10):1295–1306.

Janner, M., Du, Y., Tenenbaum, J. B., and Levine, S. (2022). Planning with diffusion for flexible behavior synthesis. *arXiv preprint arXiv:2205.09991*.

Kalman, R. (1960). A new approach to linear filtering and prediction problems. *J. Basic Eng*, 82(1):35–45.

Klyubin, A. S., Polani, D., and Nehaniv, C. L. (2005). Empowerment: A universal agent-centric measure of control. In *2005 ieee congress on evolutionary computation*, volume 1, pages 128–135. IEEE.

Knill, D. C. and Pouget, A. (2004). The bayesian brain: the role of uncertainty in neural coding and computation. *TRENDS in Neurosciences*, 27(12):712–719.

Kobayashi, K. and Kable, J. W. (2024). Neural mechanisms of information seeking. *Neuron*.

L Griffiths, T., Kemp, C., and B Tenenbaum, J. (2008). Bayesian models of cognition.

Lee, L., Eysenbach, B., Parisotto, E., Xing, E., Levine, S., and Salakhutdinov, R. (2019). Efficient exploration via state marginal matching. *arXiv preprint arXiv:1906.05274*.

Liakoni, V., Lehmann, M. P., Modirshanechi, A., Brea, J., Lutti, A., Gerstner, W., and Preuschoff, K. (2022). Brain signals of a surprise-actor-critic model: Evidence for multiple learning modules in human decision making. *NeuroImage*, 246:118780.

Little, D. Y. and Sommer, F. T. (2013). Learning and exploration in action-perception loops. *Frontiers in neural circuits*, 7:37.

Mazzaglia, P., Catal, O., Verbelen, T., and Dhoedt, B. (2022). Curiosity-driven exploration via latent bayesian surprise. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 7752–7760.

Oaksford, M. and Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychological review*, 101(4):608.

Oudeyer, P.-Y., Kaplan, F., and Hafner, V. V. (2007). Intrinsic motivation systems for autonomous mental development. *IEEE transactions on evolutionary computation*, 11(2):265–286.

Pathak, D., Agrawal, P., Efros, A. A., and Darrell, T. (2017). Curiosity-driven exploration by self-supervised prediction. In *International conference on machine learning*, pages 2778–2787. PMLR.

Rao, R. P. and Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature neuroscience*, 2(1):79–87.

Schmidhuber, J. (2010). Formal theory of creativity, fun, and intrinsic motivation (1990–2010). *IEEE transactions on autonomous mental development*, 2(3):230–247.

Thompson, W. R. (1933). On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3-4):285–294.

Tishby, N., Pereira, F. C., and Bialek, W. (2000). The information bottleneck method. *arXiv preprint physics/0004057*.

Todorov, E. (2008). General duality between optimal control and estimation. In *2008 47th IEEE conference on decision and control*, pages 4286–4292. IEEE.

Varela, F. J., Thompson, E., and Rosch, E. (1991). *The Embodied Mind: Cognitive Science and Human Experience*. MIT Press.

Von Helmholtz, H. (1867). *Handbuch der physiologischen Optik*, volume 9. Voss.

Xu, H. A., Modirshanechi, A., Lehmann, M. P., Gerstner, W., and Herzog, M. H. (2021). Novelty is not surprise: Human exploratory and adaptive behavior in sequential decision-making. *PLOS Computational Biology*, 17(6):e1009070.