Irreversibility in belief dynamics: Unraveling the link to cognitive effort

Michele Vodret*
Université Paris-Saclay, CentraleSupélec,
Laboratoire de Mathématiques et Informatique pour la Complexité et les Systèmes, 91192 Gif-sur-Yvette, France

(Received 24 October 2023; accepted 25 June 2024; published 24 July 2024)

The relationship between time irreversibility in neuronal dynamics and cognitive effort is a subject of growing interest in the scientific literature. Although correlations between proxies of both concepts have been experimentally observed, the underlying precise linkage between them remains elusive. Here we investigate the case of learning in decision-making tasks; we do so by introducing a thermodynamically grounded metric—inspired by Landauer's principle—which connects time-irreversible information processing to energy consumption. Equipped with this metric, we investigate the role of macroscopic time-reversal symmetry breaking in belief dynamics for the case of an agent with finite sensitivity while performing a static two-armed bandit task—a standard setup in cognitive neuroscience. To gain insights into the belief dynamics, we analogize it to the dynamics of an active particle subject to state-dependent noise and living in a two-dimensional space. This mapping allows an analytical description of learning-induced biases. We deeply explore the case of Q-learning with forgetting the nonchosen option. In this case, learning-induced risk aversion is formally equivalent to standard thermophoresis, i.e., the net motion towards low-temperature regions. Finally, we quantify the irreversibility of belief dynamics in the steady state for different bandit configurations, sensitivity levels, and exploitative behavior. We found a strong correlation in high-sensitivity learning between heightened irreversibility in belief dynamics and improved decision-making outcomes. Notably, as the task's difficulty increases, a greater degree of irreversibility in belief dynamics becomes necessary for having superior performances; this explicitly unravels a plausible connection between time irreversibility and cognitive effort. In conclusion, our investigation reveals that irreversibility in belief dynamics bridges out-of-equilibrium statistical physics concepts and cognitive neuroscience. In decision-making contexts, this perspective offers insights into the notion of cognitive effort, suggesting a potential mechanism driving the evolution of living systems toward out-of-equilibrium structures.

DOI: 10.1103/PhysRevE.110.014304

I. INTRODUCTION

Decision-making is a fundamental cognitive process [1,2], ubiquitous across the spectrum of living systems since they necessitate a delicate balance between exploring new opportunities and exploiting existing knowledge to thrive. Importantly, exploitative behaviors might give rise to time irreversibility. For instance, an irreversible action significantly limits future choices for an extended period [3].

A compelling correlation between established proxies of cognitive effort and time irreversibility in fMRI and MEG human-brain data across various tasks and conditions [4–9] has been discovered. In particular, for the first time, in Ref. [5], researchers estimated model-dependent irreversibility in fMRI brain data, hinting at it as a signature of consciousness states. However, the rationale behind their model-dependent estimates remained unexplained, prompting our discussion. Our work aims to show that starting from well-known learning models, it is possible to construct micro-founded irreversibility metrics in decision-making scenarios to explain the observed time irreversibility in fMRI data.

We delve into the realm of time irreversibility in belief dynamics [10] inspired by recent advancements in the neural underpinnings of. decision-making [11], particularly the brain's retention of option-specific values [12,13]. We aim to elucidate the intrinsic computational cost of decision-making systems, which has garnered increasing attention from the out-of-equilibrium physics community [14].

To distinguish time irreversibility in action dynamics and the one in belief dynamics, consider a scenario where an individual allocates limited resources between two options, A and B. Unbeknownst to the individual, both options offer unknown but statistically equivalent rewards. An initial preference for A over B paves the way for exploitation. Depending on the exploitation intensity, belief dynamics might demonstrate a cycle of self-fulfilling prophecies: A bias towards option A increases resource allocation, resulting in higher average rewards and reinforcing the initial bias. This cycle persists until negative. fluctuations in the favored option shift preference to the other. Over time, though belief dynamics are time irreversible due to resource-limited exploitation, the resultant resource allocation and acquired rewards could display time-reversible dynamics. We will formalize this observation using a stylized decision-making model.

The aforementioned self-fulfilling prophecy mechanism plays a crucial role in various social contexts, encompassing financial markets [15–17] and economics [18,19], information dissemination in social media [20–23], the dynamics of politicians and voters in election polls [24,25], up to war engagements scenarios [26]. This highlights the importance

*Contact author: mvodret@gmail.com

of studying single-individual belief dynamics to understand how collective behaviors emerge.

We take model-free reinforcement learning (RL) [27] as the framework for our case study. This approach allows us to capture how each option's subjective values, or beliefs, are independently assessed and adapted to novel opportunities without constructing an environmental model to optimize actions toward specific goals. In model-free RL, algorithms directly determine subjective value functions through environmental interactions. One prominent algorithm in model-free RL is Q-learning, which, in its fundamental form, updates the subjective value of available options based on prediction errors. This framework finds applications [28–31] in neuroscience's dopamine equals reward prediction error (DA=RPE) hypothesis. According to this hypothesis, reward expectations are stored in the corticobasal ganglia synapses and are updated based on reward prediction errors through synaptic plasticity induced by released dopamine. Notably, this framework has become the standard also in cognitive neuroscience: it is used there for modeling cognitive biases, such as positivity or confirmation bias [32,33].

We investigate the influence of time irreversibility stemming from Q-learning dynamics in belief space in a two-armed bandit problem. To study macroscopic time-reversal symmetry breaking, we derive a coarse-grained Q-learning model formally indistinguishable from standard active particle models, paving the way for applications of out-of-equilibrium statistical physics tools. The coarse-grained description of belief dynamics will also help draw analogies between phenomena observed in two distant research fields, providing analytical descriptions of phenomena previously discovered numerically in theoretical cognitive neuroscience. Specifically, we can provide a generic way to link learning-induced biases to thermophoretic phenomena.

Finally, we link time irreversibility in belief dynamics to a thermodynamically sound concept of cognitive effort via Landauer's principle [34–36]: time-irreversible information processing generates heat. We conclude the paper with a numerical investigation of irreversibility in different decision-making contexts, suggesting a close link with the concept of cognitive effort.

This study offers three primary takeaways: (i) A formal connection between emerging risk-aversion and thermophoresis—the tendency of solute particles to migrate towards cooler regions—in forgetting Q-learning; (ii) a connection between time irreversibility of intertwined belief dynamics and dissipated work; and (iii) we discern that, for difficult tasks and high enough sensitivity, intermediate exploitative behavior aligns with peak irreversibility in belief dynamics and a beneficial balance between exploration and exploitation.

The following sections are structured as follows for the reader's ease: Section II first introduces two instances of Q-learning applied to a two-arm bandit task; from these, we construct effective models that focus on the slow dynamics relevant to macroscopic time-reversal symmetry breaking. We conclude this section by distinguishing different dynamical regimes related to exploitation levels and bandit configurations. Section III explains the notion of out-of-equilibrium steady state and introduces the building blocks of the irreversibility metric, i.e., probability currents in belief space. First, we discuss these objects using a spatially coarse-grained description. Then, we introduce the valuable toolkit for investigating the coarse-grained Q-learning models in continuous time and space. The analytical tractability of the latter description allows us to link learning-induced risk aversion to thermophoresis. Section IV presents the irreversibility metric and delves into the association with cognitive effort. This section ends by sharing numerical findings related to irreversibility in different tasks. Section V concludes with a discussion of the results and suggests potential avenues for future research.

II. Q-LEARNING MODELS

First, we introduce two instances of Q-learning applied to a two-armed bandit game scenario. Then we discuss how to derive an effective description to quantify the degree of macroscopic time-reversal symmetry breaking. We conclude the discussion by highlighting qualitative differences between macroscopic belief dynamics for various exploitation levels and bandit configurations.

A. Microscopic description

Consider a two-armed bandit game scenario: At every (discrete) time step t, a decision maker invests a single unit of endowment in one between two "arms" of a slot machine, denoted as A and B. $a_t \in \{1, 0\}$ signifies whether or not the subject invested at time step t on bandit A, while $1 - a_t$ does so for bandit B. Accordingly, the rewards yielded by the arms at each time step are $R_t^A a_t$ and $R_t^B(1 - a_t)$, respectively. Both R_t^A and R_t^B are drawn by time-independent Bernoulli distributions reflecting a stable environment. We indicate the mean and standard deviation of R_t^A respectively as $\langle R^A \rangle$ and σ^A, with analogous notation for R_t^B; these pieces of information are unknown to the decision maker.

To model the possibility of exploring suboptimal options or exploiting existing knowledge, a_t is usually [33] assumed to be a Bernoulli variable whose distribution solely depends on the difference of beliefs at the current time step t, denoted respectively as $\hat{R}_t^A$ and $\hat{R}_t^B$. A convenient parametrization is given by

$$\mathbf{P}\big(a_t = 1 | \hat{R}_t^A - \hat{R}_t^B\big) = \frac{1 + \tanh\left[\Gamma\big(\hat{R}_t^A - \hat{R}_t^B\big)\right]}{2}, \qquad (1)$$

where $\Gamma \geqslant 0$ is the exploitation parameter: It dictates how belief disparities affect investments. A positive exploitation parameter Γ value enhances the inclination to invest in the arm perceived as more lucrative. The standard soft-max formulation can be recovered by setting $\Gamma \to \Gamma/2$.

Let us present the Q-learning model [37,38] in a generic way that encompasses two special cases:

$$\hat{R}_{t+1}^A - \hat{R}_t^A = \beta a_t \big(R_t^A - \hat{R}_t^A\big) - \beta^f (1 - a_t)\hat{R}_t^A, \qquad (2a)$$

$$\hat{R}_{t+1}^B - \hat{R}_t^B = \beta (1 - a_t)\big(R_t^B - \hat{R}_t^B\big) - \beta^f a_t \hat{R}_t^B. \qquad (2b)$$

The learning rates $\beta > 0$ and $\beta^f \geqslant 0$ manage two different facets: β regulates the agent's sensitivity to new data via the prediction error while β^f represents the agent's propensity to forget the value associated with the nonchosen option. We

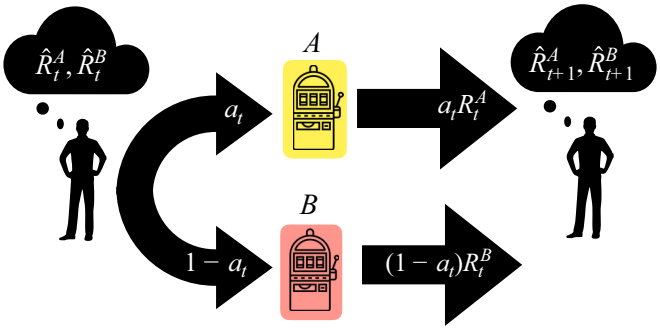

FIG. 1. Graphical representation of the model. The investment decision precedes the observation of the actual outcome.

will say that lower β corresponds to a higher sensitivity of the decision maker.

