entropy

MDPI

Article

Non-Equilibrium Relations for Bounded Rational Decision-Making in Changing Environments

Jordi Grau-Moya [1,2,3], Matthias Krüger [1,4] and Daniel A. Braun [1,2,5,*]

1 Max Planck Institute for Intelligent Systems, Stuttgart 70569, Germany; jordi.grau.mo@gmail.com (J.G.-M.); mkrueger@is.mpg.de (M.K.)
2 Max Planck Institute for Biological Cybernetics, Tübingen 72076, Germany
3 PROWLER.io, Cambridge CB2 1LA, UK
4 4th Institute for Theoretical Physics, Universität Stuttgart, Stuttgart 70569, Germany
5 Institute of Neural Information Processing, Universität Ulm, Ulm 89081, Germany
* Correspondence: daniel.braun@uni-ulm.de; Tel.: +49-731-5024150

Received: 30 July 2017; Accepted: 18 December 2017; Published: 21 December 2017

Abstract: Living organisms from single cells to humans need to adapt continuously to respond to changes in their environment. The process of behavioural adaptation can be thought of as improving decision-making performance according to some utility function. Here, we consider an abstract model of organisms as decision-makers with limited information-processing resources that trade off between maximization of utility and computational costs measured by a relative entropy, in a similar fashion to thermodynamic systems undergoing isothermal transformations. Such systems minimize the free energy to reach equilibrium states that balance internal energy and entropic cost. When there is a fast change in the environment, these systems evolve in a non-equilibrium fashion because they are unable to follow the path of equilibrium distributions. Here, we apply concepts from non-equilibrium thermodynamics to characterize decision-makers that adapt to changing environments under the assumption that the temporal evolution of the utility function is externally driven and does not depend on the decision-maker's action. This allows one to quantify performance loss due to imperfect adaptation in a general manner and, additionally, to find relations for decision-making similar to Crooks' fluctuation theorem and Jarzynski's equality. We provide simulations of several exemplary decision and inference problems in the discrete and continuous domains to illustrate the new relations.

Keywords: free energy; bounded rationality; anticipation; adaptation

1. Introduction

A number of recent studies has pointed out mathematical equivalences between thermodynamic systems described by statistical mechanics and information processing systems [1–4]. In particular, it has been suggested that decision-makers with constrained information-processing resources can be described in analogy to closed physical systems in contact with a heat bath that seek to minimize energy [1]. In this analogy, decision-makers can be thought to act in a way that minimizes a cost function or, equivalently, that maximizes a utility function in lieu of an energy function. Classic decision theory [5,6] states that, given a set of actions $\mathcal{X}$ and a set of observations $\mathcal{O}$, the perfectly rational decision-maker should choose the best possible action $x^* \in \mathcal{X}$ that maximizes the expected utility $U(x)$:

$$x^* = \underset{x}{\operatorname{argmax}}\, U(x) = \underset{x}{\operatorname{argmax}} \sum_{o \in \mathcal{O}} p(o|x) V(o), \tag{1}$$

where $p(o|x)$ is the probability of the outcome o given action x and $V(o)$ indicates the utility of this outcome. However, maximizing the expected utility is in general a costly computational operation that real decision-makers might not be able to perform.

Decision-makers that are unable to choose the best possible action x^* due to a lack of computational resources have traditionally been studied in the field of bounded rationality. Originally proposed by Herbert Simon [7,8], bounded rationality comprises a medley of approaches ranging from optimization-based approaches like bounded optimality (searching for the program that achieves the best utility performance on a particular platform) [9–11] and meta-reasoning (optimizing the cost of reasoning) [12–14] to heuristic approaches that reject the notion of optimization [15–17]. Recently, new impulses for the development of bounded rationality theory have come from information-theoretic and thermodynamic perspectives on the general organization of perception-action-systems [1,3,18–27]. In the economic and game-theoretic literature, these models have precursors that have studied bounded rationality inspired by stochastic choice rules originally proposed by Luce, McFadden and others [2,28–39]. In most of these models, decision-makers face a trade-off between the attainment of maximum utility and the required information-processing cost measured as an entropy or relative entropy. The optimal solution to this trade-off usually takes the form of a Boltzmann-like distribution analogous to equilibrium distributions in statistical physics. The decision-making process can then be conceptualized as a change from a prior strategy distribution to a posterior strategy distribution, where the change is triggered by a change in the utility landscape. However, studying changes in equilibrium distributions neglects not only the time required for this change, but also the adaptation process itself.

The main contribution of this paper is to show that the analogy between equilibrium thermodynamics and bounded-rational decision-making [1] can be extended to the non-equilibrium domain under the assumption that the temporal evolution of the utility function is externally driven and does not depend on the decision-maker's action. This allows for new predictions that can be tested in experimental setups investigating decision-makers that choose between multiple alternatives. When given sufficient time to adjust to the problem such a decision-maker may achieve a bounded optimal performance given the available precision, which may be described by an equilibrium distribution; for example, a dart thrower that has fully adapted her/his personal best performance after extensive training with prism glasses. However, if given insufficient time, the decision-maker may not achieve bounded optimal performance, but only an inferior performance biased by the specific information-processing mechanisms used by the decision-maker, which may in general be described by a non-equilibrium distribution; for example, a dart thrower that is wearing prism glasses for the first time and plays according to a non-adaptive strategy thereby "dissipating" utility. The connection between the non-equilibrium and equilibrium domains is tied with the concept of dissipation and its role in fluctuation theorems, which are important recent results in non-equilibrium thermodynamics.

The paper is organized as follows. In Section 2, we recapitulate the relation between bounded rational decision-making and equilibrium thermodynamics. In Section 3, we relate decision-making processes to non-equilibrium thermodynamics. In Section 4, we generalize concepts from non-equilibrium thermodynamics to make them applicable to a wider range of decision-making problems. In particular, we include a derivation of a generalized Jarzynski equality and a generalized Crooks' theorem for decision-making. We provide simulations to illustrate the new relations in different decision-making scenarios. In Section 5, we discuss our results.

2. Equilibrium Thermodynamics and Decision-Making

In thermodynamics, closed physical systems in thermal equilibrium with their environment are described by equilibrium distributions that do not change over time. For example, a gas in a box distributes its particles evenly over the entire space and will stay this way and not spontaneously concentrate in a corner of the box. When changing constraints of the physical system, equilibrium thermodynamics allows predicting the final state after the change has taken place. For example, when opening a divider between two boxes, the gas will expand further until it fills the entire space evenly. This way, equilibrium thermodynamics allows describing system behaviour as a change

from a prior equilibrium distribution to a posterior equilibrium distribution triggered by a change in external constraints.

On an abstract level, one can think about changes in the distribution of a random variable from a prior to a posterior distribution as the basis of information-processing. In Bayesian inference, for example, we update current prior beliefs $p_0(x)$ by means of a likelihood to obtain a posterior belief $p_1(x)$. Similarly, decision-making can be regarded as a process of changing a prior strategy $p_0(x)$ to a posterior strategy $p_1(x)$ through a process of deliberation [1], thereby emphasizing the stochastic nature of choice [40]. According to [1], such transitions from prior to posterior with information constraints can be formalized by optimizing the variational problem:

$$p_1^{\text{eq}}(x) = \underset{p}{\text{argmax}} \ \Delta F[p] \tag{2}$$

where:

$$\Delta F[p] := \sum_x p(x)\Delta U(x) - \frac{1}{\beta} D_{\text{KL}}(p||p_0), \tag{3}$$

is a free energy functional, $\Delta U(x)$ is a change in utility (analogous to the notion of gains and losses in prospect theory [15]), $D_{\text{KL}}(\cdot||\cdot)$ is the Kullback–Leibler divergence or relative entropy and β is a real-valued parameter that translates from informational units into utility units. Accordingly, Equation (3) optimizes a trade-off between utility gains and information-processing resources quantified by the "information distance" between prior and posterior. In a physical system (where the energy function corresponds to a negative utility), Equation (3) evaluated at the optimum p_1^{eq} quantifies the negative free energy difference $\Delta F[p_1^{\text{eq}}]$ between the final state 1 and the initial state 0 assuming an isothermal process with respect to the inverse temperature β and a negative energy difference of $\Delta U = U_1 - U_0$.

For a given information cost parameter β, the bounded rational decision-maker optimally trades off utility gain against informational resources according to Equation (2), thereby following the strategy:

$$p_1^{\text{eq}}(x) = \frac{1}{Z_\beta} p_0(x) e^{\beta \Delta U(x)} \tag{4}$$

with partition function $Z_\beta = \sum_x p_0(x) e^{\beta \Delta U(x)}$. When inserting the optimal strategy $p_1^{\text{eq}}(x)$ into Equation (3), the certainty-equivalent value of strategy p_1^{eq} is determined by

$$\Delta F^{\text{eq}} := \Delta F[p_1^{\text{eq}}] = \frac{1}{\beta} \log Z_\beta. \tag{5}$$

For $\beta \to 0$, the cost of computation dominates, and the optimal strategy is given by the prior strategy $p_1^{\text{eq}}(x) = p_0(x)$ with the value $\lim_{\beta \to 0} \Delta F[p_1^{\text{eq}}] = \langle \Delta U(x) \rangle_{p_0(x)}$. This models a decision-maker that cannot afford any information-processing. When information costs are low ($\beta \to \infty$), the optimal strategy $p_1^{\text{eq}}(x)$ places all the probability mass on the maximum of $\Delta U(x)$, and the value of the strategy is $\lim_{\beta \to \infty} \Delta F[p_1^{\text{eq}}] = \max_x \Delta U(x)$. This models a perfectly rational decision-maker that can hand pick the best action. While this model includes maximum (expected) utility decision-making of Equation (1) as a special case, note that conceptually, the formulation of the decision problem as a variational problem in the probability distribution is very different from traditional approaches that define an optimization problem directly in the space of actions.

One possible objection to the strategy (4) is that it requires computing the partition sum Z_β over all possible actions, which is in general an intractable operation; even though Equation (4) could still be of descriptive value. It should be noted, however, that the decision-maker is not required to explicitly compute $p_1^{\text{eq}}(x)$; it suffices to produce a sample from $p_1^{\text{eq}}(x)$ to generate a decision. This can be achieved, for example, by Markov Chain Monte Carlo (MCMC) methods that are specifically designed to avoid the explicit computation of partition sums [41]. In the following, we recapitulate

two simple MCMC examples in the context of decision-making: a bounded rational decision-maker that uses a rejection sampling scheme and a bounded rational decision-maker that uses a variant of the Metropolis–Hastings scheme [42].

Exemplary Bounded Rational Decision-Makers

The optimal distribution (4) can be implemented, for example, by a decision-maker that follows a probabilistic satisficing strategy with aspiration level $T \geq \max_x \Delta U(x)$. Such a decision-maker optimizes the utility $\Delta U(x)$ by drawing samples from the prior distribution $x_s \sim p_0(x)$ and accepts with certainty the first sample x_s with utility $\Delta U(x_s) \geq T$ reaching the aspiration level T or any sample with utility below the aspiration level with acceptance probability $p_{\text{accept}} = \exp(\beta(\Delta U(x_s) - T))$. The most efficient samplers use $T = \max_x \Delta U(x)$. For samplers with $T > \max_x \Delta U(x)$, the probability distribution (4) is still recovered, but more samples are required, as the acceptance probability p_{accept} is decreased in this case. This strategy is a particular version of the rejection sampling algorithm and is shown in pseudo-code in Algorithm 1. We can see the direct connection between informational resources ("distance away from the prior") and the average number of samples required until acceptance, as the expected number of required samples from p_0 to obtain one accepted sample from p_1^{eq} is given by $\bar{n}_\beta = \exp(\beta T)/Z_\beta \geq \exp D_{\text{KL}}(p||p_0)$ [43]. In the limit of zero information-processing with $D_{\text{KL}}(p||p_0) = 0$ in the high-cost regime $\beta \to 0$, the sampling complexity tends to its minimum $\bar{n}_{\beta \to 0} \to 1$.