A graphical description of the dynamics between time step t and $t+1$ is given in Fig. 1: The action taken by the agent at time step t is a function only of the current beliefs. Then, based on the obtained rewards, the agent updates his or her beliefs with Eq. (2). Note that the updated beliefs depend only on previous ones, i.e., the beliefs dynamics is Markovian. In the following, we specialize the model above into two special cases.

1. Standard Q-learning

This learning rule requires $\beta^f = 0$, i.e., the decision-maker remembers the value of the nonchosen option virtually forever. The belief dynamics for $\Gamma = 0$ is at long times in the neighborhood of their respective averages (see top left plot in Fig. 2, where the case of symmetric arms is investigated). For large exploitation parameter Γ and at long times the behavior is different: One of the two beliefs, say, B, is pushed to the lower boundary of the support of the belief distribution and the other, say, A, is in the neighborhood of $\langle \hat{R}^A \rangle$, implying that $a_t = 1$ for a long time (see the bottom left plot in Fig. 2).

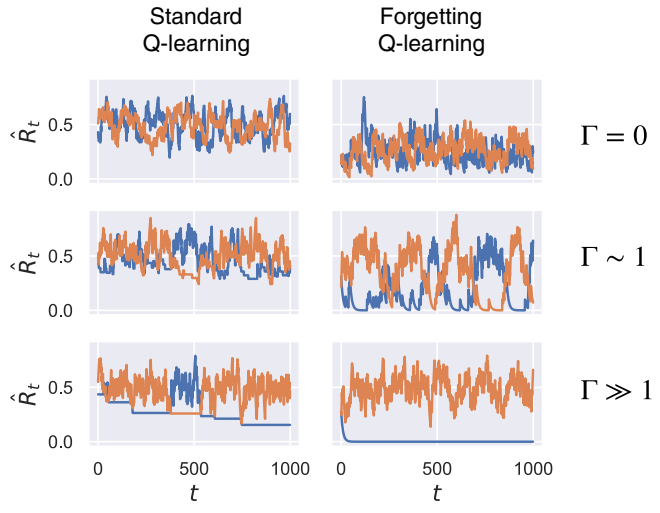

FIG. 2. Example of belief dynamics for standard ($\beta^f = 0$) and forgetting ($\beta^f = \beta$) Q-learning models, respectively given by left and right plots. $\beta = 0.1$. The two arms are symmetric and $\langle R^A \rangle = \langle R^B \rangle = 0.5$. The exploitation level is increasing from top to bottom. Left plots are obtained with $\Gamma = 0$, $\Gamma = 5$, and $\Gamma = 40$, while right plots are obtained with $\Gamma = 0$, $\Gamma = 2.5$, and $\Gamma = 7.5$. Initial conditions are $\hat{R}_0^A = \hat{R}_0^B = 0.5$ for left plots and $\hat{R}_0^A = \hat{R}_0^B = 0.25$ for right plots.

2. Forgetting Q-learning

Forgetting Q-learning requires $\beta^f > 0$. In the following, we refer to the well-known case with $\beta^f = \beta$, for simplicity; in this case, the linear term in the right-hand side of the equations above is time independent and equal to β for both arms, i.e.,

$$\hat{R}_{t+1}^A - \hat{R}_t^A = -\beta \hat{R}_t^A + \beta a_t R_t^A, \tag{3a}$$

$$\hat{R}_{t+1}^B - \hat{R}_t^B = -\beta \hat{R}_t^B + \beta a_t R_t^B. \tag{3b}$$

Examples of belief dynamics are depicted on right plots in Fig. 2 which show a certain degree of similarity to the case without forgetting. We highlight two differences. First, for $\Gamma = 0$ the belief dynamics oscillates around $\langle R^A \rangle/2 = \langle R^B \rangle/2$. This is because forgetting modifies the structure of the subjective value, which is no longer given by the moving average of prediction errors. Second, for large Γs, one of the two beliefs is pushed to zero.

There are two reasons behind the choice of a zero reference for the forgetting term: The first one is that forgetting has been suggested to be implemented as a decay of synaptic strengths storing learned values in the context of the DA=RPE hypothesis, and the second is that from a normative point of view forgetting usually wants to model a notion of fast goal-reaching. The more the reference point is smaller than the average return of the bandit's arms, the faster the decision-maker will decide to invest consistently in one of the two arms. From a purely qualitative point of view, one can see that the two dynamics for large Γs are consistent; this would not be the case if the forgetting takes the value 0.5 as a reference, for instance.

In conclusion, by shifting the belief related to the nonchosen option towards zero, the forgetting Q-learning model effectively reduces the time needed to reach stationary dynamics concerning standard Q-learning. This is particularly evident in the central plots of Fig. 2, where exploitation allows us to distinguish more between the two options in the case of forgetting Q-learning with respect to the case of standard Q-learning.

Before proceeding, we would like the reader to recognize that the mathematical structure of the forgetting Q-learning model is simpler than that of the standard Q-learning model. The reason is that the deterministic belief-dependent term in the former dynamics resembles an elastic spring, whereas, in the standard model, the spring is both belief and arm dependent. While this complication does not qualitatively alter the main results of the present paper, I decided to thoroughly analyze the simpler forgetting Q-learning model quantitatively and leave the detailed analysis of more complex learning rules for future investigations.

B. Effective description at low frequencies

Now we derive an effective forgetting Q-learning model to understand the belief dynamics behavior at a macroscopic

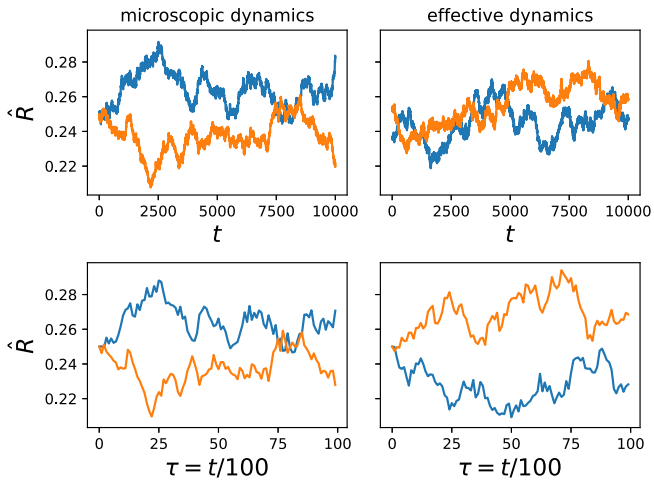

FIG. 3. Coarse-graining procedure on forgetting Q-learning and checks. Top left: Belief dynamics for $\beta = 0.001$ and $\Gamma = 1.5$. Bottom left: Belief dynamics sampled every $L = 100$ steps, given by Eq. (4). Bottom right: Coarse-grained dynamics at macroscopic frequency, given by Eq. (7). Coarse-grained belief dynamics at microscopic frequency, given by Eq. (8).

level [4]; a completely analogous treatment can be made for the more general model given by Eq. (2). This approach aligns with the study of complex systems, where understanding how microscopic patterns give rise to macroscopic behavior is a central objective.

Consider the setup where the two arms A and B are very similar; in this situation, it is difficult to distinguish them and to play consistently on the most lucrative one. To do that, one should average over many realizations, meaning that the sensitivity to new information should be high, i.e., $\beta \ll 1$. In this scenario, beliefs are slowly varying, and the description provided by Eq. (2) [as well as Eq. (3)] can be simplified by looking at the dynamics at a coarser timescale.

The starting point to obtain such a description is the microscopic dynamics introduced in the previous section; in Fig. 3, in the top left plot, we show the dynamics for $\beta = 0.001$ and $\Gamma = 1.5$.

Then we define the following lagged temporal differences:

$$\Delta_L \hat{R}^A_t = \hat{R}^A_{t+L} - \hat{R}^A_t, \tag{4a}$$

$$\Delta_L \hat{R}^B_t = \hat{R}^B_{t+L} - \hat{R}^B_t, \tag{4b}$$

where L is greater than one but smaller than β^{-1}; we choose $L = 100$, and we plot the resulting time series in the bottom left plot of Fig. 3, which displays less variability at high frequency.

By injecting Eq. (3) in the equations above and rearranging them, one obtains

$$\Delta_L \hat{R}^A_t = -\beta \sum_{l=0}^{L-1} \hat{R}^A_{t+l} + \beta \sum_{l=0}^{L-1} a_{t+l} R^A_{t+l}, \tag{5a}$$

$$\Delta_L \hat{R}^B_t = -\beta \sum_{l=0}^{L-1} \hat{R}^B_{t+l} + \beta \sum_{l=0}^{L-1} (1 - a_{t+l}) R^B_{t+l}, \tag{5b}$$

where the first terms in the right-hand sides of both equations above can be approximated by L times the arithmetic mean of the beliefs over the window $[t, t+L)$, which can be approximated by $\hat{R}^A_t$ and $\hat{R}^B_t$, respectively. The second terms on the right-hand sides of the equations above can be simplified as follows. Let us focus on Eq. (5a). The product $a_t R^A_t$ is a time-dependent Bernoulli variable with mean $\langle a_t \rangle \langle R^A \rangle$ since a_t and R^A_t are independent. The summation of L Bernoulli variables can be approximated for large L by a Gaussian variable with mean and variance given by L times the associated mean and variance of the Bernoulli variable under the summation; therefore, this Gaussian random variable can be written as L times $\eta^A_t|_L$, where

$$\langle \eta^A_t \rangle = \langle a_t \rangle \langle R^A \rangle, \tag{6a}$$

$$\left\langle \left(\eta^A_t - \langle \eta^A_t \rangle\right)^2 \right\rangle\Big|_L = \langle a_t \rangle \langle R^A \rangle (1 - \langle a_t \rangle \langle R^A \rangle)/L. \tag{6b}$$

Equation (5b) can be treated similarly, giving a Gaussian variable η^B_t, with mean and variances obtained from the ones above by the following modifications: $\langle a_t \rangle \to (1 - \langle a_t \rangle)$ and $\langle R^A \rangle \to \langle R^B \rangle$. After these modifications are taken into account, we obtain the effective version of Eq. (5), given by

$$\Delta_L \hat{R}^A_t = -\beta L \hat{R}^A_t + \beta L \eta^A_t\big|_L, \tag{7a}$$

$$\Delta_L \hat{R}^B_t = -\beta L \hat{R}^B_t + \beta L \eta^B_t\big|_L. \tag{7b}$$

The resulting time series are shown in the bottom right plot of Fig. 3.

Let us recap what we have obtained. We showed that the coarse-grained description of the belief dynamics deviates from the original one [Eq. (3)] in two respects. The first modification is that, in the coarse-grained description, rewards are Gaussian variables instead of Bernoulli variables. The second modification is that in the coarse-grained model, the noise (the second term in the right-hand side of equations above) is generally state dependent.

Interestingly, after these modifications are considered, the slow learning dynamics makes predictions that align with recent empirical findings. In particular, in passive learning scenarios ($\Gamma = 0$), where the investment is split equally on both arms ($\langle a \rangle \sim 1/2$), the beliefs dynamics given by Eq. (7) results in two independent first-order autoregressive (AR) processes with additive Gaussian noises, which are time reversible in the long run. Therefore, the coarse-grained description in passive learning scenarios aligns with the time reversibility observed in Alzheimer's Disease-related fMRI brain dynamics [39]. Note that this is not the case in the microscopic version of the model because Bernoulli rewards and actions break the time-reversal symmetry of the related autoregressive process [40].

1. Langevin equation

Finally, to return to the original microscopic timescale, one can set $L \to 1$ in Eq. (7). The resulting equations can be formally rewritten as

$$\frac{d\hat{R}^A_t}{dt} = -\beta \hat{R}^A_t + \beta \eta^A_t, \tag{8a}$$

$$\frac{d\hat{R}^B_t}{dt} = -\beta \hat{R}^B_t + \beta \eta^B_t, \tag{8b}$$

where η^A_t has mean and variance given by Eq. (6) with $L = 1$ and similarly for η^B_t.

To simulate the above systems of equations in a computer, one relies on the so-called Euler method: Introducing a scale h and pushing it to zero allows one to simulate effectively the dynamics in continuous time. Accordingly, the set of equations above is implemented for simulations hereafter in the following way:

$$\hat{R}^A_{t+h} - \hat{R}^A_t = -h\big(\beta \hat{R}^A_t + \beta \langle a_t\rangle \langle R^A\rangle\big) + \sqrt{h}\beta\sqrt{\langle a_t\rangle\langle R^A\rangle(1-\langle a_t\rangle\langle R^A\rangle)}\xi^A_t, \tag{9a}$$

$$\hat{R}^B_{t+h} - \hat{R}^B_t = -h\big[\beta \hat{R}^B_t + \beta \langle (1-\langle a_t\rangle)\rangle \langle R^B\rangle\big] + \sqrt{h}\beta\sqrt{\langle (1-\langle a_t\rangle)\rangle\langle R^B\rangle[1-(1-\langle a_t\rangle)\langle R^B\rangle]}\xi^B_t, \tag{9b}$$

where ξ^A_t and ξ^B_t are two independent Gaussian variables with zero mean and unit variance. $h = 0.1$ for the rest of the paper. For instance, the top right plot shows the effective dynamics related to the microscopic dynamics in the top left plot in Fig. 3.

2. *Simplified forgetting Q-learning model*

In what follows, we will be interested in understanding quantitatively whether there are net movements in the belief dynamics at the steady states. These are a signature of the time irreversibility of belief dynamics. To facilitate this task, we will consider a simplified forgetting Q-learning model, where we retain only one noise source, the one coming from the stochastic rewards. The microscopic model is therefore given by Eq. (3) with the substitution $a_t \to \langle a_t\rangle$: The role played by a_t is taken by $\langle a_t\rangle$ which can be considered as a deterministic variable given the difference $\hat{R}^A_{t,L} - \hat{R}^B_{t,L}$, instead of being the outcome of a Bernoulli distribution [see Eq. (1)]. With this modification, we let $a_{t,L}$ vary continuously, incorporating a notion of confidence [41] in the model.

Finally, after performing a time coarse-graining procedure as in the previous section, the Euler method prescribes the following discrete-time dynamics:

$$\hat{R}^A_{t+h} - \hat{R}^A_t = -h\big(\beta \hat{R}^A_t + \beta \langle a_t\rangle \langle R^A\rangle\big) + \sqrt{h}\beta\langle a_t\rangle\sqrt{\langle R^A\rangle(1-\langle R^A\rangle)}\xi^A_t, \tag{10a}$$

$$\hat{R}^B_{t+h} - \hat{R}^B_t = -h\big[\beta \hat{R}^B_t + \beta \langle (1-\langle a_t\rangle)\rangle \langle R^B\rangle\big] + \sqrt{h}\beta(1-\langle a_t\rangle)\sqrt{\langle R^B\rangle(1-\langle R^B\rangle)}\xi^B_t, \tag{10b}$$

which deviates from Eq. (9) because of the variance of the noise term; in particular, in the simplified forgetting Q-learning model, as expected, the noise variance is always lower. This will become convenient for determining net movements at the steady states, i.e., irreversibility, in the belief dynamics. The reason is that the determination of net movements relies on a signal-to-noise ratio; extracting the signal related to net movements of beliefs in the steady state will be easier if we reduce the noise. Therefore, changes to the forgetting Q-learning model are only quantitative, but the behavior remains the same qualitatively.

To give empirical validity to the present model, experiments must be performed in which the decision maker can split the investment at each timestep between both arms instead of being restricted to invest solely in one of the two arms at each timestep.

C. Preliminary analysis

We summarize below the main properties of the coarse-grained Q-learning models for different exploitation levels and bandit configurations.

$\Gamma = 0$: The beliefs dynamics is passive because the agent's action is decoupled from his or her own beliefs. In particular, in the modified forgetting Q-learning model, the agent will always split the investment equally. Another way of formulating this concept is by saying that in the case of passive learning, there is no feedback between actions and beliefs. Therefore, the coarse-grained dynamics of the beliefs are completely decoupled, and they evolve in time as independent Ornstein-Uhlenbeck processes, implying that, as we stressed in Sec. II B, the coarse-grained belief dynamics is time reversible in the steady state [40].

$\Gamma \sim 1/\langle R^A\rangle, 1/\langle R^B\rangle$: Investments influence the belief dynamics; this is because the difference in beliefs determines them. Effectively, this is a state-dependent, i.e., multiplicative, noise. This region is the most interesting for us; let us mention here two reasons why: first, it is with a Γ in this region that the decision-maker will gain the most on average in realistic scenarios [32,33] in cases where $\langle R^A\rangle \sim \langle R^B\rangle$. Second, as we will show later, in this region, the belief dynamics is time irreversible even in the steady state.

$\Gamma \gg 1/\langle R^A\rangle, 1/\langle R^B\rangle$: In this situation, the decision maker plays initially in a consistent manner on one of the two arms, leading the belief of the nonchosen option to go to the lower boundary of the support of the belief distribution in the case of standard Q-learning and to zero in the case of (modified) forgetting Q-learning (see respectively the bottom left and right panel of Fig. 2). Let us analyze the modified forgetting Q-learning case in depth. Depending on the ratio between the fluctuations of the chosen belief and the distance of the belief from zero, the dynamics might or might not be stable. To be more precise, let us define the following noise-to-signal ratio:

$$\chi^A = \frac{\sigma^A_{\max}}{\mathrm{gap}^A_{\max}} = \sqrt{\frac{\beta(1-\langle R^A\rangle)}{\langle R^A\rangle}}, \tag{11}$$

where $\mathrm{gap}^A_{\max}$ is the maximum average distance between the beliefs and the final equality is justified for the modified forgetting model, for which $\sigma^A_{\max} = \sqrt{\beta R^A(1-R^A)}$ and

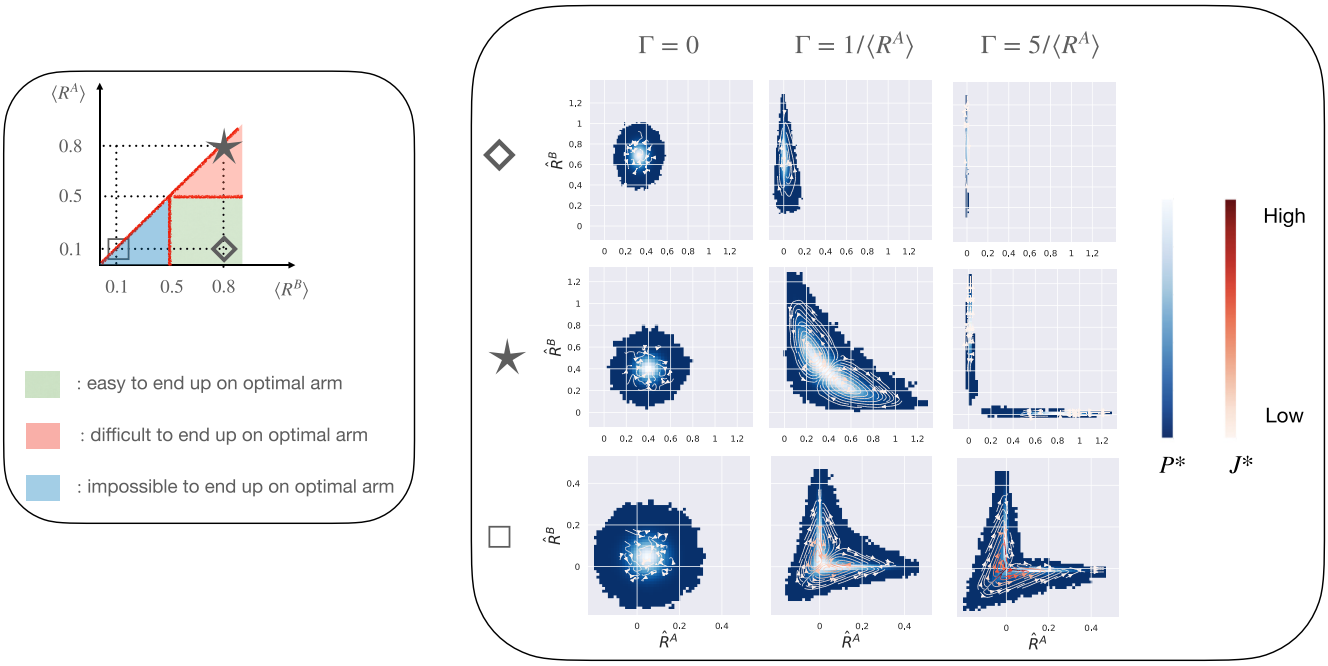

FIG. 4. Preliminary analysis of the modified forgetting Q-learning model. $\beta = 0.25$. Left: Sketch the belief dynamics' phase space when $\Gamma \gg 5/\langle R^A \rangle$. Right: Steady-state probability distribution on coarse-grained state space $P_*[\hat{R}]$, represented by the blue or white background color, and stationary probability currents among coarse-grained states J_* shown by red or white arrows. The scale of the color bars is purely qualitative. Symbols on the left refer to the combination of $\langle R^A \rangle$, $\langle R^B \rangle$ defined in the left plot.