Algorithm 1 Rejection sampling.

repeat
  $x \sim p_0(x)$
  $u \sim \text{Uniform}[0,1]$
  if $u \leq \exp(\beta(\Delta U(x) - T))$ **then** accept
until accept
return x

In case we do not want to set an absolute aspiration level T, an incremental version of such a decision-maker can be realized by the Metropolis–Hastings scheme. Given a current action proposal x, the decision-maker generates a novel proposal x' from $p_0(x)$. If $\Delta U(x') \geq \Delta U(x)$, then the sample is accepted with certainty. An inferior sample is accepted with probability $p_{\text{accept}} = \exp(\beta(\Delta U(x') - \Delta U(x))$. The aspiration level in this case is variable and always given by the utility of the previous sample. This corresponds to a Markov chain with transition probability $p(x'|x) = p_0(x') \min\{1, \exp(\beta(\Delta U(x') - \Delta U(x)))\}$ and stationary distribution $p_1^{\text{eq}}(x)$. This Markov chain fulfils detailed balance, i.e., $p_1^{\text{eq}}(x)p(x'|x) = p_1^{\text{eq}}(x')p(x|x')$, which implies that after infinitely many repetitions, the samples x will follow the stationary distribution. This Markov chain is a particular version of the Metropolis–Hastings algorithm and is shown in pseudo-code in Algorithm 2. The longer the chain runs, the further the distribution of x will move away from the prior, i.e., the higher the informational resources will be. Finally, the chain reaches the equilibrium distribution.

Algorithm 2 Metropolis–Hastings sampling.

$x \sim p_0(x)$
repeat
  $x' \sim p_0(x')$
  $u \sim \text{Uniform}[0,1]$
  if $u \leq \exp(\beta(\Delta U(x') - \Delta U(x)))$ **then** accept $x \leftarrow x'$
until chain has converged to equilibrium
return x

3. Non-Equilibrium Thermodynamics and Decision-Making

If decision-making is emulated by a Markov chain that converges to an equilibrium distribution and one wants to be absolutely certain that the chain has reached equilibrium, then one has to wait for an infinitely long time. For finite times, when considering only a limited number of samples from the chain, we are dealing in general with non-equilibrium any time process models, i.e., computational processes that can be interrupted at any time to deliver an answer; a representative example being the Metropolis–Hastings dynamics when Algorithm 2 is run for $k \in \mathbb{N}$ steps. The same holds true for a rejection sampling decision-maker. Even though Algorithm 1 generates equilibrium samples with a finite expected number of samples $\bar{n}_\beta$, before running the algorithm, it is unknown whether after a particular number of steps k, a sample will be accepted or not; to have certainty, we would have to allow for an infinite amount of time ($k \to \infty$). In an any time version of rejection sampling, the probability of not accepting a sample after k tries is given by $q_k = [1 - Z(\beta)\exp(-\beta T)]^k$, in which case the sample x_s will be distributed according to the prior distribution $p_0(x)$. The probability of accepting a sample that is distributed according to $p_1^{\text{eq}}(x)$ after k tries is given by $1 - q_k$. Accordingly, the action at time k is a mixture distribution of the form:

$$p_k^{\text{neq}}(x) = (1 - q_k)p_1^{\text{eq}}(x) + q_k p_0(x). \tag{6}$$

The distribution $p_k^{\text{neq}}(x)$ is a non-equilibrium distribution that reaches equilibrium $p_k^{\text{neq}}(x) \to p_1^{\text{eq}}(x)$ for $k \to \infty$. In the following, we ask how far the tools of non-equilibrium thermodynamics are applicable to such any time decision-making processes.

3.1. Non-Equilibrium Thermodynamics

In thermodynamics, non-equilibrium processes are often modelled in the presence of an external parameter $\lambda(t) \in [0, 1]$ that determines how the energy function $E_\lambda(x)$ changes over time; for example, when switching on a potential in a linear fashion, the energy would be $E_\lambda(x) = E_0(x) + \lambda\,(E_1(x) - E_0(x))$. When the change in the parameter λ is done infinitely slowly (quasi-statically), the system's probability distribution follows exactly the path of equilibrium distributions (for any λ) $p_\lambda(x) = \frac{1}{Z_\lambda} e^{-\beta E_\lambda(x)}$. Importantly, when the switching of the external parameter λ is done in finite time, the trajectory in phase space of the evolving thermodynamic system can potentially be very different from the quasi-static case. In particular, the non-equilibrium path of probability distributions is going to be, in general, different from the equilibrium path. We define the trajectory of an evolving system as a finite sequence of states $\mathbf{x} := (x_0, x_1, \ldots x_N)$ at times $t_0, t_1, \ldots, t_N$, and the probability of the trajectory as $p(\mathbf{x}) := p(x_0|t_0)\prod_{n=1}^N p(x_n|x_{n-1}, t_n)$ that follows Markovian dynamics. Since λ is then a function of time $\lambda(t_n)$, we can effectively consider the energy as a function of state and time $E(x_n, t_n) := E_{\lambda(t_n)}(x_n)$. Accordingly, the internal energy of the system can change in two ways depending on changes in the two variables t_n and x_n. Assuming discrete time steps, an energy change due to a change in the external parameter is defined as the work [24,44]:

$$w(x_{n-1}, t_{n-1} \to t_n) = E(x_{n-1}, t_n) - E(x_{n-1}, t_{n-1})$$

and an energy change due to an internal state change is defined as the heat [24,44]:

$$q(x_{n-1} \to x_n, t_n) = E(x_n, t_n) - E(x_{n-1}, t_n).$$

For an entire process trajectory $x_0, x_1, \ldots, x_N$ measured at times $t_0, t_1, \ldots, t_N$, the extracted work is $W(\mathbf{x}) = -\sum_{n=1}^N w(x_{n-1}, t_{n-1} \to t_n)$, and the heat transferred to the environment by relaxation steps is $Q(\mathbf{x}) = -\sum_{n=1}^N q(x_{n-1} \to x_n, t_n)$. The sum of work and heat is the total energy difference $\Delta E(\mathbf{x}) := -(E(x_N, t_N) - E(x_0, t_0)) = W(\mathbf{x}) + Q(\mathbf{x})$. In expectation with respect to $p(\mathbf{x})$, we define the average work $W := \langle W(\mathbf{x})\rangle_{p(\mathbf{x})}$, the average heat $Q := \langle Q(\mathbf{x})\rangle_{p(\mathbf{x})}$ and the average energy change

$\Delta E := \langle \Delta E(\mathbf{x}) \rangle_{p(\mathbf{x})}$. With these averaged quantities, we obtain the first law of thermodynamics in its usual form:

$$\begin{aligned} \Delta E &= W + Q \\ &= W + T\Delta S + W^{\text{diss}}. \end{aligned} \tag{7}$$

The heat Q can be decomposed into a reversible and an irreversible part given by the entropy difference $\Delta S = -(S(t_N) - S(t_0))$, which is multiplied by the temperature T and the average dissipation W^{diss}. The concept of dissipation will be particularly useful later to quantify inefficacies in decision-making processes with limited time. By identifying the equilibrium free energy difference with $\Delta F := -(F(t_N) - F(t_0)) = \Delta E - T\Delta S$, we can then write the first law as:

$$W = \Delta F - W^{\text{diss}}. \tag{8}$$

In case of a quasi-static process, the extracted work W exactly coincides with the equilibrium free energy difference (thus, $W^{\text{diss}} = 0$). In the case of a finite time process, we can express the average dissipated work as [45–47]:

$$W^{\text{diss}} := \left\langle W^{\text{diss}}(\mathbf{x}) \right\rangle_{p(\mathbf{x})} = \Delta F - W = \frac{1}{\beta} D_{\text{KL}} \left(p(\mathbf{x}) || p^{\dagger}(\mathbf{x}) \right) \tag{9}$$

where D_{KL} is the relative entropy that measures in bits the distinguishability between the probability of the forward in time trajectory $p(\mathbf{x})$ and the probability of the backward in time trajectory $p^{\dagger}(\mathbf{x}) := p(x_N|t_N) \prod_{n=1}^{N} p(x_{n-1}|x_n, t_{n-1})$. From the positivity of the relative entropy, we can immediately see the non-negativity of entropy production $W^{\text{diss}} \geq 0$, which allows stating the second law of thermodynamics in the form:

$$W \leq \Delta F. \tag{10}$$

3.1.1. Crooks' Fluctuation Theorem

Equation (9) can be given in a more general form without averages. It is possible to relate the reversibility of a process with its dissipation at the trajectory level. Given a protocol $\Lambda = (\lambda_0, \lambda_1, \ldots \lambda_N)$, i.e., a sequence of external parameters, the probability $p(\mathbf{x})$ of observing a trajectory of the system in phase space compared with its time-reversal conjugate $p^{\dagger}(\mathbf{x})$ (when using the time-reversal protocol $\Lambda^{\dagger} = (\lambda_N, \lambda_{N-1}, \ldots \lambda_0)$) depends on the dissipation of the trajectory in the forward direction according to the following expression:

$$\frac{p(\mathbf{x})}{p^{\dagger}(\mathbf{x})} = e^{\beta W^{\text{diss}}(\mathbf{x})},$$

where $W^{\text{diss}}(\mathbf{x}) = \Delta F - W(\mathbf{x})$ is the dissipated work of the trajectory. For this relation to be true, both backward and forward processes must start with the system in equilibrium. Intuitively, this means that the more the entropy production (measured by the dissipated work), the more distinguishable are the trajectories of the forward protocol compared to the backward protocol.

3.1.2. Jarzynski Equality

Additionally, another relation of interest in non-equilibrium thermodynamics has recently been found transforming the inequality of Equation (10) into an equality, the so-called Jarzynski equality [48]:

$$\left\langle e^{\beta W(\mathbf{x})} \right\rangle_{p(\mathbf{x})} = e^{\beta \Delta F} \tag{11}$$

where the angle brackets denote an average over all possible trajectories $\mathbf{x}$ of a process that drives the system from an equilibrium state at $\lambda = 0$ to another state at $\lambda = 1$. Specifically, the above equality says that, no matter how the driving process is implemented, we can determine equilibrium quantities from work fluctuations in the non-equilibrium process; or in other words, this equality connects non-equilibrium thermodynamics with equilibrium thermodynamics. In the following, we are interested in the question whether there exist similar relations such as the Jarzynski equality or Crooks' fluctuation theorem and similar underlying concepts such as dissipation and time reversibility for the case of decision-making.

3.2. Non-Equilibrium Thermodynamics Applied to Bounded Rational Decision-Making

In direct analogy to the previous section, in the following, we consider decision-makers faced with the problem of optimizing a changing utility function. We assume that time is discretized into N steps $t_0, \ldots, t_N$. For each time step t_n, the utility is assumed to be constant, but it can change between time steps, such that we have a sequence of decision problems expressed by the changes in utility $\Delta U(x, t_0 \to t_1), \ldots, \Delta U(x, t_{N-1} \to t_N)$. At each time point t_n, the decision-maker chooses action x_n, such that we can summarize the decision-maker's choices by a vector $\mathbf{x} := (x_0, \ldots, x_N)$. The behaviour of the decision-maker is characterized by the probability $p(\mathbf{x}) := p(x_0|t_0) \prod_{n=1}^{N} p(x_n|x_{n-1}, t_n)$ with $p(x_0|t_0) = p_0(x_0)$, assuming that the initial strategy is a bounded rational equilibrium strategy. In this setup, we assume that the changes in the utility function are externally driven, i.e., the decision-maker's actions cannot change the temporal evolution of the utility function. Furthermore, note that the decision-maker does not know how the utility changes over time. Accordingly, the best the decision-maker can do is to optimize the current utility as much as possible.