$\text{gap}^A_{\max} = \langle R^A \rangle$. Let us consider three different situations, depending on the χ^A value and the analogous one for arm B. The discussion will be made by identifying a tipping point where $\chi^X = 1/3$, for which Gaussian fluctuations do not allow the belief to reach zero.

(i) Let us first consider the case $\chi^B \ll 1/3$ but $\chi^A > 1/3$. In the steady state, the decision maker will invest consistently in the optimal arm (χ^X is monotonously decreasing for $\langle R^X \rangle$). This is because the beliefs dynamics with $\hat{R}^B_t \gg \hat{R}^A_t$ is stable while the reverse condition is not.

(ii) The situation where $\chi^A, \chi^B \ll 1/3$ allows either conditions ($\hat{R}^A_t \gg \hat{R}^B_t$ and the reverse one) to be stable in the steady state, meaning that the decision-maker might be stuck virtually forever on the less lucrative arm.

In these first two cases, the noise can be treated effectively as an additive one in the steady state since the dynamics is stable. This means that the dynamics is time reversible (given by a single Ornstein-Uhlenbeck process, since the nonchosen belief is frozen at zero).

(iii) Finally, when $\chi^A, \chi^B > 1/3$, even if for a certain period one of the two arms is consistently chosen, the fluctuations will inevitably allow the belief to approach zero, disanchoring the—up to that point—nonchosen arm. In this case, the noise cannot be treated as additive even in the limit $\Gamma \to \infty$. Therefore, the belief dynamics will always be coupled if $\Gamma \neq 0$.

These considerations are resumed in the left panel of Fig. 4. There, we fix $\beta = 0.25$, and we let vary the ratios χ^A and χ^B by moving $\langle R^A \rangle$ and $\langle R^B \rangle$ (only the lower triangle of the phase space is shown, since the beliefs dynamics it's symmetric between $\hat{R}^A$ and $\hat{R}^B$). The threshold where $\chi^A, \chi^B = 1/3$ is used to divide the phase space into three regions, which correspond approximately to $\langle R^A \rangle, \langle R^B \rangle = 0.5$; note that if one lowers β, then the value of the corresponding threshold in $\langle R^A \rangle$ and $\langle R^B \rangle$ would diminish. In other words, increasing the sensitivity allows the decision-maker to distinguish two similar options. The region where $\chi^A, \chi^B > 1/3$ is given by the blue region, where the average rewards of both bandits are low, while the opposite red region refers to the case where $\chi^A, \chi^B < 1/3$ and corresponds to the case where the average rewards of both bandits are high. Finally, the "safe" region corresponds to the case where one average reward is so greater than the other that no fluctuation can destabilize the belief related to the most lucrative arm in the long run. Below, we will analyze in depth one example belonging to each of these regions, identified by symbols ($\star$, $\square$, $\diamond$) in the left plot of Fig. 4.

Alongside the visual inspection of the trajectories of the beliefs, an object worth analyzing is the associated probability density function (PDF) indicated as $P_t = P_t[\hat{R}_t]$. A spatially coarse-grained version of it is shown for the steady state of the system for different Γs in the right panel of Fig. 4 for the three bandit's configuration given by $\star$, $\square$ and $\diamond$ highlighted in the left plot; note that in Fig. 4 and in the following we will identify steady-state properties by the subscript $_*$. One observes the transition to bimodality of P_* as Γ increases in the $\star$ configuration; the bimodality is related to the emergence of trapping states. Most importantly for the remainder of the paper, for moderate Γ values, P_* spans the two-dimensional space maximally, while for large Γs the beliefs dynamics is mostly constrained onto a one-dimensional space. Note that the bimodality is not present as Γ increases for the $\diamond$ configuration, given that here the optimal belief is stable, and the agent can always recognize it and play consistently on it. On the other hand, the $\square$ configuration shows that even for large

Γ, the beliefs are always coupled (the visited belief space is always connected).

The case of Q-learning without forgetting is more mathematically subtle, given that there is no reference point in the dynamics (as the zero of the forgetting Q-learning model). In particular, this implies that the denominator in Eq. (11) is always of the order of the fluctuation themselves if $\langle R^A\rangle \sim \langle R^B\rangle$. Given the already rich behavior of the dynamics of forgetting Q-learning, we will not deal any longer with Q-learning without forgetting in this paper, leaving a detailed analysis of it for future research.

III. NONEQUILIBRIUM BELIEF DYNAMICS

This section introduces the concept of out-of-equilibrium dynamics in the steady state, together with key metrics that quantify the departure from equilibrium and, therefore, time-reversible dynamics. First, we tackle the case where we divide the continuous belief space into a finite-size grid because of its high interpretability. Then, we will introduce the techniques necessary to tackle the case with continuous state-space.

A. Coarse-grained space

To monitor net movements in the spatially coarse-grained picture of the model one can compute the transition matrix $\mathcal{T}[\hat{R}^i \to \hat{R}^j]$ from state $\hat{R}^i$ to state $\hat{R}^j$, where $i, j \in \{0, \ldots, N\}$, N^2 is the cardinality of the coarse-grained state space and $\mathcal{T}[\hat{R}^i \to \hat{R}^j]$ represents the probability of going to the coarse-grained state $\hat{R}^j$ starting from $\hat{R}^i$. From the transition matrix, one can define the associated probability current as

$$J_t[\hat{R}^i \to \hat{R}^j] = P_t[\hat{R}^i]\,\mathcal{T}[\hat{R}^i \to \hat{R}^j] - P_t[\hat{R}^j]\,\mathcal{T}[\hat{R}^j \to \hat{R}^i]. \tag{12}$$

Note that in doing this spatial coarse-graining, one loses information about the dynamics at smaller spatial scales.

A useful classification of the system's dynamical state in the steady state is contingent on the value of the probability current:

$J_* = 0$: In equilibrium steady states (ESS) [42] there are no probability currents. This indicates time-reversal symmetry (TRS), i.e., in these states, there is a complete absence of net movements in the system; this condition is known in the physics literature as detailed balance. In the dynamics of interest here, ESSs are observed in two distinct regimes of the exploitation parameter: $\Gamma = 0$ and Γ large for the regions where χ^A, $\chi^B \ll 1/3$ or $\chi^B \ll 1/3$ and $\chi^A > 1/3$, respectively given by the red and green regions in the left plot of Fig. 4.

$J_* \neq 0$: Nonequilibrium steady states (NESS) exhibit probability currents. In systems with a compact state space, these flows lead to net circulating movements in the belief space, a clear signature of time-reversal symmetry breaking (TRSB). This behavior is notably prevalent in the regime $\Gamma \sim 1/\langle R^A\rangle, 1/\langle R^B\rangle$ for the regions where χ^A, $\chi^B \ll 1/3$, and the one where $\chi^B \ll 1/3$ and $\chi^A > 1/3$ (red and green region in the left plot of Fig. 2), and for all $\Gamma \neq 0$ for the region where χ^A, $\chi^B > 1/3$ (blue region in the left plot of Fig. 4).

Probability currents among coarse-grained states are depicted on top of the plots on the right panel of Fig. 4. As can be visually appreciated, the NESSs (center-center, center-bottom, and right-bottom plots) are characterized by a structure of probability currents similar to dipole currents [43,44]; the rise and fall of self-fulfilling prophecies anticipated in section I is now evident from these plots. Let us now comment in detail on the case of forgetting Q-learning: A small initial bias towards arm A with respect to the equilibrium condition of passive learning leads to an increase in the value of $\hat{R}^A$ and a decrease of $\hat{R}^B$; eventually, $\hat{R}^A_t$ reaches the bottom right angle of the belief space and from there $\hat{R}^A$ will diminish; when $\hat{R}^A \sim \hat{R}^B$ two things can happen: The initial bias is restored, and the cycle repeats itself, or there is an inversion such that $\hat{R}^B > \hat{R}^A$. In this latter case, $\hat{R}_t$ will follow the cycle in the upper triangle of the plot, completely analogous to the cycle in the lower triangle of the plot.

B. Continuous space

In the previous section, we estimated probability currents between coarse-grained states. It is well known that estimating the probability currents J on a spatially coarse-grained version of the system's state space provides only lower bound estimates on these [45]. To properly estimate probability density currents, and therefore—as we will see in Sec. IV—time irreversibility, in a continuous-state system, a useful framework is given by Fokker-Planck equations; the reason for this is related to a technical simplification: the Fokker-Planck equation associated with a stochastic process is the deterministic dynamic equation for its PDF.

1. Fokker Planck equation

The coupled Langevin equations (8) articulate how beliefs evolve in time due to drifts—or systematic tendencies—and diffusions, which refer to random fluctuation; the former is represented in our system by the forgetting term and the average noise-related contributions, while the latter relates to the deviation from the mean of the noise term in our model. The starting point to discuss Fokker Planck equation is the system of Langevin equation given by Eq. (8), which can be rewritten in a more compact form given by

$$\frac{d\hat{R}_t}{dt} = \mathcal{F}_t + \xi_t, \quad \text{with} \quad \langle \xi_t^T \xi_{t'}\rangle = 2\mathcal{D}_t \delta_{t-t'}, \tag{13}$$

where $\mathcal{F}_t$ is the drift vector and $\mathcal{D}_t$ is the diffusion matrix, both of which depend on the current belief $\hat{R}_t$. These mathematical objects are defined for the case of modified forgetting Q-learning in Appendix B.