At time t_0, the decision-maker starts with selecting an action x_0 from the distribution $p(x_0|t_0)$ and the utility changes instantly by $\Delta U(x, t_0 \to t_1)$. The decision-maker can then adapt to this utility change with the distribution $p(x_1|x_0, t_1)$ and select the action x_1 at time t_1, but at this point, the utility is already changing again by $\Delta U(x, t_1 \to t_2)$. The adaptation from $p(x_0|t_0)$ to $p(x_1|x_0, t_1)$ is analogous to a physical relaxation process and implies a strategy change between x_0 and x_1. In general, at each time point t_{n-1}, the decision-maker chooses action x_{n-1} while the current utility changes by:

$$\Delta U(x_{n-1}, t_{n-1} \to t_n) = U(x_{n-1}, t_n) - U(x_{n-1}, t_{n-1}).$$

This way, the decision-maker is always lagging behind the changes in utility, just like a physical system would lag behind the changes in the energy function. The utility $\Delta U(x_{n-1}, t_{n-1} \to t_n)$ gained by the decision-maker at time point t_{n-1} parallels the concept of work in physics. For a whole trajectory, we define the total utility gain due to changes in the environment as $\mathcal{U}(\mathbf{x}) = \sum_{n=1}^{N} \Delta U(x_{n-1}, t_{n-1} \to t_n)$. Note that the last decision x_N can be ignored in this notation, as it does not contribute to the utility.

In Figure 1 (left column), we illustrate the setup for a one-step decision problem $\Delta U(x, t_0 \to t_1)$ with behaviour vector $\mathbf{x} = (x_0, x_1)$. An instantaneous change in the environment occurs at time t_0 represented by a vertical jump from λ_0 to λ_1 in the upper panels that translates directly into a change in free energy difference represented by ΔF in the lower panels. The system's previous state at t_0 is given by $p_0^{\text{eq}}(x)$, i.e., the equilibrium distribution for U_0. The new equilibrium is given by $p_1^{\text{eq}}(x)$, i.e., the equilibrium distribution for U_1. In this case, the behaviour vector is $\mathbf{x} = (x_0, x_1)$ with $x_0 \sim p_0^{\text{eq}}(x)$, and x_1 is ignored.

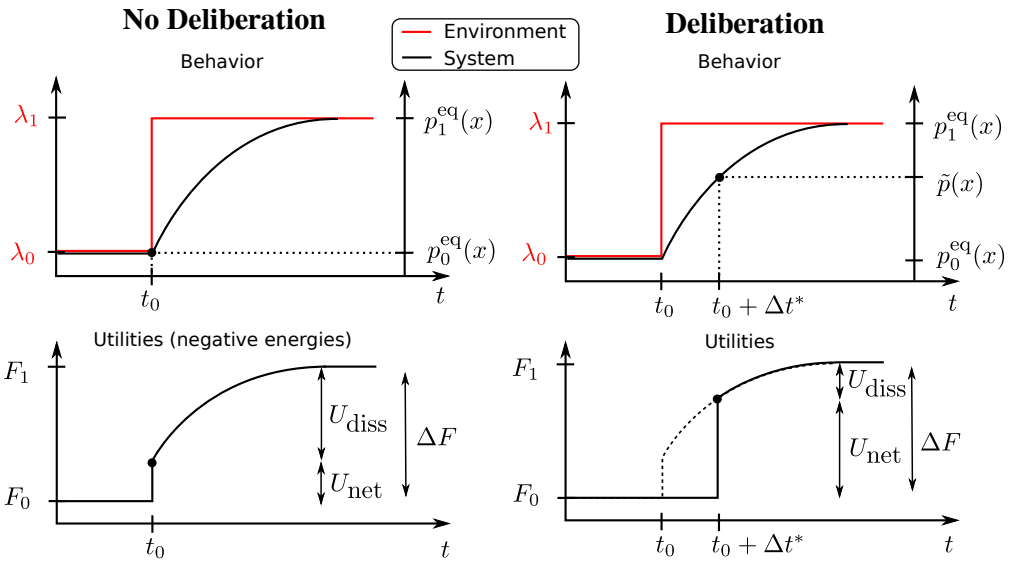


Figure 1. Temporal structure of the one-step decision problem. An instantaneous change in the environment occurs at time t_0 represented by a vertical jump from λ_0 to λ_1 in the upper panels that translates directly into a change in free energy difference represented by ΔF in the lower panels. The system's previous state at t_0 is given by $p_0^{\mathrm{eq}}(x)$, i.e., the equilibrium distribution for $U_{\lambda_0}(x)$. The new posterior equilibrium is given by $p_1^{\mathrm{eq}}(x)$, i.e., the equilibrium distribution for $U_{\lambda_1}(x)$. When given unlimited time, the decision-maker will eventually evolve to $p_1^{\mathrm{eq}}(x)$. Deliberative and non-deliberative decision-makers differ in how much time they get to adapt to the change in utility before they have to choose an action x that provides them with the utility gain $\Delta U(x) = U_{\lambda_1}(x) - U_{\lambda_0}(x)$. Left: In direct analogy to physical thermodynamics, the non-deliberative decision-maker has to emit an action before it can adapt to any changes in utility and therefore acts according to the previous strategy $p_0^{\mathrm{eq}}(x)$ at time t_0. On average, with such a strategy, the utility gained is $\mathcal{U}^{\mathrm{net}} = \sum_x p_0^{\mathrm{eq}}(x)\Delta U(x)$ at t_0 and the dissipation is $\mathcal{U}^{\mathrm{diss}} = \Delta F - \mathcal{U}^{\mathrm{net}}$. Right: The deliberative decision-maker is allowed to adapt to the change in utility for a certain time Δt^* before the action has to be emitted. This deliberation period allows the decision-maker to compute a better strategy $\tilde{p}(x)$. In this case, the net utility is $\mathcal{U}^{\mathrm{net}} = \sum_x \tilde{p}(x)\Delta U - \frac{1}{\beta} D_{\mathrm{KL}}\left(\tilde{p}(x)||p_0^{\mathrm{eq}}(x)\right)$.

Similarly to Equation (8), we can now formulate the first law for decision-making as:

$$\mathcal{U} = \Delta F - \mathcal{U}^{\mathrm{diss}}$$

stating that the total average utility $\mathcal{U} := \langle \mathcal{U}(\mathbf{x}) \rangle_{p(\mathbf{x})}$ is the difference between the bounded optimal utility (following the equilibrium strategy with precision β) expressed by the equilibrium free energy difference ΔF and the dissipated utility $\mathcal{U}^{\mathrm{diss}}$. The dissipation for a trajectory $\mathcal{U}^{\mathrm{diss}}(\mathbf{x}) := \Delta F - \mathcal{U}(\mathbf{x})$ measures the amount of utility loss due to the inability of the decision-maker to act according to the equilibrium distribution. This is because the decision-maker cannot anticipate the changes in the environment. At most, the decision-maker could act according to the equilibrium distributions of the previous environment. Thus, even with full adaptation, the decision-maker will always lag behind one time step and will therefore always dissipate.

Due to an equivalent version of Equation (9), we can also state the second law for decision-making $\mathcal{U}^{\mathrm{diss}} \geq 0$, which implies that a purely adaptive decision-maker can gain a maximum utility that cannot be larger than the free energy difference:

$$\mathcal{U} \leq \Delta F.$$

Similarly, we can obtain equivalent relationships to the Crooks fluctuation theorem:

$$\frac{p(\mathbf{x})}{p^{\dagger}(\mathbf{x})} = e^{\beta \mathcal{U}^{\text{diss}}(\mathbf{x})}, \tag{12}$$

and the Jarzynski equality:

$$\left\langle e^{\beta \mathcal{U}(\mathbf{x})} \right\rangle_{p(\mathbf{x})} = e^{\beta \Delta F} \tag{13}$$

which both have the same implications as in the physical scenario and can be derived in the same way as in the physical counterpart [44]. In summary, we can say that an adaptive decision-maker, which has to act without knowing that the utility function has changed, follows the same laws as a thermodynamic physical system that is lagging behind the equilibrium.

3.3. Examples

In this section, we illustrate the applicability of thermodynamic non-equilibrium concepts in a series of simulations for different decision-making scenarios. In particular, we study two model classes: the first one contains simple one-step lag models of adaptation where equilibrium is always reached with one time step delay, and the second one contains more complex models of adaptation that do not necessarily equilibrate after one time step. In the first model class, we can easily study the relation between dissipation and the rate of information-processing, whereas in the second class of models, we can study more complex non-equilibrium phenomena such as learning hysteresis.

3.3.1. One-Step Lag Models of Adaptation

Consider a learner that is adapted to their environment such that their behaviour can be described by the equilibrium distribution $p_0(x)$. For this idealized scenario, we assume that the learner can adapt their behaviour to any environment perfectly after a time lapse of Δt. This also means that before the lapse of Δt, the learner continues to follow their old strategy and is inefficient during this time span. We now consider two scenarios: first, where the environment changes suddenly by $\Delta U(x)$, and second, where the environment changes slowly in N small steps of $\Delta U(x)/N$. In the first case, the learner is going to dissipate the utility:

$$\mathcal{U}^{\text{diss}} = \frac{1}{\beta} D_{\text{KL}} \left(p_0(x) || p_1^{\text{eq}}(x) \right),$$

in the first time step. In all subsequent time steps, no more utility is wasted, assuming the environment does not change any more. In the second case, the utility function can be written as $U_t(x) = U_0(x) + \frac{t}{N} \Delta U(x)$ for $t \in \mathbb{N} : 0 \leq t \leq N$. To compute the dissipated utility, we need to compare the learner's behaviour in time step t to the bounded optimal behaviour, which is:

$$p^{\text{eq}}(x,t) = \frac{1}{Z} p^{\text{eq}}(x,t-1) e^{\frac{\beta}{N} \Delta U(x)}$$

for $t > 0$. The overall average dissipated utility for the whole process is then

$$\mathcal{U}_N^{\text{diss}} = \frac{1}{\beta} \sum_{t=1}^{N} D_{\text{KL}} \left(p^{\text{eq}}(x,t-1) || p^{\text{eq}}(x,t) \right).$$

The net utility gain for the N-step scenario is $\mathcal{U}_N^{\text{net}} = \Delta F - \mathcal{U}_N^{\text{diss}}$. Note that:

$$\mathcal{U}_N^{\text{diss}} \geq \mathcal{U}_{N+1}^{\text{diss}}$$

and consequently, in direct analogy to a quasi-static change in a thermodynamic system, we get vanishing dissipation ($\mathcal{U}_N^{\text{diss}} \to 0$) if the utility changes infinitely slowly ($N \to \infty$ and $\Delta U(x)/N \to 0$), such that the net utility equals the free energy difference $\mathcal{U}_N^{\text{net}} = \Delta F$.

3.3.2. Bayesian Inference as a One-Step Lag Process

Bayesian inference mechanisms naturally have step by step dynamics that update beliefs with new incoming observations. Again, we can consider two scenarios: first where the learner updates their belief abruptly by processing a huge chunk of data in one go, and second, where belief updates are incremental with small chunks of data at each time step. Here, we show how the size of the chunks of data affect the overall surprise of the decision-maker and how this relates to dissipation applying the free energy principle to Bayesian inference.

Traditionally, Bayes' rule is obtained directly from the product rule of probabilities $p(\theta, \mathcal{D}) = p(\theta)p(\mathcal{D}|\theta) = p(\mathcal{D})p(\theta|\mathcal{D})$ where θ correspond to the different available hypotheses and $\mathcal{D}$ corresponds to the dataset. However, Bayes' rule can also be considered to be a consequence of the maximization of the free energy difference with the log-likelihood as a utility function [49–51]. In this view, the posterior belief $p(\theta|\mathcal{D})$ is a trade-off between maximizing the likelihood $p(\mathcal{D}|\theta)$ and minimizing the distance from the prior $p_0(\theta)$ such that:

$$p(\theta|\mathcal{D}) = \underset{\tilde{p}}{\operatorname{argmax}}\, \Delta F[\tilde{p}] = \underset{\tilde{p}}{\operatorname{argmax}} \int \tilde{p}(\theta|\mathcal{D}) \log p(\mathcal{D}|\theta) d\theta - \frac{1}{\beta} \int \tilde{p}(\theta|\mathcal{D}) \log \frac{\tilde{p}(\theta|\mathcal{D})}{p_0(\theta)} d\theta \tag{14}$$

$$= \frac{1}{Z} p_0(\theta) e^{\beta \log p(\mathcal{D}|\theta)} = \frac{1}{Z} p_0(\theta) p(\mathcal{D}|\theta)^\beta \tag{15}$$

is identical to Bayes' rule when $\beta = 1$. For $\beta \to \infty$, we recover the maximum likelihood estimation method as the density update is $p(\theta|\mathcal{D}) = \delta(\theta - \theta_{\mathrm{MLE}})$ with $\theta_{\mathrm{MLE}} = \operatorname{argmax}_\theta \log p(\mathcal{D}|\theta)$.

Such a Bayesian learner with prior $p_0(\theta)$ that incorporates all the data X at once is going to experience the expected surprise $\mathcal{S} = -\int p_0(\theta) \log p(\mathcal{D}|\theta) d\theta$. In contrast, a Bayesian learner that incorporates the data slowly in N steps (thus, the dataset $\mathcal{D} = (X_1, \ldots, X_N)$ is divided in N parts) experiences an expected surprise of $\mathcal{S} = -\sum_{n=1}^{N} \int p(\theta|X_1, \ldots, X_{n-1}) \log p(X_n|\theta) d\theta$. Here, the surprise $\mathcal{S}$ corresponds to the thermodynamic concept of work. The first law can then be written as:

$$\Delta F + \mathcal{S} = \mathcal{U}^{\mathrm{diss}}$$

where the equivalent of dissipation corresponds to:

$$\mathcal{U}^{\mathrm{diss}} = \frac{1}{\beta} D_{\mathrm{KL}}(p_0(\theta) || p_{\mathrm{eq}}(\theta|\mathcal{D})).$$

when processing all the data at once and to:

$$\mathcal{U}^{\mathrm{diss}} = \frac{1}{\beta} \sum_{n=1}^{N} D_{\mathrm{KL}}(p(\theta|X_{<n}) || p_{\mathrm{eq}}(\theta|X_{\leq n})).$$

when processing the data in N steps where $X_{<n} = (X_1, \ldots, X_{n-1})$ and $X_{\leq n} = (X_1, \ldots, X_n)$. Thus, given that the equilibrium free-energy difference ΔF is a state function independent of the path (that means independent of whether data are processed all in one go or in small chunks), a system acquiring data slowly will have a reduced surprise $\mathcal{S}$ and therefore have less dissipation $\mathcal{U}^{\mathrm{diss}}$.

In Figure 2, we show how the number of data chunks has an effect on the overall surprise and dissipation. In particular, we have a dataset $\mathcal{D} = (x_1, \ldots, x_T)$ consisting of $T = 100$ data points Gaussian distributed $x \sim \mathcal{N}(x; \mu_d = 5, \sigma_d^2 = 4)$ that we divide into batches of different sizes $b \in \{100, 50, 25, 20, 10, 5, 2, 1\}$. The decision-maker has prior belief $p_0(\theta)$ about the mean $\theta = \mu_d$ and incorporates the data of every batch of data according to Bayes' rule until all the data are incorporated. In general, the Bayesian learner processes the data in T/b steps; for example in the case of $b = 100$, all data are processed at once (having thus high surprise), and in the case of $b = 1$, it incorporates the data in T updates with an overall smaller surprise. In Figure 2, we show for different batch sizes

the free energy optimum $\Delta F = \log \int p_0(\theta) p(\mathcal{D}|\theta)$, the surprise $\mathcal{S}$ and the dissipation $\mathcal{U}^{\text{diss}} = \Delta F - \mathcal{S}$. It can be seen that when acquiring the data in small chunks, the surprise of the decision-maker and the dissipation are lower.

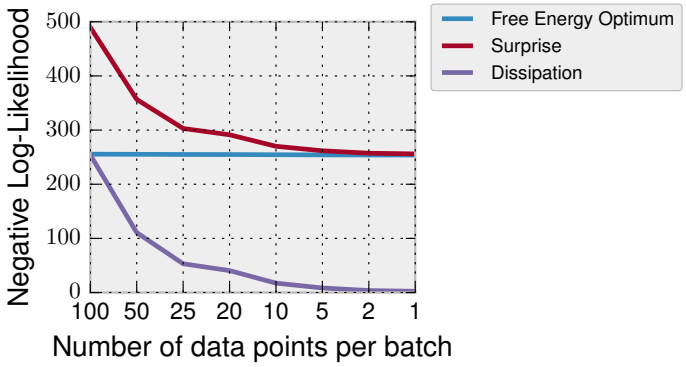


Figure 2. Surprise, dissipation and free energy optimum as a function of the number of data points per batch in a Bayesian inference task. When the decision-maker processes all the data in one step, it has maximum surprise and dissipation. However, when incorporating the data slowly, the surprise and dissipation are humble. The free energy optimum is only a function of the data independent of how they are incorporated.

3.4. Dissipation and Learning Hysteresis

A common paradigm to study how humans learn is through adaptation tasks where subjects are exposed to changes in an environmental variable that they can counteract by changing an internal variable. Sensorimotor adaptation in humans has been extensively studied in these error-based paradigms, for example where subjects have to adapt their hand position (internal variable) to change a virtual end effector position represented by a dot on a screen (external variable).

Consider a utility function $U_v(x) = -(x - \mu_v)^2$. For $v = 0$, we determine the prior behaviour of a decision-maker with $p_0(x) = \frac{e^{\beta U_0(x)}}{Z}$. Initially, the decision-maker obtains an average utility of $\langle U_0 \rangle_{p_0}$, which corresponds to zero mismatch between the decision-maker and the environmental variable. A change of the environmental variable to $v = 1$ effectively changes the utility function to $U_1(x) = -(x - \mu_1)^2$, making p_0 non-optimal. This forces the decision-maker to reduce error adapting to the environmental variable by changing its probability distribution over his/her actions. When fully adapted to the new environment, the decision-maker again makes no errors (other than the errors due to motor noise). We illustrate this adaptation paradigm with a decision-maker that adapts according to the Metropolis–Hastings algorithm, which follows Markovian dynamics [52].

Crooks Theorem and Hysteresis Effects in Adaptation Tasks

Limited adaptation capabilities not only have an effect on the amount of obtained utility through the second law for decision-making $\mathcal{U}^{\text{net}} \leq \Delta F$, but also induce a time asymmetry in sequential decision-making processes. Hysteresis loops are a typical example of this asymmetry. Hysteresis is the phenomenon in which the path followed by a system due to an external perturbation, e.g., from state A to B, is not the same as the path followed in the reverse perturbation, e.g., from state B to A. When the system follows the same path for the forward perturbation and for the reverse perturbation, we say that the process is time symmetric (and therefore, it is not subject to hysteresis effects).

In the two left panels of Figure 3, we show a simulated trajectory of actions composed of 80 trials for an adaptation task using the Metropolis–Hastings algorithm with $\beta = 22.5$, a Gaussian proposal $g(x'|x) = \mathcal{N}(x'; \mu = x, \sigma_p = 0.1)$ and acceptance criterion $\alpha(x'|x) = \min\left(\frac{e^{\beta U(x')} g(x|x')}{e^{\beta U(x)} g(x'|x)}, 1\right)$, when changing the environmental variable from $\mu_0 = 0.0$ to $\mu_1 = 1.0$. In blue, we show the trajectory for

the forward-in-time perturbation, which converges after a few dozen trials to the new equilibrium. In brown, we show the trajectory for the reversed perturbation where the process starts with the last trial (80) and ends with the initial trial (0). In the left panel, the perturbation is made instantaneously in one step at Trial 40 and in the right panel in multiple steps ($N = 23$). The hysteresis effect is clearly seen in the instantaneous perturbation where the path of actions followed by the decision-maker in the forward perturbation is clearly different from a typical trajectory of actions taken when applying the reversed perturbation. When the perturbation is made in multiple steps, both typical backward and typical forward trajectories become more similar denoting a smaller hysteresis effect. In this way, hysteresis effects are tightly connected to the concept of dissipation.

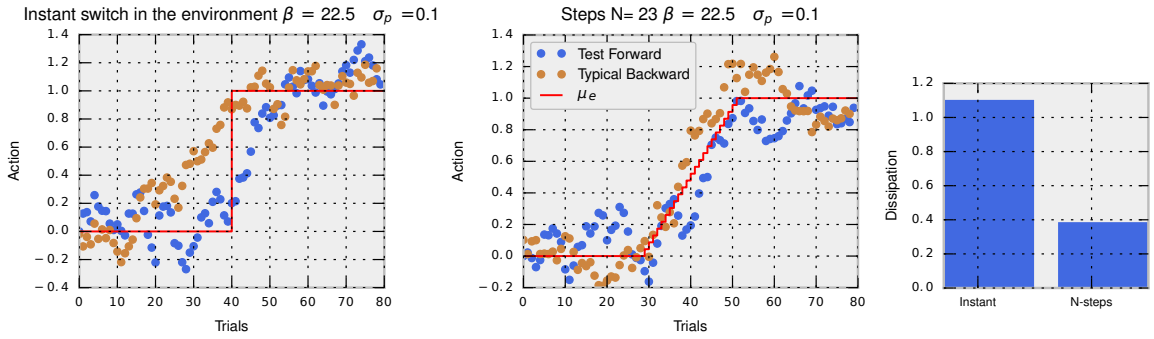


Figure 3. Trajectories of actions from the Metropolis–Hastings algorithm with $\beta = 22.5$ and the proposed standard deviation $\sigma_p = 0.1$ in a forward (blue) or backward (brown) protocol for an instant change in the environment (first panel) and for a slow change in the environment (second panel). In both cases, the total change in the environment is $\mu_0 = 0$ to $\mu_1 = 1$. The last panels shows the dissipation for the forward protocol (blue) in both the instant or the slow change in the environment. The difference in probability densities of forward and backward trajectories relates directly to dissipation and to hysteresis effects.

Dissipation and the ratio between forward and backward probabilities of trajectories of actions correspond exactly to the Crooks theorem for decision-making:

$$\frac{p(\mathbf{x})}{p^{\dagger}(\mathbf{x})} = e^{\beta \mathcal{U}^{\mathrm{diss}}(\mathbf{x})}.$$

The probability of observing a trajectory of accepted actions $\mathbf{x} = (x_0, x_1, \ldots x_T)$ for the Metropolis–Hastings algorithm is easily computed with $p(\mathbf{x}) = p(x_0)\prod_{t=1}^{T} g(x_t|x_{t-1})\alpha(x_t|x_{t-1})$. Similarly, the probability of observing the same trajectory in the backward protocol is $p(\mathbf{x}^{\dagger}) = p_{\mathrm{eq}}(x_T)\prod_{t=1}^{T} g(x_{T-t}|x_{T-t+1})\alpha(x_{T-t}|x_{T-t+1})$. The dissipated utility is $\mathcal{U}^{\mathrm{diss}} = \Delta F - U_{\mathrm{tot}}$ where the free energy difference is computed between the final $p_1(x) = \frac{1}{Z}e^{\beta U_1(x)}$ and initial equilibrium distributions $p_0(x) = \frac{1}{Z}e^{\beta U_0(x)}$, and the total utility gained U_{tot} is the sum of the utilities $\Delta U(x, t_n \to t_{n+1})$ at each environmental change at time t_n. In the third panel of Figure 3, we show that the protocol with the instantaneous perturbation has higher dissipation (related to higher hysteresis) compared to the protocol with multiple small perturbations.

4. Generalized Non-Equilibrium Thermodynamics for Decision-Making with Deliberation

So far, we have studied decision-makers that were forced to select an action with no opportunity to respond to a change in the utility function. This could correspond, for example, to a scenario of trial-and-error learning, where the best available strategy is the prior strategy adapted to the environment before the utility changed. However, this restriction may not always be suitable. Consider for example a chess player that is shown a particular board configuration (corresponding to a change in utility) and now has a certain amount of time to decide on the next move. Similarly, consider the

two introductory examples in Section 3, where we allow a sampling algorithm to run for a certain number of steps, and then, we stop and evaluate the action after the algorithm has adapted to the new utility. In general, such deliberation processes are expensive, and we assume in the following that the Kullback–Leibler divergence is an appropriate measure of this computational expense, as outlined in the Introduction.