The Fokker Planck equation describes the deterministic dynamics of $P_t = P_t(\hat{R})$ as

$$\frac{\partial P_t}{\partial t} = -\nabla \cdot J_t, \tag{14}$$

where the probability density current J_t is given by

$$J_t = \mathcal{F}_t P_t - \nabla(\mathcal{D}_t P_t). \tag{15}$$

J_t represents here the local net flow of probability in the beliefs space $\hat{R}$; note that it is the continuous—in time and space—analog of the probability current introduced in Eq. (12).

The detailed balance condition $J = 0$ in the Fokker-Planck framework can be rewritten, leading to a condition that can be easily checked analytically and deeper insights into the fundamental causes of TRSB in NESSs. It can be rewritten using the equation above:

$$\nabla \times \mathscr{F} = 0, \tag{16}$$

where the so-called thermodynamic force $\mathscr{F}$ is given by

$$\mathscr{F} = \mathcal{D}^{-1}(\mathcal{F} - \nabla \cdot \mathcal{D}). \tag{17}$$

From direct computation of Eq. (16), one sees that the detailed balance is broken as soon as $\Gamma \neq 0$.

Therefore, the curl of the thermodynamic force plays the role of the electric density current in magnetostatic, where it induces the magnetic field. Here $\nabla \times \mathscr{F}$ is the source of the NESS [46,47]. A succinct way to rephrase the above intuition is that NESSs are related to a topological symmetry breaking.

Since we cannot easily construct the steady-state distribution due to the absence of detailed balance for $\Gamma \neq 0$, determining $P_*(\hat{R})$ for generic Γ values remains a challenge. In the following, we show how interesting insights can still be garnered from the steady-state PDF of the beliefs difference $\hat{R}^A_t - \hat{R}^B_t$.

2. *Emergent risk aversion as thermophoresis*

At first glance, using point estimates in the update equation for the beliefs given by Eq. (2) appear overly simplistic, especially when considering its lack of direct reference to well-documented human behavioral tendencies, such as risk aversion. If everything else is equal, then risk-averse individuals demonstrate a preference for less variable options, reducing the associated risk. However, early numerical analysis on related models revealed that using point estimates in the update equation of the beliefs does not neglect these tendencies; instead, risk aversion is an emerging property of the beliefs dynamics [48] analyzed in this paper.

Here we show that the Fokker-Planck framework allows us to derive this result explicitly for the case of forgetting Q-learning. This is possible because the belief dependency in the noise term is solely on the difference $\hat{R}^A_t - \hat{R}^B_t$ [see Eq. (1)], and the spring is elastic [see Eq. (3) and subsequent related ones]. To show this explicitly, let us introduce the coordinate transformation $(\hat{R}^A_t, \hat{R}^B_t) \to (\hat{R}^A_t + \hat{R}^B_t, \hat{R}^A_t - \hat{R}^B_t)$ and similarly for the rewards (R^A_t, R^B_t). Of particular interest is the observation that the update equation for $\delta\hat{R}_t = \hat{R}^A_t - \hat{R}^B_t$ in the case of forgetting Q-learning remains independent of the coordinate $\hat{R}^A_t + \hat{R}^B_t$, thus implying that the detailed balance for $\delta\hat{R}$ holds; note that this is not true in the low-frequency regime of the standard model because of the extra-non-linear couplings induced by the belief- and state-dependent spring. The TRS of $\delta\hat{R}_t$ in the steady state allows for an analysis of the associated Fokker-Planck equation. In particular, the thermodynamic force $\tilde{\mathscr{F}} = \tilde{\mathscr{F}}[\delta\hat{R}]$ is given by

$$\tilde{\mathscr{F}} \sim \frac{1}{\beta} \frac{-2\delta\hat{R} + \langle \delta R \rangle + \langle R \rangle \tanh[\Gamma \delta\hat{R}]}{\sigma_A^2 a^2 + \sigma_B^2 (1-a)^2}, \tag{18}$$

where for conciseness, we have not reported the second term $(\nabla \cdot \mathcal{D}/\mathcal{D})$ because it is of order zero in β and therefore subleading in the regime $\beta \ll 1$. From the equation above it is clear that for intermediate Γs, the multiplicative noise implies risk aversion: In fact, the denominator is smaller in the case of $\delta\hat{R} > 0$ for $\sigma_A^2 < \sigma_B^2$; this implies a stronger thermodynamic force towards region with $\delta\hat{R} > 0$, i.e., to belief states where the agent invests mostly on the less variable arm A.

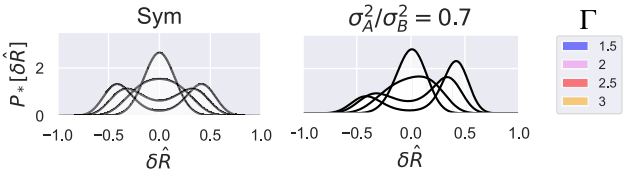

FIG. 5. Comparison of stationary PDF in modified forgetting Q-learning computed from the analytical prediction (black solid lines) with the one obtained from numerical simulations (colored histograms). Left: Symmetric bandits: $\langle R^A \rangle = \langle R^B \rangle = 0.5$. Right: Bandits with symmetric rewards and asymmetric variances: $\sigma_A^2 = \sigma_B^2 * 0.7$. The total simulation time is 10^4, and we retain only the second half of the trajectories.

Interestingly, for the present discussion, the form of detailed balance given by Eq. (16) is known as potential condition [42]. The reason is apparent for the dynamics of $\hat{R}^A - \hat{R}^B_t$ we are discussing. In fact, one has $P_*[\delta\hat{R}] \propto \exp[\int \tilde{\mathscr{F}}] = \exp[-\tilde{\Phi}]$, i.e., since $\tilde{\mathscr{F}}$ is curl-free then the thermodynamic potential $\tilde{\Phi}$ can be constructed by a simple integration of the thermodynamic force. $P_*[\delta\hat{R}]$ obtained from simulations and the one predicted from the theoretical argument above are shown in Fig. 5. In particular, the right plot shows visually the learning-induced risk aversion. In fact, the more Γ increases, the more the body of the distribution shifts in favor of the less-varying alternative. Notably, how we recover emergent risk-aversion is exactly in line with how standard thermophoresis—the particles' tendency to move to cooler regions in a solution with a nonvanishing temperature gradient—arises in physical systems [44].

Let us remark here that it is possible to compute analytically not only $P_*[\delta\hat{R}]$ but also the steady state PDF related to the average cumulated earned reward, represented by $R^A_t a_t + R^B_t(1-a_t)$. In fact, the earned reward at time t is governed by the difference in beliefs [see Eq. (1)]. This leads to an interesting insight, anticipated without proof in Sec. I: An irreversible sequence of belief updates may—and do in the present model—generate a time-reversible sequence of actions; furthermore, in the case of a stable environment like the one of the present setup, also the sequence of earned rewards is time reversible in the long run.

IV. TIME IRREVERSIBILITY IN BELIEFS DYNAMICS

This section is divided into two parts. First, we use Landauer's principle to define the irreversibility of belief dynamics, and then we argue using theoretical arguments that it is optimized in the steady state. Finally, numerical analysis relates the time irreversibility in the steady state to the exploration-exploitation trade-off in the coarse-grained modified forgetting Q-learning model.

A. Irreversibility as a measure of cognitive effort

The fundamental discovery encapsulated in Landauer's principle is that the average work dissipated by an actual machine to make the shift from $\hat{R}_t$ to $\hat{R}_{t+1}$ is bounded from below by the irreversibility rate Φ in units of kT, where k is the Boltzmann constant and T is the temperature of the room in which the system performing this operation is working. Below we detail how this statement can be formally established. This will lead naturally to the notion of irreversibility of belief dynamics we use in this discussion.

The irreversibility rate Φ is defined as the Kullback-Leibler divergence between the probability of observing a jump and its time-reversed counterpart [49,50], i.e.,

$$\Phi_t = D^{\mathrm{KL}}[P_t[\hat{R}_t \to \hat{R}_{t+1}]|P_t[\hat{R}_{t+1} \to \hat{R}_t]], \tag{19}$$

where $D^{\mathrm{KL}}[P|Q] = \int_x P(x) \log P(x)/Q(x)$. This divergence is appropriate for Markovian processes (like those we are considering in this work) [51]. Let us note that Φ_t is non-negative by construction and invariant under a homogeneous dilation of the state space.

The irreversibility rate can be exactly computed in the continuous-time limit for systems described by Langevin equations like Eqs. (13) using path integrals techniques [44,52]. One obtains

$$\Phi_t = \langle v_t \cdot \mathscr{F}_t \rangle, \tag{20}$$

where $v_t = J_t/P_t$ is the net directed velocity of the beliefs in the two-dimensional space $\hat{R}$ and $\langle\cdot\rangle$ stands for the average over P_t. Therefore, Φ is the dissipated power from the thermodynamic force $\mathscr{F}$ in units of kT. Hence, we identify the irreversibility rate Φ_t with the fundamental cognitive effort needed to perform a shift from $\hat{R}_t$ to $\hat{R}_{t+1}$ on average across all possible transitions.

Let us recover a previous result anticipated in Sec. III B 1, related to the NESS being generated by $\nabla \times \mathscr{F}$. Given the new quantity Φ we have introduced, $\Phi_* \neq 0$ for $\Gamma \neq 0$. This result can be recovered as follows. In the steady state, the velocity follows circulating lines (see the currents in Fig. 4 again and remember that $v_t = J_t/P_t$). One can calculate the average over the whole state space in Eq. (20) as an average over these closed lines. The dissipated power by the thermodynamic force on a closed loop is in general positive in the steady state for $\Gamma \neq 0$ because, by applying Stokes's theorem, the line integral receives a nonzero contribution from the surface integral of $\nabla \times \mathscr{F}$.

Equation (20) gives another interesting insight: in the steady state the velocity has to be aligned to the nonconservative part of $\mathscr{F}$ since we know that Φ_t is non-negative by construction. Appendix A will prove that, in the steady state, the velocity is maximally aligned with the nonconservative thermodynamic force compatible with a minimal dissipation along closed lines.

B. Numerical results

We now turn to analyzing the irreversibility rate Φ in the steady state of the coarse-grained modified forgetting Q-learning model.

First, we investigate whether Φ computed from the model-dependent formula given by Eq. (20) is compatible with available model-independent estimates. Later, we will investigate in deep the whole phase space sketched already in the left panel of Fig. 4. Note that to avoid incurring a degenerate diffusion matrix for large Γs, we add a small exogenous noise to the update equations (see Appendix B).