In the following, we consider again decision-makers facing a sequence of decision problems expressed by the utility changes $\Delta U(x, t_0 \to t_1), \ldots, \Delta U(x, t_{N-1} \to t_N)$. In contrast to the previous section where decision-makers had to decide before they could adapt to the utility change, decision-makers that deliberate select their action x_n after they have (partially) adapted to the utility change:

$$\Delta U(x_n, t_{n-1} \to t_n) = U(x_n, t_n) - U(x_n, t_{n-1}).$$

Using this notation, we are able to summarize the decision-maker's choice by a vector $\mathbf{x} := (x_0, \ldots, x_N)$ and characterize its behaviour by the probability $p(\mathbf{x}) := p(x_0|t_0) \prod_{n=1}^{N} p(x_n|x_{n-1}, t_n)$ with $p(x_0|t_0) = p_0(x_0)$, assuming that the initial strategy is a bounded rational equilibrium strategy. Note that in the deliberation scenario, the initial state x_0 does not constitute a decision, but instead, we include the last decision x_N.

This setup is illustrated again in Figure 1 (right column) for a one-step decision problem $\Delta U(x, t_0 \to t_1)$ with behaviour vector $\mathbf{x} = (x_0, x_1)$ and with an instantaneous change in the environment occurring at time t_0. In the deliberation scenario, the utility is determined after the deliberation time. During deliberation, the decision-maker has changed the strategy distribution from $p_0^{\mathrm{eq}}(x)$ to a non-equilibrium distribution $\tilde{p}(x)$ (for example, the distribution (6) in the rejection sampling scheme) spending in the process a certain amount of resources and achieving an average net utility of $\mathcal{U}^{\mathrm{net}} = \Delta F[\tilde{p}(x)]$ according to Equation (3). In this case, the behaviour vector is $\mathbf{x} = (x_0, x_1)$ with x_0 ignored and $x_1 \sim \tilde{p}(x)$. In such a scenario with a single decision problem, we define, in analogy with the previous section, the average dissipated utility as [24,53]:

$$\begin{aligned} \mathcal{U}^{\mathrm{diss}} &:= \Delta F - \mathcal{U}^{\mathrm{net}} \\ &= \frac{1}{\beta} D_{\mathrm{KL}} \left(\tilde{p}(x) || p_1^{\mathrm{eq}}(x) \right). \end{aligned} \tag{16}$$

See Appendix for a derivation of (16) from (9). It readily follows from the positivity of the relative entropy $D_{\mathrm{KL}}(p||q) \geq 0$ that:

$$\mathcal{U}^{\mathrm{net}} \leq \Delta F \tag{17}$$

with equality when $\tilde{p}(x) = p_1^{\mathrm{eq}}(x)$. In the case of the rejection sampling decision-maker of Equation (6), this would correspond to an infinite amount of samples $k \to \infty$. The inequality (17) shows that we cannot obtain more utility than the equilibrium free energy difference.

Let us now look at the general case. In contrast to an agent without deliberation capabilities, an agent that deliberates will be able to act according to a different distribution than the prior strategy. This means that when facing the utility change $\Delta U(x, t_{n-1} \to t_n)$ at time t_n, the agent chooses the action x_n sampled from the posterior strategy, contrary to an agent without deliberation that chooses x_{n-1} sampled from the prior strategy. The deliberation process incurs a computational cost that is measured (in a similar fashion to stochastic thermodynamics [54] and previous formulations of bounded rationality given in the introduction) with the difference between the conditional stochastic entropies from prior to posterior:

$$s(x_n|x_{n-1}, t_n) - s(x_n|x_{n-1}, t_{n-1}) := -\log \frac{p(x_n|x_{n-1}, t_n)}{p(x_n|x_{n-1}, t_{n-1})}.$$

Note that the prior distribution $p(x_n|x_{n-1}, t_{n-1})$ is the previous posterior distribution evaluated at x_n instead of x_{n-1}. Basically, this measures the change in probability from prior behaviour to posterior behaviour of the newly chosen action x_n.

Taking into account the computational cost of deliberation, we define the net utility of action x_n due to a change in the environment as

$$u(x_n, t_{n-1} \to t_n) = \Delta U(x_n, t_{n-1} \to t_n) - \frac{1}{\beta} \log \frac{p(x_n|x_{n-1}, t_n)}{p(x_n|x_{n-1}, t_{n-1})},$$

which generalizes the concept of work from the previous section. The expected change in net utility is the objective function that the decision-maker optimizes at each time step. The total net utility $\mathcal{U}^{\text{net}}(\mathbf{x}) = \sum_{n=1}^{N} u(x_n, t_{n-1} \to t_n)$ takes the form of a non-equilibrium free energy:

$$\mathcal{U}^{\text{net}}(\mathbf{x}) = \sum_{n=1}^{N} \Delta U(x_n, t_{n-1} \to t_n) - \frac{1}{\beta} \sum_{n=1}^{N} \log \frac{p(x_n|x_{n-1}, t_n)}{p(x_n|x_{n-1}, t_{n-1})}. \tag{18}$$

at the trajectory level. Similarly to Equation (8), the first law for decision-making with deliberation costs is:

$$\mathcal{U}^{\text{net}} = \Delta F - \mathcal{U}^{\text{diss}}$$

and states that the total net utility $\mathcal{U}^{\text{net}} = \langle \mathcal{U}^{\text{net}}(\mathbf{x}) \rangle_{p(\mathbf{x})}$ is the difference between the bounded optimal utility (following the equilibrium strategy with precision β) expressed by the equilibrium free energy difference ΔF and the dissipated utility $\mathcal{U}^{\text{diss}}$. The dissipation:

$$\mathcal{U}^{\text{diss}}(\mathbf{x}) := \Delta F - \mathcal{U}^{\text{net}}(\mathbf{x}) \tag{19}$$

measures the amount of utility loss if the decision-maker's plan does not manage to produce an action from the equilibrium distribution, for example due to the lack of time for deliberation. However, a decision-maker with infinite deliberation time will not have this problem and therefore will not dissipate by wasting utility.

To investigate the counterpart of the second law, we need to determine whether $\mathcal{U}^{\text{diss}} \geq 0$ holds. This can be achieved, for example, by first deriving the counterpart of the Crooks fluctuation theorem or the counterpart of the Jarzynski equation with subsequent application of Jensen's inequality. In the following two theorems, we assume that the decision-makers satisfy the detailed balance condition. The detailed balance condition ensures two important characteristics. First, the stochastic process reaches equilibrium, and second, it ensures time-reversibility when in equilibrium. In a decision-making scenario, this translates into the following. First, when given enough computation time, the decision-makers manage to sample actions from the correct equilibrium distributions. Second, ideal decision-makers in equilibrium should not produce any entropy, which is exactly what happens if detailed balance is satisfied.

Theorem 1. *Crook'sfluctuation theorem for decision-making with deliberation costs states that:*

$$\frac{p(\mathbf{x})}{p^{\dagger}(\mathbf{x})} = e^{\beta \mathcal{U}^{diss}(\mathbf{x})} \tag{20}$$

where the dissipated utility of a particular trajectory is $\mathcal{U}^{diss}(\mathbf{x}) = \Delta F - \mathcal{U}^{net}(\mathbf{x})$ *as defined in Equation* (18) *and the probability of the trajectory using the backward protocol is* $p^{\dagger}(\mathbf{x}) = p^{\dagger}(x_0|x_1, t_0)\, p^{\dagger}(x_1|x_2, t_1) \cdots p^{\dagger}(x_N|t_N)$ *for N decision problems starting at time* t_N *and going backwards up to* t_0*. For the relation to be valid, we must assume that the starting distribution in the backward process is also in equilibrium,* $p(x_N|t_N) \propto e^{\beta U(x_N, t_N)}$.

Proof. Here, we derive the relationship between reversibility and dissipation.

$$\begin{aligned}\frac{p(\mathbf{x})}{p^{\dagger}(\mathbf{x})} &= \frac{p(x_0|t_0)p(x_1|x_0,t_1)\cdots p(x_N|x_{N-1},t_N)}{p^{\dagger}(x_0|x_1,t_0)p^{\dagger}(x_1|x_2,t_1)\cdots p^{\dagger}(x_N|t_N)} \\ &= \frac{e^{\beta U(x_0,t_0)}}{Z_0}\frac{1}{e^{\beta U(x_0,t_0)}}\frac{p(x_1|x_0,t_1)}{p(x_1|x_0,t_0)}\frac{e^{\beta U(x_1,t_0)}}{e^{\beta U(x_1,t_1)}}\cdots\frac{p(x_N|x_{N-1},t_N)}{p(x_N|x_{N-1},t_{N-1})}\frac{e^{\beta U(x_N,t_{N-1})}}{e^{\beta U(x_N,t_N)}}Z_N \\ &= \frac{Z_N}{Z_0}e^{\beta\frac{1}{\beta}\log\frac{p(x_1|x_0,t_1)}{p(x_1|x_0,t_0)}}e^{-\beta\Delta U(x_1,t_0\to t_1)}\cdots e^{\beta\frac{1}{\beta}\log\frac{p(x_N|x_{N-1},t_N)}{p(x_N|x_{N-1},t_{N-1})}}e^{-\beta\Delta U(x_N,t_{N-1}\to t_N)} \\ &= e^{\beta\Delta F-\beta\mathcal{U}^{\mathrm{net}}(\mathbf{x})} = e^{\beta\mathcal{U}^{\mathrm{diss}}(\mathrm{x})}\end{aligned}$$

where in the second line, we have substituted $p^{\dagger}(x_{n-1}|x_n,t_{n-1})$ using the identity:

$$p^{\dagger}(x_{n-1}|x_n,t_{n-1}) = \frac{e^{\beta U(x_{n-1},t_{n-1})}}{e^{\beta U(x_n,t_{n-1})}}p(x_n|x_{n-1},t_{n-1})$$

from detailed balance, and we assumed the initial distribution to be in equilibrium $p(x_0|t_0) = \frac{e^{\beta U(x_0,t_0)}}{Z_0}$ and that in the backward process the decision-maker starts also using the equilibrium strategy $p^{\dagger}(x_N|t_N) = \frac{1}{Z_N}e^{\beta U(x_N,t_N)}$. In the third line, we cancel out terms and apply the following two equalities $\frac{p(x_n|x_{n-1},t_n)}{p(x_n|x_{n-1},t_{n-1})} = e^{\beta\frac{1}{\beta}\log\frac{p(x_n|x_{n-1},t_n)}{p(x_n|x_{n-1},t_{n-1})}}$ and $\Delta U(x_n,t_{n-1}\to t_n) = U(x_n,t_n) - U(x_n,t_{n-1})$. Finally, in the last line, we employ the definition of the net utility in Equation (18) and $\frac{Z_N}{Z_0} = e^{\beta\Delta F}$. □

Although at first sight, Equation (20) looks the same as the previous Crooks' relation for the no-deliberation case (12), it is not the same. Here, the net utility is defined by Equation (18), which takes into account both the gain in utility and the computational costs of deliberating.

Theorem 2. *The Jarzynski equality for decision-making with deliberation costs states that:*

$$\left\langle e^{\beta\mathcal{U}^{net}(\mathbf{x})}\right\rangle_{p(\mathbf{x})} = e^{\beta\Delta F}. \tag{21}$$

Proof.

$$\begin{aligned}&\left\langle \exp\left(\beta\sum_{n=1}^{N}\left[\Delta U(x_n,t_{n-1}\to t_n) - \frac{1}{\beta}\log\frac{p(x_n|t_n,x_{n-1})}{p(x_n|t_{n-1},x_{n-1})}\right]\right)\right\rangle_{p(\mathbf{x})} = \\ &\overset{(1.)}{=} \sum_{x_0,\,x_n,\cdots x_N} p(x_0|t_0)\prod_{n=1}^{N}p(x_n|t_n,x_{n-1})\prod_{n=1}^{N}\frac{\exp(\beta U(x_n,t_n))}{\exp(\beta U(x_n,t_{n-1}))}\prod_{n=1}^{N}\frac{p(x_n|t_{n-1},x_{n-1})}{p(x_n|t_n,x_{n-1})} \\ &\overset{(2.)}{=} \sum_{x_0,\cdots x_n,\cdots x_N} p(x_0|t_0)\frac{\exp(\beta U(x_1,t_1))}{\exp(\beta U(x_1,t_0))}\prod_{n=2}^{N}\frac{\exp(\beta U(x_n,t_n))}{\exp(\beta U(x_n,t_{n-1}))}p(x_1|t_0,x_0)\prod_{n=2}^{N}p(x_n|t_{n-1},x_{n-1}) \\ &\overset{(3.)}{=} \frac{1}{Z_0}\sum_{x_1\cdots x_n,\cdots x_N}\exp(\beta U(x_1,t_1))\prod_{n=2}^{N}\frac{\exp(\beta U(x_n,t_n))}{\exp(\beta U(x_n,t_{n-1}))}\prod_{n=2}^{N}p(x_n|t_{n-1},x_{n-1}) \\ &\overset{(4.)}{=}= \frac{1}{Z_0}\sum_{x_2\cdots x_n,\cdots x_N}\prod_{n=2}^{N}\frac{\exp(\beta U(x_n,t_n))}{\exp(\beta U(x_n,t_{n-1}))}\prod_{n=3}^{N}p(x_n|t_{n-1},x_{n-1})\underbrace{\sum_{x_1}\exp(\beta U(x_1,t_1))p(x_2|t_1,x_1)}_{=\exp(\beta U(x_2,t_1))\text{(Detailed Balance)}} \\ &\overset{(5.)}{=} \frac{1}{Z_0}\sum_{x_N}\exp(\beta U(x_N,t_N)) = \frac{Z_N}{Z_0} = e^{\beta\Delta F}\end{aligned}$$

In (1.), we unfold the expression and exploit the equality $e^{\log p+\log q} = pq$ for the summation inside the exponential. In (2.), we cancel the trajectory probabilities $\prod_{n=1}^{N}p(x_n|t_n,x_{n-1})$ and then take one term out of the two remaining products. In (3.), first, we use the equivalence $\exp(\beta U(x_1,t_0)) = Z_0 p_{\mathrm{eq}}(x_1|t_0)$

(because at time t_0, the decision-maker is acting according to the equilibrium distribution) that allows us to cancel with $p(x_1|t_0, x_0) = p_{\text{eq}}(x_1|t_0)$, and second, we sum over x_0 with the only term that depends on it being $p(x_0|t_0)$. In (4.), we take one term of the second product and perform the sum over x_1 to obtain by detailed balance $\exp(\beta U(x_2, t_1))$ that will allow us to cancel with the term in the denominator of the first product. We perform Steps (3.) and (4.) repeatedly until obtaining the last equivalence that proves the theorem. □

Again, we note that the previously-proven Jarzynski relation from Equation (21) is not the same equation as in the no-deliberation case (13). In the deliberation case, the definition of the net utility is different and takes into account both the utility gain and the computational cost of deliberating.

We can now state the second law of decision-making with deliberation costs as:

$$\left\langle \mathcal{U}^{\text{diss}}(\mathbf{x}) \right\rangle_{p(\mathbf{x})} = \frac{1}{\beta} D_{\text{KL}} \left(p(\mathbf{x}) || p^{\dagger}(\mathbf{x}) \right) \geq 0 \tag{22}$$

from Equation (20) by rearranging and taking expectations. The same inequality can be obtained from Equation (21) by applying Jensen's inequality $\langle \exp x \rangle \geq \exp \langle x \rangle$ to recover $\langle \mathcal{U}^{\text{net}}(\mathbf{x}) \rangle_{p(\mathbf{x})} \leq \Delta F$. Equation (21) connects finite with infinite time decision-making. That is, there is a relation between the equilibrium free-energy differences that is the maximum attainable net utility with unlimited computation time and the net utility obtained by decision-makers with limited computation time. In the next section, we will provide examples of how to use these relations to extract useful information from decision-making processes.

4.1. Examples

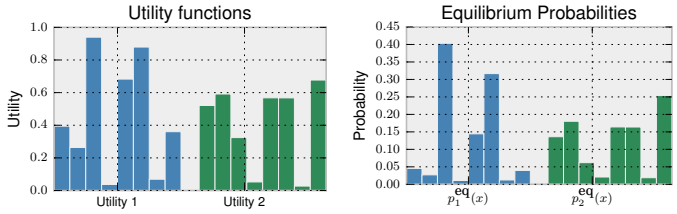
For the deliberation scenario, we illustrate the novel Jarzynski equality and Crooks theorem for decision-making in two decision-making scenario with clearly defined independent episodes: the first case is a discrete decision-making problem, and the second case is a continuous decision-making problem.

4.1.1. Jarzynski and Crooks Relations for Episodic Decision-Making with Deliberation

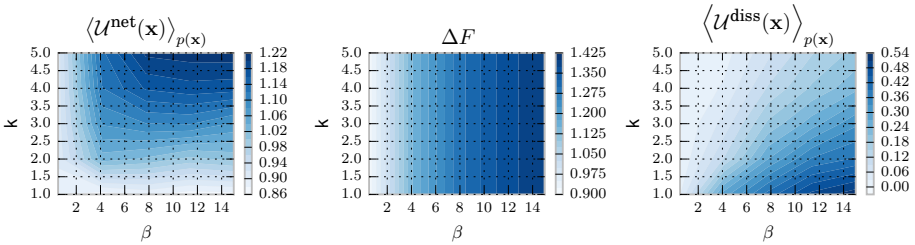
Choice-reaction-time experiments aimed to study information-processing in humans typically consider episodic tasks consisting of many trials; see [55] for a recent example. Here, we take a variation of Hicks episodic task with discrete action space, commonly used in the decision-making literature. In our variation of Hicks task, the decision-maker is shown a set of eight light bulbs. Initially, all light bulbs are turned off. Upon stimulus presentation, all light bulbs are turned on with different light intensities (representing different utilities) for a limited amount of time in which the decision-maker must choose the brightest light associated with the highest utility. The choice task is repeated many times, each time with different light intensities. For simplicity, our example contains only two stimuli: compare Utility 1 and Utility 2 in Figure 4A. When given enough time, a decision-maker with prior $p_0(x)$ chooses its actions according to the equilibrium distribution from Equation (4), as illustrated in Figure 4A for the uniform prior $p_0(x) = \frac{1}{8}$ that we assume in our example. In this case, the precision β specifies how well the light intensities can be told apart by a bounded optimal decision-maker.

In Figure 4, we model a decision-maker using the rejection sampling algorithm with the most efficient aspiration level given by the maximum utility $\max_x \Delta U(x)$. In particular, we simulate the rejection sampling algorithm with a limited number of samples (parameterized by k), where the choice strategy is given by non-equilibrium probability distribution in Equation (6) from the Introduction, because we assume that a response has to be produced within a fixed amount of time.

A



B



C

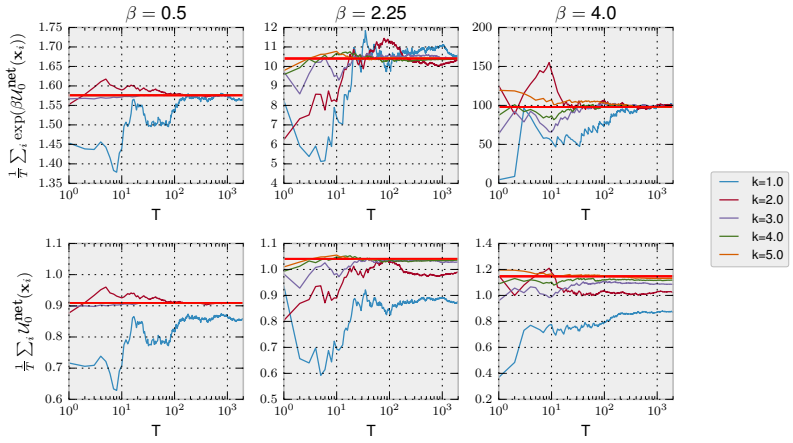


Figure 4. Episodic decision-making with deliberation. (**A**) Utility functions and equilibrium distributions for the two decision problems; (**B**) we show for different β and k (left) the average net utility, (middle) the free energy difference and (right) the average dissipated utility; (**C**) top panels: empirical averages approximating the Jarzynski expression in dependence of the number of trajectories T using different β and different number of available samples k; bottom panels: the associated expected net utility gain, which in the limit $T \to \infty$ is lower than the free energy difference (horizontal light red line).

In this kind of episodic task, the decision-maker always starts with the same prior $p_0(x)$ over the possible choices x. The probability of a trajectory of decisions $\mathbf{x}$ is defined as $p(\mathbf{x}) := \prod_{n=1}^{N} p(x_n|t_n)$ for each episode n, and the net utility for a trajectory is:

$$\mathcal{U}_0^{\text{net}}(\mathbf{x}) := \sum_{n=1}^{N} \left[\Delta U(x_n, t_{n-1} \to t_n) - \frac{1}{\beta} \log \frac{p(x_n|t_n)}{p_0(x_n)} \right].$$

Consequently, the equilibrium free energy is defined as $\Delta F := \max_{\tilde{p}(\mathbf{x})} \langle \mathcal{U}_0^{\text{net}}(\mathbf{x}) \rangle_{\tilde{p}(\mathbf{x})}$, which can also be decomposed into the sum of N independent equilibrium free energies $\Delta F = \sum_{n=1}^{N} \left\langle \Delta U(x_n, t_{n-1} \to t_n) - \frac{1}{\beta} \log \frac{p^{\text{eq}}(x_n|t_n)}{p_0(x_n)} \right\rangle_{p^{\text{eq}}(x_n|t_n)}$ where:

$$p^{\text{eq}}(x_n|t_n) = \frac{p_0(x_n) \exp(\beta \Delta U(x_n, t_{n-1} \to t_n))}{Z_n}$$

and the dissipated utility for a trajectory is $\mathcal{U}^{\text{diss}}(\mathbf{x}) := \Delta F - \mathcal{U}_0^{\text{net}}(\mathbf{x})$.

We simulate trajectories with $N = 2$ by sampling repeatedly from Equation (6). In the first panel of Figure 4B, we show that, as expected, the more samples k a decision-maker can afford, the higher the average net utility $\langle \mathcal{U}_0^{\text{net}} \rangle_{p(\mathbf{x})}$. In the second panel, it can be seen that the equilibrium free energy difference is invariant with respect to k and increases with higher precision β. Lastly, in the third panel, we plot the average dissipated utility $\langle \mathcal{U}^{\text{diss}} \rangle_{p(\mathbf{x})}$ that measures how much utility is lost due to the limited number of available samples. The highest dissipation occurs for high β and few samples k because such a high-precision decision-maker can potentially obtain high utility, but the limited amount of samples restrain it. In the following, we consider both a Jarzynski-like relation and a fluctuation theorem valid for a fixed prior.

Jarzynski Equality for Decision-Making with Fixed Prior p_0

For a fixed prior, it can readily be shown that the following relation is valid:

$$\left\langle e^{\beta \mathcal{U}_0^{\text{net}}(\mathbf{x})} \right\rangle_{p(\mathbf{x})} = e^{\beta \Delta F}. \tag{23}$$

To illustrate the validity of Equation (23), we simulated a decision-maker that faces T times the same two decision problems from Figure 4A. We can estimate the left-hand side of Equation (23) with the empirical average $\frac{1}{T} \sum_i \exp(\beta \mathcal{U}_0^{\text{net}}(\mathbf{x}_i))$ with the T trajectories of decisions, where $\mathbf{x}_i \sim p(\mathbf{x})$. In the top row of Figure 4C, we show the empirical average converging to $\exp(\beta \Delta F)$ (as expected by the law of large numbers) depending on the number of simulated trajectories T and precision β, empirically validating Equation (23). In the bottom row, we show how the second law for decision-making is fulfilled as the average net utility is less than the equilibrium free energy, thus satisfying the inequality (17).