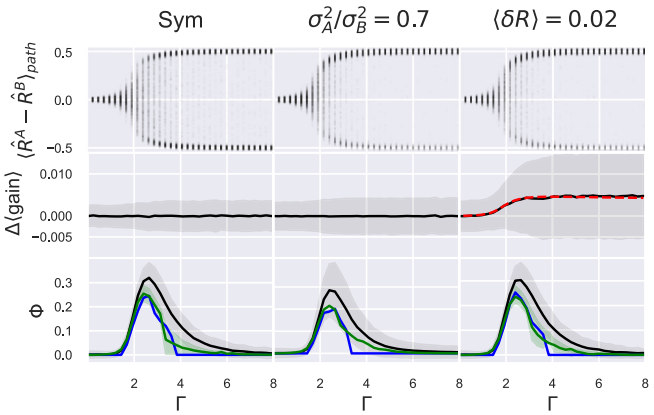

FIG. 6. Comparison of model-dependent irreversibility metric with model-independent estimations, for the case of modified forgetting Q-learning. Left: Symmetric bandits. Center: Asymmetric bandits in the variance of the rewards. Right: Asymmetric bandits in the average rewards: $\langle R^A \rangle = 0.51$ and $\langle R^B \rangle = 0.49$. Displayed metrics include (from top to bottom) average difference in beliefs, average earned reward minus the one related to $\Gamma = 0$, and irreversibility rate. The red line in the central plot is related to the theoretical value of this metric. The lower panel's black, blue, and green lines represent Φ calculated from Monte Carlo simulations, a neural network, and a gradient boosting approach, respectively. These estimators are based on the same set of trajectories. Initial conditions for larger Γs are given by the equilibrium belief distribution obtained from the previous one. The total simulation time is 10^4, and we retain only the second half of the trajectories.

1. *Consistency checks*

We consider three different scenarios belonging to the case $\chi^A, \chi^B \ll 1$ (red region in the left plot of Fig. 4): the case of completely symmetric arms, the case with asymmetric variances (respectively shown already in the left and right plots of Fig. 5), and finally the case of asymmetric average rewards. For each scenario, three metrics are exhibited in Fig. 6 and discussed below, from top to bottom. Before proceeding, let us remark that to have a clear picture of steady states, we used an iterative numerical scheme to choose the initial condition of the simulations.

$\langle \hat{R}^A - \hat{R}^B \rangle_{\mathrm{path}}$: Each point corresponds to the average difference in beliefs for fixed trajectory.

By looking at this metric, one can see that trapping states emerge at high exploitation levels. Moreover, this metric clearly shows the average fraction of time passed in a given belief state.

$\Delta\langle \mathrm{gain} \rangle$: Each point corresponds to the average earned reward across the trajectories minus the one obtained with passive learning ($\Gamma = 0$). The red line is obtained analytically starting from Eq. (18) (see the discussion at the end of Sec. III B 2).

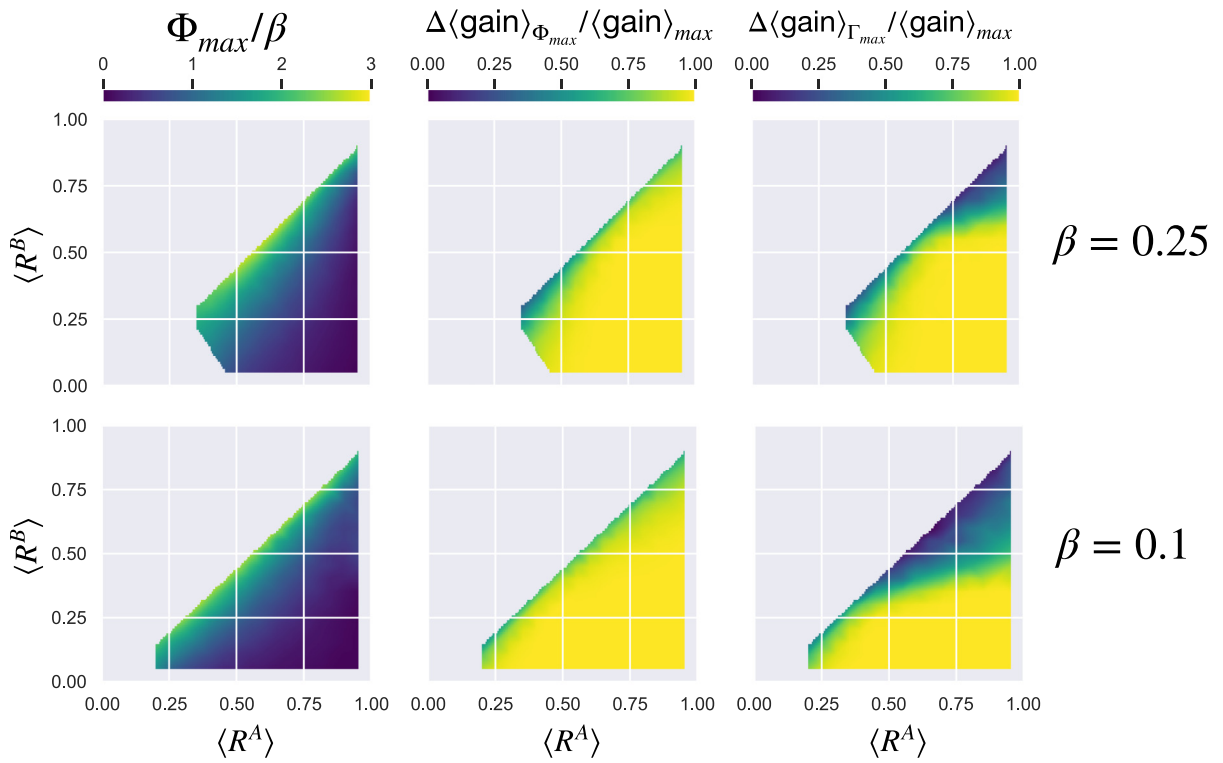

FIG. 7. Analysis of irreversibility rate and average earned reward in the steady state of the modified forgetting Q-learning model, varying the bandit's configuration. From left to right, three key metrics are exhibited: the maximal irreversibility $\Phi_{\max}$ (divided by β), the excess normalized gain for the Γ related to $\Phi_{\max}$, and the one for $\Gamma = 5/\langle R^B\rangle$, the normalization being with respect to passive learning. Top and bottom plots, respectively, refer to different β values. Initial conditions for the beliefs dynamics are uniformly distributed. The total simulation time is 10^4, and we retain only the second half of the trajectories.

This metric reflects the agent's average reward earned $[R_t^A a_t + R_t^B(1-a_t)]$, thereby quantifying the system's operational efficiency. Note that when the arms have different average rewards (right plot), a high average earned reward is obtained for moderate Γs.

Φ: Each point corresponds to the average irreversibility rate across the trajectories.

To compute the irreversibility rate numerically from Monte Carlo simulations, we note that Eq. (20) after an integration by parts, leads to [53]

$$\Phi_t = \langle (\mathcal{F}_t - \nabla\cdot\mathcal{D}_t)^{\mathsf{T}}\mathcal{D}_t^{-1}(\mathcal{F}_t - \nabla\cdot\mathcal{D}_t)\rangle - \langle \nabla\cdot(\mathcal{F}_t - \nabla\cdot\mathcal{D}_t)\rangle, \tag{21}$$

where $\mathcal{F}_t$ and $\mathcal{D}_t$ are derived explictely in Eq. (B2) and Eq. (B3), respectively.

On top of the black line provided by Eq. (21), two additional benchmarks calculated directly from Eq. (19) are presented: The blue line is based on a recently proposed neural network approach [54], while the green line is provided by an algorithm [55] that maps the problem of calculating the irreversibility rate onto a classification task [56], by leveraging on gradient boosting techniques. Crucially, these additional estimators do not need any information about the model except the tacitly assumed Markovian property assumed in the definition of the irreversibility rate given by Eq. (19). The reason why the Monte Carlo estimator is consistently above the others is related to the fact that no spatial coarse-graining is applied in this case since full knowledge of the underlying model is provided.

Let us now comment on these results. The irreversibility rate Φ is null at both exploitation parameter extremities, in sync with previous analyses done in this paper. However, the numerical analysis revealed a noteworthy crest at median exploitation parameters; this indicates a belief dynamics propitious to humans in asymmetric bandit scenarios, being comfortably distant from bifurcation-prone zones, as one can appreciate from the top panels depicting the average distance in beliefs. A similar conclusion can be reached by looking at the top and center plots in the asymmetric bandit scenarios, where one can see that the exploitation level related to the maximal irreversibility rate corresponds to a heightened average earned reward and lowered average earned reward variability, respectively.

2. Irreversibility across bandit configurations and sensitivity levels

In order to shed light on the dynamics across the whole bandit's configuration space, in Fig. 7, we show the results of extensive simulations for $\Gamma \in [0, 5/\langle R^B\rangle]$, for the cases with $\beta = 0.25$ (top plots) and $\beta = 0.1$ (bottom plots). From left to right we reported three metrics: the ratio between the maximal irreversibility rate Φ and the learning rate β for the given configuration of average rewards and varying Γ, the normalized excessive gain when Φ is at its maximum, and when $\Gamma = 5/\langle R^B\rangle$. We excluded the top half triangle since it is

completely symmetric with respect to the diagonal. Moreover, as we explained in Sec. II C, if $\chi^A, \chi^B > 1/3$ (bottom left corner), then the beliefs dynamics never decouples. As a consequence, the more Γ increases, the more the irreversibility rate increases. We excluded this region from the one reported in Fig. 7 for plot visibility reasons.

Interestingly, from the left plots, one can recognize that lowering the difficulty of the task (i.e., increasing the distance between $\langle R^B \rangle$ and $\langle R^A \rangle$) lowers the associated maximum irreversibility. From the comparison between the center and right plots, one can conclude that the region with maximal irreversibility is more beneficial than the one for larger Γs. The reason for this is that, as we explained earlier, for very large Γs the condition $\hat{R}^A \gg \hat{R}^B$ and the reverse one are both stable (i.e., the belief dynamics is nonergodic) and therefore the chances to end up on the suboptimal arm increase. In particular, one can compare the center plots with the sketch of the phase space we draw purely from theoretical consideration (see left panel in Fig. 4). Finally, the reason why we plotted the ratio Φ/β is to have another consistency check: Due to dimensional analysis considerations, if we let vary β for fixed Γs, then the irreversibility rate will scale as β if one does not enter in the region where $\chi^A, \chi^B > 1/3$ (bottom left corner).