Crooks' Fluctuation Theorem for Decision-Making with Fixed Prior p_0

For the fixed prior, it can readily be shown that the following fluctuation relation holds:

$$\frac{\tilde{p}(\mathbf{x})}{p^{\text{eq}}(\mathbf{x})} = e^{\beta(\Delta F - \mathcal{U}_0^{\text{net}}(\mathbf{x}))} = e^{\beta \mathcal{U}^{\text{diss}}(\mathbf{x})} \tag{24}$$

where $p^{\text{eq}}(\mathbf{x}) := \prod_{n=1}^{N} p^{\text{eq}}(x_n|t_n)$ is the optimal equilibrium distribution over trajectories $\mathbf{x}$. Note in this case that the probability distribution of the backward process $p^{\dagger}(\mathbf{x})$ coincides with the optimal equilibrium distribution $p^{\dagger}(\mathbf{x}) = p^{\text{eq}}(\mathbf{x})$ because of the independence of the decision problems. More specifically, the original Crooks theorem for decision-making from Equation (20) is valid only when the backward process starts in equilibrium. In our episodic task, all decision problems are independent, which makes the starting equilibrium distributions for all the backward processes coincide with the posterior equilibrium distributions of the forward process.

The fluctuation relation (24) for episodic tasks adopts a different meaning than the conventional relation. Specifically, the ratio between probabilities is now between the probability of observing a trajectory of actions when having finite time to make a decision (a sequence of non-equilibrium probabilities) and the probability of observing the same trajectory when having infinite time (a sequence

of equilibrium probabilities). This ratio is governed by the exponential of the dissipated utility $\mathcal{U}^{\text{diss}}(\mathbf{x})$ similarly to the original Crooks equation.

Equation (24) can be rewritten by re-arranging the terms and averaging over $p(\mathbf{x})$ as

$$\frac{1}{\beta} D_{\text{KL}}\left(p(\mathbf{x}) || p^{\text{eq}}(\mathbf{x})\right) = \left\langle \mathcal{U}^{\text{diss}}(\mathbf{x}) \right\rangle_{p(\mathbf{x})}.$$

Consequently, we see that purely from the trajectories of actions, we can obtain the average dissipated utility. We can test this relation in human experiments by comparing the trajectories of actions in two different conditions, first when having finite time and second when having as much time as needed. Then, from the probabilities of action trajectories, we can extract the average dissipated utility.

4.1.2. Jarzynski and Crooks Relations for Deliberating Continuous Decisions

Since many decision tasks take place in the continuous domain (for example, sensorimotor tasks), we now consider continuous state space problems. In particular, we repeat the same analysis as in the previous section by validating our Jarzynski equation, but this time in the continuous domain. Moreover, in this example, we allow for adaptive changes in the prior, such that the prior in one trial is equal to the posterior of the previous trial. In the following, we model decision-making as a diffusion process with Langevin dynamics that stops after a certain time t and emits an action x. The diffusion process uses gradient information to find the optimum utility and will converge to an equilibrium distribution for $t \to \infty$. In our example, we will employ quadratic utility functions that allow for a closed form solution of the non-equilibrium probability density that changes over time.

Let $x(t) \in \mathbb{R}$ be the dynamics of computation that a decision-maker carries out when deliberating. The differential equation that describes the dynamics is:

$$\frac{\partial x}{\partial t} = \alpha \frac{\partial U(x)}{\partial x} + \alpha \xi(t) \tag{25}$$

where $\xi(t)$ is white Gaussian noise with mean $\langle \xi(t) \rangle = 0$ and correlation $\langle \xi(t)\xi(t') \rangle = 2D\delta(t - t')$. Note that Equation (25) is closely related to learning algorithms that use gradient information such as Stochastic Gradient Descent (SGD). These algorithms find the minimum of a cost function by taking steps in the state space in the opposite direction of the gradient. Here, we see that the learning rate corresponds to the parameter α, which, in contrast with plain GD, not only multiplies the gradient, but also the noise term.

Equation (25) gives the dynamics of the decision-making process in terms of a stochastic differential equation, which can equivalently be expressed by the evolution of the probability $p(x,t)$ described by the Fokker–Planck equation [56]:

$$\frac{\partial p(x,t)}{\partial t} = -\alpha p(x,t) \frac{\partial^2 U(x)}{\partial x^2} - \alpha \frac{\partial U(x)}{\partial x} \frac{\partial p(x,t)}{\partial x} + D\alpha^2 \frac{\partial^2 p(x,t)}{\partial x^2}. \tag{26}$$

In order to compute the net utility, we need the probability of the non-equilibrium distribution up to a desired time t; thus, we need to solve the Fokker–Planck equation. For quadratic utility functions $U_y(x) = -(a_y x^2 + b_y x)$ with coefficients a_y and b_y for environment y and initial Gaussian distribution with mean μ_0 and variance σ_0^2, the solution is (see Appendix):

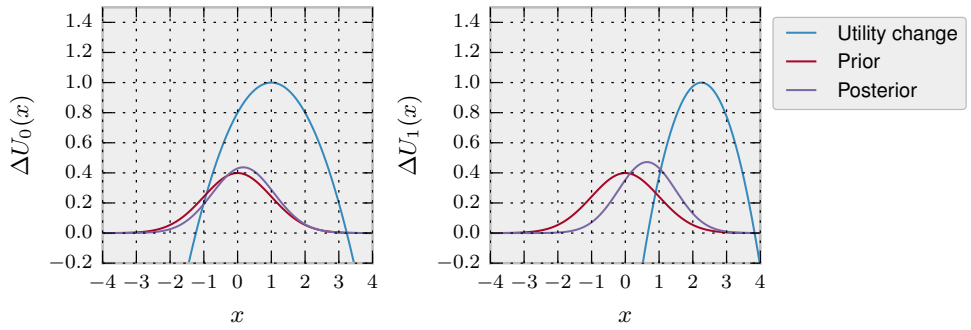
$$p(x,t) = \frac{1}{\sqrt{2\pi\sigma^2(t)}} e^{\frac{-(x-\mu(t))^2}{2\sigma^2(t)}} \tag{27}$$

with:

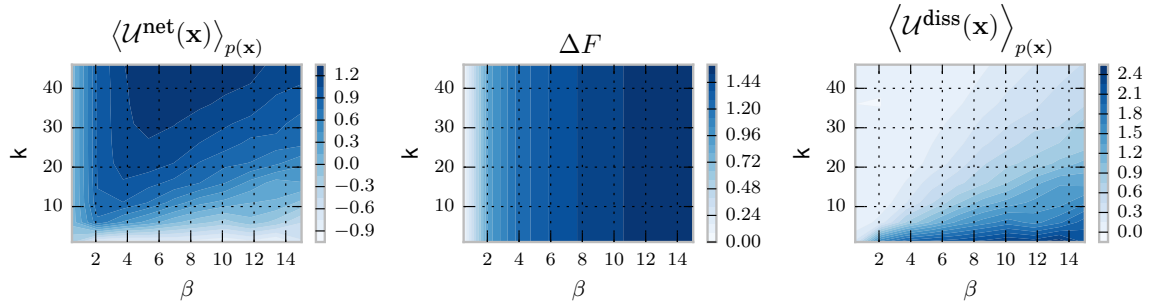
$$\sigma^2(t) = \frac{\alpha^2 D}{2c}\left(1 - e^{-2ct}\right) + \sigma_0^2 e^{-2ct}$$
$$\mu(t) = e^{-ct}\mu_0 - \frac{b_1}{2a_1}(1 - e^{-ct})$$

where $c = 2\alpha a_1$, and we assumed that the prior strategy is Gaussian distributed with mean μ_0 and variance σ_0^2. The precision parameter relates to the other parameters with the relation $\beta = \frac{2\alpha}{D}$, which means that the higher the α, the more we take into account the gradient leading to a higher β, and the lower the noise D, also the higher β.

Following a similar approach as in the previous section, we expose a decision-maker to two utility functions given by $U_1(x) = 0.2x^2 - 0.4x - 0.8$ and $U_2(x) = 0.4x^2 - 1.8x + 1.025$ shown in Figure 5A. The prior for the first utility is given by $\mu_0 = 0$ and $\sigma_0^2 = 1$. In Figure 5B, we show the net utility, equilibrium free-energy differences and dissipated utility (according to Equations (18) and (19)) for different values of β and number of steps k; corresponding to time $t = k\Delta t$ in Equation (27) for a given reference Δt. In Figure 5C, we show the convergence of the Jarzynski term towards the true equilibrium free energy difference term depending on the number of trajectories to make the estimation. We can see on the bottom row that the second law for decision-making represented by the inequality (17) is fulfilled.

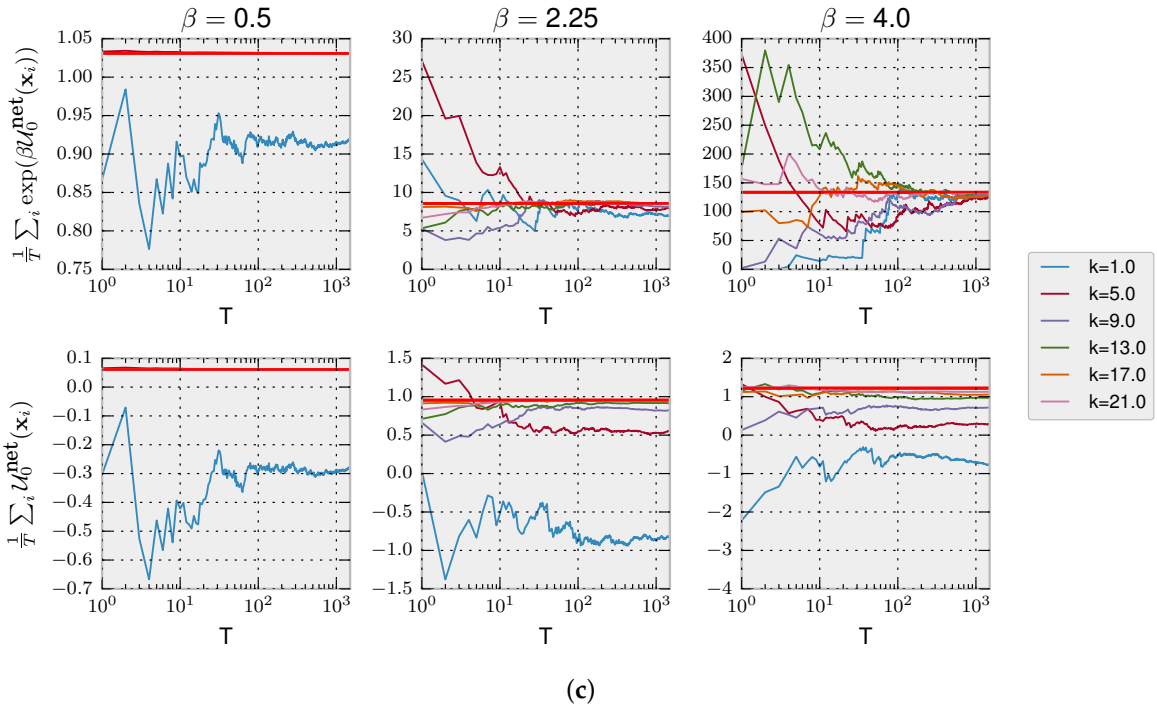


(**a**)



(**b**)

Figure 5. Cont.



(**c**)

Figure 5. Langevin dynamics simulations. (**a**) In blue, the different utility changes ΔU_1 and ΔU_2, in red the prior p_0 and in purple the posterior for $\beta = 0.5$; (**b**) We show for different β and time $t = k\Delta t$ directly depending on k, (left) the average net utility, (middle) the free energy difference and (right) the average dissipated utility; (**c**) top panels: convergence of the empirical Jarzynski estimate depending on the number of trajectories T using different β and different numbers of update steps k. Bottom panels: the associated expected net utility gain, which in the limit $T \to \infty$ is lower than the free energy difference (horizontal light red line). With these simulations, we validate Equation (21).

5. Discussion

In this paper, we highlighted the similarities between non-equilibrium thermodynamics and bounded rational decision-making in the case of agents that can deliberate before selecting an action and agents that cannot. Additionally, we derived a novel Jarzynski equality and a Crooks fluctuation theorem for decision-making scenarios with deliberation. We have shown how to use Jarzynski's and Crooks' equations in different scenarios to extract relevant variables of the decision-making process such as the equilibrium free energy difference, the average dissipated utility and the action-path probabilities for both equilibrium posterior distributions and distributions of the backward-in-time protocol. We have provided a number of examples for the no-deliberation and deliberation scenario, such as one-step lag dynamics, discrete choice tasks and continuous decision-making tasks that may be applicable both to cognitive and sensorimotor experiments [57].