V. DISCUSSION

According to Landauer's principle, we linked the irreversibility rate associated with beliefs dynamics to a thermodynamically consistent measure of cognitive effort in a simple but paradigmatic setup: forgetting Q-learning dynamics [37] in two-armed bandit tasks.

First, we provide a general mapping of the decision-making model onto one with active particles, i.e., particles able to spend energy to move. A side result is the formal identification of learning-induced risk-aversion and standard thermophoresis, providing an analytical description of this phenomenon already known in the cognitive neuroscience literature for the first time. The combination of theoretical and numerical analysis has shown that intermediate exploitative behavior produces maximum irreversibility in belief dynamics for sufficiently high sensitivity levels. Moreover, this peak in irreversibility aligns with a beneficial trade-off between exploration and exploitation. Despite increasing the sensitivity might be useful in distinguishing similar bandit configurations, it comes at an intrinsic penalty: it increases the chances of being stuck on the suboptimal arm. Therefore, irreversibility acts as a thermometer for belief coupling. If beliefs are coupled, then the decision maker oscillates stochastically between the two arms but spends more time playing on the best option, if available. Moreover, extensive numerical analysis allowed us to conclude that less irreversibility is required to excel in the game if the task difficulty is reduced. This finding fosters the link between irreversibility in belief dynamics and a plausible proxy of cognitive effort. Interestingly, the higher the sensitivity, the more crucial it becomes to search for the exploitation parameter that maximizes the irreversibility rate to achieve superior gains. Therefore, this stylized model suggests a plausible evolutionary mechanism that underscores the likelihood of biological entities being optimized to function in maximally out-of-equilibrium states [57].

As mentioned in the main sections, we do not expect our findings to be qualitatively modified in the case of standard Q-learning. However, analyzing belief dynamics with a belief and arm-dependent spring requires extra care. Subsequent analysis will explore this and other interesting learning rules, such as those affected by confirmation or positivity bias. Interestingly, these latter models will yield more exotic thermoporetic effects [33,58,59]: for instance, positivity bias is known to lead to emergent risk-seeking behavior [32] which, under the lens of the present paper, could be well accounted for by negative thermophoresis-like dynamics.

ACKNOWLEDGMENTS

I am indebted to Christian Bongiorno for his indispensable assistance with the numerical aspects of this research and to Stefano Palminteri for suggesting relevant literature in the cognitive neuroscience domain. The discussion was significantly enriched by Damien Challet's broader perspectives on the topic. I also acknowledge fruitful discussions with Cristiano Pacini, Matteo Marsili, Massimo Vergassola, Stefano Celani, Edgar Roldan, Matteo Sireci, Daniel Busiello, and Walter Quattrociocchi. The author is supported by the Agence National de la Recherche (CogFinAgent: ANR-21-CE23-0002-02).

APPENDIX A: ANALYSIS OF LYAPUNOV FUNCTION

In order to have insights about how Φ is optimized as the NESS is reached, it is useful to study a particular Lyapunov function of the dynamics. A Lyapunov function is such that its temporal derivative is always nonpositive, meaning that its fixed point corresponds to the steady state of the dynamics.

Consider the function $\mathcal{L}$ given by

$$\mathcal{L} = D^{\mathrm{KL}}[P_t|P_*]. \tag{A1}$$

By following Sireci and Busiello [60], by taking the time derivative of $\mathcal{L}$ and inserting the Fokker-Planck equation one obtains:

$$\frac{d\mathcal{L}_t}{dt} = -\Pi_t + \langle v_t \mathcal{D}^{-1} v_* \rangle \tag{A2}$$

$$= -\left\langle (v_t - v_*)\mathcal{D}_t^{-1}(v_t - v_*) \right\rangle \leqslant 0, \tag{A3}$$

where $\Pi_t = \langle v_t \mathcal{D}^{-1} v_t \rangle$ in the first equality is the so-called entropy production in the stochastic thermodynamics literature [45,49,61,62]. The second equality can be established by noting that [63] $\langle v_t \mathcal{D}^{-1} v_* \rangle = \langle v_* \mathcal{D}^{-1} v_* \rangle$. The final inequality in Eq. (A3), trivially follows since the final term is quadratic in $v_t - v_*$ and $\mathcal{D}$ is semipositive definite by construction. This proves that $\mathcal{L}_t$ is a Lyapunov function of the dynamics.

Let us make an important remark: $\Pi_t \geqslant 0$ by definition because it is quadratic in the thermodynamic velocities v_t and inversely proportional to the diffusion matrix, which is semipositive definite by construction. In ESSs, $\Pi = 0$ because $v = 0$ by definition; therefore, $\Pi > 0$, i.e., the case where currents are present, is a clear marker of irreversible dynamics.

Following a similar reasoning, one can see that in Eq. (A2) the negative time derivative of the Lyapunov function has been written as the sum of a nonpositive and a non-negative term (remember that $\langle v \mathcal{D}^{-1} v_* \rangle = \langle v_* \mathcal{D}^{-1} v_* \rangle$), suggesting that in

the vicinity of the steady state, the first is maximized and the second is minimized.

Interestingly, the second term in Eq. (A2) can be rewritten by simply using the identity $v_* = J_*/P_*$ and the definition of J given by Eq. (15). One obtains

$$\langle v\mathcal{D}^{-1}v_*\rangle = \Phi_t - \langle v\cdot\boldsymbol{\nabla}\log[P_*]\rangle. \tag{A4}$$

In the steady state, the second term in the right-hand side of the equation above can be rewritten after a partial integration as $\langle\boldsymbol{\nabla}\cdot v_*\rangle_*$, where $\langle\cdot\rangle_*$ indicates an average over the steady state PDF P_*; this term has to be zero in a NESS with a compact state space since the occupied state space in the steady state is no longer contracting or expanding. Therefore, along the dynamics, $d\mathcal{L}_t/dt$ goes to zero by minimizing the entropy production $\Pi_t = \langle v_t\mathcal{D}^{-1}v_t\rangle$ while maximizing the dissipation of the thermodynamic force along the closed lines created in the vicinity of the steady state by probability currents.

From Eq. (A2) evaluated in the steady state one obtains the well-known result $\Phi_* = \Pi_*$, i.e., in the steady state the irreversibility rate, also known as entropy flux, is equal to the entropy production. The equation $\Phi_* = \Pi_*$ can be interpreted as a form of energy conservation, echoing the interpretation in physics. In fact, the entropy flux is the average dissipated power in units of kT done by the thermodynamic force $\mathscr{F}$, as previously emphasized. On the other hand, Π is analogous to the kinetic energy of the active particle with velocity field v_t and mass $\mathcal{D}^{-1}$; this is tantamount to saying that the inertia of the particle is lower in a noisier environment.

Let us recapitulate what we have obtained.

The main result of this section is that the combination of Eq. (A2), (A3), and (A4) implies that the NESS is the least dissipative state compatible with a velocity that is maximally aligned with the nonconservative part of thermodynamic force $\mathscr{F}$, therefore suggesting an efficient (thermodynamically speaking) information processing in the steady state [44].

APPENDIX B: MODEL USED FOR SIMULATIONS

The model for which we are going to investigate quantitatively Φ_* is given by:

$$\begin{aligned}\frac{d\hat{R}_t^A}{dt} &= -\beta\left(\hat{R}_t^A + a_t R_t^A + \eta_t^A\right)\\ \frac{d\hat{R}_t^B}{dt} &= -\beta\left[\hat{R}_t^A + (1-a_t)R_t^B + \eta_t^B\right],\end{aligned} \tag{B1}$$

where we made one modification with respect to Eq. (8): We added two small exogenous white noises, η_t^A and η_t^B, which are needed in order to have a well-defined two-dimensional diffusion matrix in the large-Γ region, where, in the absence of such noises, it would become a singular matrix. I set the variances of η_t^A and η_t^B so that $\mathrm{var}[\eta^A] = \mathrm{var}(\eta^B) = \sigma_\eta^2 \ll \sigma_A^2, \sigma_B^2$.

The derivation of the Fokker-Planck equation (see Sec. III B) leads to:

$$\mathcal{F}_t = \beta\begin{bmatrix} -\hat{R}_t^A + a_t\langle R^A\rangle \\ -\hat{R}_t^B + (1-a_t)\langle R^B\rangle\end{bmatrix} \tag{B2}$$

and

$$\mathcal{D}_t = (\beta/2)^2\begin{bmatrix}\sigma_A^2 a_t^2 + \sigma_\eta^2, & 0\\ 0, & \sigma_B^2(1-a_t)^2 + \sigma_\eta^2\end{bmatrix}. \tag{B3}$$

These are the expressions of $\mathcal{F}$ and $\mathcal{D}$ we use to quantify the irreversibility rate from Monte Carlo simulations by means of Eq. (21) in Fig. 6.

[1] A. Rangel, C. Camerer, and P. R. Montague, A framework for studying the neurobiology of value-based decision making, Nat. Rev. Neuro. **9**, 545 (2008).

[2] G. Piccinini and A. Scarantino, Information processing, computation, and cognition, J. Biol. Phys. **37**, 1 (2011).

[3] C. Henry, Investment decisions under uncertainty the "irreversibility effect", Am. Econ. Rev. **64**, 1006 (1974).

[4] C. W. Lynn, E. J. Cornblath, L. Papadopoulos, M. A. Bertolero, and D. S. Bassett, Broken detailed balance and entropy production in the human brain, Proc. Natl. Acad. Sci. USA **118**, e2109889118 (2021).

[5] Y. Sanz Perl, H. Bocaccio, C. Pallavicini, I. Pérez-Ipiña, S. Laureys, H. Laufs, M. Kringelbach, G. Deco, and E. Tagliazucchi, Nonequilibrium brain dynamics as a signature of consciousness, Phys. Rev. E **104**, 014411 (2021).

[6] G. Deco, Y. Sanz Perl, H. Bocaccio, E. Tagliazucchi, and M. L. Kringelbach, The INSIDEOUT framework provides precise signatures of the balance of intrinsic and extrinsic dynamics in brain states, Commun. Bio. **5**, 572 (2022).

[7] M. Gilson, E. Tagliazucchi, and R. Cofré, Entropy production of multivariate Ornstein-Uhlenbeck processes correlates with consciousness levels in the human brain, Phys. Rev. E **107**, 024121 (2023).

[8] P. K. Tewarie, R. Hindriks, Y. M. Lai, S. N. Sotiropoulos, M. Kringelbach, and G. Deco, Non-reversibility outperforms functional connectivity in characterisation of brain states in MEG data, NeuroImage **276**, 120186 (2023).

[9] D. Bernardi, D. Shannahoff-Khalsa, J. Sale, J. A. Wright, L. Fadiga, and D. Papo, The time scales of irreversibility in spontaneous brain activity are altered in obsessive compulsive disorder, Front. Psychiatr. **14**, 1158404 (2023).

[10] A. Safron, D. A. Sakthivadivel, Z. Sheikhbahaee, M. Bein, A. Razi, and M. Levin, Making and breaking symmetries in mind and life, Interface Focus **13**, 20230015 (2023).

[11] C. Padoa-Schioppa and K. E. Conen, Orbitofrontal cortex: A neural circuit for economic decisions, Neuron **96**, 736 (2017).

[12] S. Bavard, A. Rustichini, and S. Palminteri, Two sides of the same coin: Beneficial and detrimental consequences of range adaptation in human reinforcement learning, Sci. Adv. **7**, eabe0340 (2021).

[13] S. Bavard and S. Palminteri, The functional form of value normalization in human reinforcement learning, Elife **12**, e83891 (2023).

[14] D. Wolpert, J. Korbel, C. Lynn, F. Tasnim, J. Grochow, G. Kardeş, J. Aimone, V. Balasubramanian, E. de Giuli, D. Doty

et al., Is stochasticthermodynamics the key to understanding the energy costs of computation? arXiv:2311.17166.
[15] M. Marsili, Market mechanism and expectations in minority and majority games, Phys. A **299**, 93 (2001).
[16] M. Wyart and J.-P. Bouchaud, Self-referential behaviour, overreaction and conventions in financial markets, J. Econ. Behav. Organ. **63**, 1 (2007).
[17] M. Vodret, I. Mastromatteo, B. Tóth, and M. Benzaquen, Microfounding GARCH models and beyond: A Kyle-inspired model with adaptive agents, J. Econ. Interact. Coord. **18**, 599 (2023).
[18] R. E. Farmer, *The Macroeconomics of Self-fulfilling Prophecies* (MIT Press, Cambridge, MA, 1999).
[19] J.-P. Bouchaud and R. E. Farmer, Self-fulfilling prophecies, quasi nonergodicity, and wealth inequality, J. Political Econ. **131**, 947 (2023).
[20] M. Marsili, F. Vega-Redondo, and F. Slanina, The rise and fall of a networked society: A formal model, Proc. Natl. Acad. Sci. USA **101**, 1439 (2004).
[21] J. da Gama Batista, J.-P. Bouchaud, and D. Challet, Sudden trust collapse in networked societies, Eur. Phys. J. B **88**, 55 (2015).
[22] M. Cinelli, G. De Francisci Morales, A. Galeazzi, W. Quattrociocchi, and M. Starnini, The echo chamber effect on social media, Proc. Natl. Acad. Sci. USA **118**, e2023301118 (2021).
[23] D. Mocanu, L. Rossi, Q. Zhang, M. Karsai, and W. Quattrociocchi, Collective attention in the age of (mis) information, Comput. Hum. Behav. **51**, 1198 (2015).
[24] L. Frisell, A theory of self-fulfilling political expectations, J. Public Econ. **93**, 715 (2009).
[25] D. Rothschild and N. Malhotra, Are public opinion polls self-fulfilling prophecies? Res. Polit. **1**, 1 (2014).
[26] R. K. Merton, The self-fulfilling prophecy, Antioch Rev. **8**, 193 (1948).
[27] N. D. Daw, *Advanced Reinforcement Learning*, 2nd ed. (Academic Press, 2014), Chap. 16, pp. 299–320.
[28] J. P. O'Doherty, A. Hampton, and H. Kim, Model-based fmri and its application to reward learning and decision making, Ann. N.Y. Acad. Sci. **1104**, 35 (2007).
[29] K. Morita and A. Kato, Striatal dopamine ramping may indicate flexible reinforcement learning with forgetting in the cortico-basal ganglia circuits, Front. Neural Circ. **8**, 36 (2014).
[30] A. Kato and K. Morita, Forgetting in reinforcement learning links sustained dopamine signals to motivation, PLoS Comput. Biol. **12**, e1005145 (2016).
[31] W. Schultz, Dopamine reward prediction error coding, Dialog. Clinic. Neurosci. **18**, 23 (2016).
[32] S. Palminteri and M. Lebreton, The computational roots of positivity and confirmation biases in reinforcement learning, Trends Cogn. Sci. **26**, 607 (2022).
[33] S. Palminteri, Choice-confirmation bias and gradual perseveration in human reinforcement learning, Behav. Neurosci. **137**, 78 (2023).
[34] R. Landauer, Irreversibility and heat generation in the computing process, IBM J. Res. Dev. **5**, 183 (1961).
[35] M. P. Frank, Physical foundations of Landauer's principle, in *Reversible Computation* (Springer, Berlin, 2018), pp. 3–33.
[36] A. Bérut, A. Arakelyan, A. Petrosyan, S. Ciliberto, R. Dillenschneider, and E. Lutz, Experimental verification of Landauer's principle linking information and thermodynamics, Nature (Lond.) **483**, 187 (2012).
[37] K. Katahira, The relation between reinforcement learning parameters and the influence of reinforcement history on choice behavior, J. Math. Psychol. **66**, 59 (2015).
[38] S. E. Seidenbecher, J. I. Sanders, A. C. von Philipsborn, and D. Kvitsiani, Reward foraging task and model-based analysis reveal how fruit flies learn value of available options, PLoS One **15**, e0239616 (2020).
[39] J. Cruzat, R. Herzog, P. Prado, Y. Sanz-Perl, R. Gonzalez-Gomez, S. Moguilner, M. L. Kringelbach, G. Deco, E. Tagliazucchi, and A. Ibañez, Temporal irreversibility of large-scale brain dynamics in Alzheimer's disease, J. Neurosci. **43**, 1643 (2023).
[40] G. Weiss, Time-reversibility of linear stochastic processes, J. Appl. Probab. **12**, 831 (1975).
[41] N. Salem-Garcia, S. Palminteri, and M. Lebreton, Linking confidence biases to reinforcement-learning processes, Psychol. Rev. **130**, 1017 (2023).
[42] C. W. Gardiner *et al.*, *Handbook of Stochastic Methods*, Vol. 3 (Springer, Berlin, 1985).
[43] M. Mendler and B. Drossel, Predicting properties of the stationary probability currents for two-species reaction systems without solving the Fokker-Planck equation, Phys. Rev. E **102**, 022208 (2020).
[44] D. M. Busiello, S. Liang, and P. De Los Rios, Emergent thermophoretic behavior in non-equilibrium chemical systems, Bull. Am. Phys. Soc. **68**, N14.00004 (2023).
[45] D. M. Busiello, J. Hidalgo, and A. Maritan, Entropy production for coarse-grained dynamics, New J. Phys. **21**, 073004 (2019).
[46] T. Chou, K. Mallick, and R. K. Zia, Non-equilibrium statistical mechanics: From a paradigmatic model to biological transport, Rep. Prog. Phys. **74**, 116601 (2011).
[47] X. Fang, K. Kruse, T. Lu, and J. Wang, Nonequilibrium physics in biology, Rev. Mod. Phys. **91**, 045004 (2019).
[48] J. G. March, Learning to be risk averse, Psychol. Rev. **103**, 309 (1996).
[49] U. Seifert, Stochastic thermodynamics, fluctuation theorems and molecular machines, Rep. Prog. Phys. **75**, 126001 (2012).
[50] É. Roldán, Facultad de ciencias físicas, Ph.D. thesis, Universidad Complutense de Madrid, 2013.
[51] We note that Eq. (19) can be derived by a more general formulation valid for non-Markovian processes, where the primary object is the Kullback-Leibler divergence between the probability of an actual path and its time-reversed twin [64].
[52] L. F. Cugliandolo and V. Lecomte, Rules of calculus in the path integral representation of white noise Langevin equations: The Onsager–Machlup approach, J. Phys. A: Math. Theor. **50**, 345001 (2017).
[53] Interestingly, the irreversibility rate is composed by two terms: The first corresponds to exploration, while the second is associated with the contraction of the belief space due to forgetting [65].
[54] D.-K. Kim, Y. Bae, S. Lee, and H. Jeong, Learning entropy production via neural networks, Phys. Rev. Lett. **125**, 140604 (2020).

[55] M. Vodret, C. Pacini, and C. Bongiorno, Functional decomposition and estimation of irreversibility in time series via machine learning, arXiv:2407.06063.

[56] A. Seif, M. Hafezi, and C. Jarzynski, Machine learning the thermodynamic arrow of time, Nat. Phys. **17**, 105 (2021).

[57] L. M. Martyushev, Entropy and entropy production: Old misconceptions and new breakthroughs, Entropy **15**, 1152 (2013).

[58] V. Chambon, H. Théro, M. Vidal, H. Vandendriessche, P. Haggard, and S. Palminteri, Information about action outcomes differentially affects learning from self-determined versus imposed choices, Nat. Hum. Behav. **4**, 1067 (2020).

[59] M. Sugawara and K. Katahira, Dissociation between asymmetric value updating and perseverance in human reinforcement learning, Sci. Rep. **11**, 3574 (2021).

[60] M. Sireci and D. M. Busiello, Dissipative symmetry breaking in non-equilibrium steady states, arXiv:2212.03353.

[61] U. Seifert, Entropy production along a stochastic trajectory and an integral fluctuation theorem, Phys. Rev. Lett. **95**, 040602 (2005).

[62] T. Tomé, Entropy production in nonequilibrium systems described by a Fokker-Planck equation, Braz. J. Phys. **36**, 1285 (2006).

[63] C. Van den Broeck and M. Esposito, Three faces of the second law. II. Fokker-Planck formulation, Phys. Rev. E **82**, 011144 (2010).

[64] É. Roldán and J. M. R. Parrondo, Estimating dissipation from single stationary trajectories, Phys. Rev. Lett. **105**, 150607 (2010).

[65] D. Daems and G. Nicolis, Entropy production and phase space volume contraction, Phys. Rev. E **59**, 4000 (1999).