In Section 3, we started out by directly translating physical non-equilibrium concepts to the decision-making domain in the case of decision-makers that cannot deliberate before acting and therefore lag behind changes in the utility landscape. In analogy to physical systems, we assumed that such decision-makers adapt to each utility change even though they are lagging behind, i.e., even after they have already chosen their action and there is no benefit of this adaptation at the current time step, but to improve their prior for the next choice. In physical systems, this does not constitute an issue, because there is a continuous adaptation to the energy gradient at every instant independent of how time is discretized. However, in the decision-making scenario, we assumed a single distinguished moment where the action is issued and the utility is evaluated. Therefore: Why should such decision-makers adapt at all after the action has been selected? Following the

argument of no-free lunch theorems, there would be no benefit in adapting to arbitrary changes. Having a closer look at our examples in Section 3.3, it becomes evident that we implicitly assumed that the utility changes in each step were small, so there is a benefit in adapting the prior for the next trial. Such assumptions are typically made in learning scenarios, for example the i.i.d. assumption for inference problems or assumptions that utility changes in each time step are limited to a finite interval in decision-making problems. However, none of the non-equilibrium relations we discussed necessarily assume small utility changes. It should therefore be noted that, while the discussed non-equilibrium relations hold for arbitrary utility changes, in the context of non-deliberative decision-making, we would have to make additional assumptions such that utility changes in each step are small and can accumulate so that adaptation is beneficial. Importantly, the appropriateness of adaptation is not an issue when we assume a deliberation process where adaptation occurs before emitting an action, as there is a direct benefit of adaptation in the current trial. This is the general decision problem discussed in Section 4.

While we have considered mainly non-sequential decision-making problems here for simplicity, the same formalism could also be applied to sequential decision-making problems. In that case, one would replace the notion that an action corresponds to a discrete or continuous state x with the notion that an action might consist of choosing an entire trajectory $x_{1:\tau}$. In this case also, the utility $U(x_{1:\tau}, t)$ would be defined over trajectories, and these utilities would change over episodes t. Again, one would have to assume that the utility function does not change while the trajectory $x_{1:\tau}$ is generated. This corresponds to the fact that we assume that the utility is constant for each single episode t (cf. Figure 1), while the deliberative decision-maker can, as it were, sample the new utility function before emitting an action. An example would be finding a trajectory for a pendulum swing-up or a sequence of actions to navigate a maze. A path integral controller [58] would for example exactly produce such trajectories. A deliberative decision-maker would sample many such trajectories until time is up and one trajectory has to be selected, then the utility changes again, and the path integral controller samples new trajectories that have a different shape in line with the new utility function. Our assumption that the temporal evolution of the utility function does not depend on the decision-maker's action implies that consecutive episodes are independent and can have different utility functions, but the decision-maker can carry its prior from one episode over to the next.

Recently, there has been a renewed interest in modelling decision-making with computational constraints [59,60] both in the computer science and the neuroscience literature, where there is growing evidence that the human brain might exploit sampling [22,61–65] for approximate inference and decision-making [66,67]. Such sampling models have been used for example to explain anchoring biases in choice tasks, because MCMC has finite mixing times and therefore exhibits a dependence on the prior distribution [68,69]. In particular, the idea of using the (expected) relative entropy or the mutual information as a computational cost has been suggested several times in the literature [2,3,23,33,70–72]. In [33] and similarly in [20], the authors derive the relative entropy as a control cost from an information-theoretic point of view, under axioms of monotonicity and invariance under relabelling and decomposition. In other fields such as robotics, the relative entropy has also been used as a control cost [18,21,25,58,73,74] to regularize the behaviour of the controller by penalizing controls that are far from the uncontrolled dynamics of the system or to deal with model uncertainty [75]. Naturally, questions regarding the generality of entropic costs as information-processing costs and their potential relation to algorithmic space-time resource constraints carry over to the non-equilibrium scenario and remain a topic for future investigations.

So far, only very few studies have established connections between non-equilibrium thermodynamics and decision-making in the literature, even though non-equilibrium analysis might provide a promising way to relate mechanistic dynamical models to conceptually simpler utility-based models that are often employed as normative models. Jarzynski-like and Crooks-like relations have been noted in the economics literature in gambling scenarios [76] and when studying the arrow of time for decision-making [77,78]. We reported preliminary results for the one-step delayed decision-making

in [79,80]. In the machine learning literature, generalized fluctuation theorems have recently been used in [81] to train artificial neural networks with efficient exploration. In general, fluctuation theorems and Jarzynski equalities allow one to estimate free energy differences, which are very important in decision-making because the free energy directly relates to the value function, which is a central concept in control and reinforcement learning. Fluctuation theorems typically make the assumption that the temperature parameter is constant (isothermal transformations) and that initial states are in equilibrium. In our paper, we also made these assumptions, which may limit the generality of our results. Loosening these restrictions (cf. for example [82,83]) might be an important next step for future investigations of non-equilibrium relations in the decision-making context.

Regarding the connection between predictive power and dissipation, [24] has found that non-predictive systems are also systems that are highly dissipative. In [24], the authors consider the effects of a stochastic driving signal x mediated by an energy function $E(x, s)$ on the state s of a Markov system with fixed transition probability $p(s'|s, x)$. They regard the Markov system as a computing device and study how much information the state s carries about the driving signal x. They find a fundamental relationship between dissipation (energy efficiency) and lack of predictive power. Their results concern non-equilibrium trajectories when x changes at every time point. The intuition is that when a system naturally moves in the direction of a changing energy landscape, then this is not only more efficient energetically, but it can also be interpreted in the sense that the system predicts the changing energy landscape. Once the system equilibrates, the energy landscape (i.e., the external variable x) does not change any more, and the mutual information between state and external variable xvanishes, as does the dissipation. Therefore, the equilibrium state is of no particular interest in this analysis. If one were to apply this framework to a decision-maker, the decision-maker would be represented by the system with the state s, and the driving signal x would be the input provided to the decision-maker. One important difference between [24] and our formulation is that in [24], the driving signal x is stochastic and is sampled from a stationary probability distribution, whereas in our formulation, we assume a fixed deterministic driving signal (the sequence of utility functions) without an underlying probability distribution. Assuming such a fixed input does prohibit an analysis in terms of mutual information between s and x. Nevertheless, it would be straightforward to allow for stochastic changes in the utility function also in our formulation, and the results of [24] would be applicable and complementary. While in [24], the equilibrium is of no particular interest, in our analysis, we are interested in the approach to equilibrium and in the resources spent on the way, that is the time that is spent during deliberating where the environment is assumed to be roughly constant, i.e., it does not change too much on the short time scale of deliberating, then the environment changes again, and the decision-maker can adapt to this change by deliberation (in contrast, in [24], the decision-maker follows a fixed dynamics and does not adapt).

In conclusion, the results presented here bring the fields of stochastic thermodynamics and decision-making closer together by studying decision-making systems as statistical systems just like in thermodynamics. In this analogy, the energy function in physics corresponds to the utility functions in decision-making. Importantly, the statistical ensembles of both decisions and physical states can be conceptualized as non-equilibrium ensembles that reach equilibrium after a finite time adaptation process.

Acknowledgments: This study was supported by the ERC Starting Grant BRISC 678082, by the DFG Grant BR4164/1-1, the DFG Grant KR 3844/2-1 and the Max Planck Society. We thank our funding sources for supporting this study.

Author Contributions: J.G.-M.and D.A.B. conceived of the research. J.G.-M. did the derivations and performed the simulations. J.G.-M., M.K. and D.A.B. contributed to the main ideas of the paper. J.G.-M., M.K. and D.A.B. wrote the manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Dissipation for a One-Step Decision Problem

In the following, we derive Equation (16) from Equation (9) for a one-step decision problem. Let $\mathbf{x} = (x_0, x_1)$. The probabilities $p(\mathbf{x})$ of the forward trajectory are $p(\mathbf{x}) = p_0^{\text{eq}}(x_0)p(x_1|x_0, t_1)$, and the probabilities $p^\dagger(\mathbf{x})$ of the backward trajectory are $p^\dagger(\mathbf{x}) = p_1^{\text{eq}}(x_1)p^\dagger(x_0|x_1, t_0)$. The detailed balance condition allows us to re-write $p^\dagger(x_0|x_1, t_0)$ as $p^\dagger(x_0|x_1, t_0) = \frac{e^{\beta U(x_0,t_0)}}{e^{\beta U(x_1,t_0)}} p(x_1|x_0, t_0)$ with $e^{\beta U(x_0,t_0)} = Z_0 p_0^{\text{eq}}(x_0)$ and $e^{\beta U(x_1,t_0)} = Z_0 p_0^{\text{eq}}(x_1)$. With our notation in the deliberation scenario, x_1 is the decision, and x_0 is arbitrary and can be ignored. This effectively implies independence between x_0 and x_1, such that $p(x_1|x_0, t_1) = p(x_1|t_1)$ and $p^\dagger(x_0|x_1, t_0) = \frac{p_0^{\text{eq}}(x_0)}{p_0^{\text{eq}}(x_1)} p(x_1|t_0) = p_0^{\text{eq}}(x_0)$. Substituting the previous identities in the KL-divergence, we obtain:

$$\begin{aligned}\frac{1}{\beta} D_{\text{KL}}\left(p(\mathbf{x})||p^\dagger(\mathbf{x})\right) &= \frac{1}{\beta} \sum_{x_0,x_1} p_0^{\text{eq}}(x_0)p(x_1|t_1) \log \frac{p_0^{\text{eq}}(x_0)p(x_1|t_1)}{p_1^{\text{eq}}(x_1)p_0^{\text{eq}}(x_0)} \\ &= \frac{1}{\beta} D_{\text{KL}}\left(p(x_1|t_1)||p_1^{\text{eq}}(x_1)\right) = \frac{1}{\beta} D_{\text{KL}}\left(\tilde{p}(x)||p_1^{\text{eq}}(x)\right)\end{aligned}$$

where we have made the replacement $p(\cdot|t_1) = \tilde{p}(\cdot)$ to obtain the notation from Figure 1.

Appendix B. Fokker-Planck Solution of Continuous Decision-Making Problem

A solution of the Fokker-Planck Equation (26) for known initial state x_0 can be found in [84]. Here, we sketch the solution when the initial state is Gaussian distributed.

Consider the following dynamics:

$$\frac{dx}{dt} = A(x, t) + B(x, t)\xi(t)$$

where $A(x, t) = \alpha \frac{\partial U_1}{\partial x}$, $B(x, t) = \alpha$. When imposing a quadratic utility function:

$$U_y(x) = -(a_y x^2 + b_y x)$$

for an environment indexed by $y = 1$, the associated Fokker–Planck equation is

$$\frac{\partial P}{\partial t} = 2\alpha a_1 \frac{\partial}{\partial x} xP + \alpha b_1 \frac{\partial}{\partial x} P + \alpha^2 D \frac{\partial^2}{\partial x^2} P.$$

We will solve this equation by first taking the Fourier transform in the variable x and then solving by the method of characteristics. The Fourier transform is:

$$\begin{aligned}\frac{\partial \hat{P}}{\partial t} &= -cs\frac{\partial \hat{P}}{\partial s} - \alpha^2 D s^2 \hat{P} + \alpha b_1 is\hat{P} \\ &= -cs\frac{\partial \hat{P}}{\partial s} + \hat{P}\left(c_2 is - \alpha^2 D s^2\right)\end{aligned}$$

where $c = 2\alpha a_1$ and $c_2 = \alpha b_1$. Now, applying the method of characteristics:

$$\frac{d\hat{P}}{dx} = \frac{\partial \hat{P}}{\partial s}\frac{ds}{dx} + \frac{\partial \hat{P}}{\partial t}\frac{dt}{dx}$$

we obtain that $dt = dx$, $s = s_0 e^{ct}$, and applying these relations, we get:

$$\frac{d\hat{P}}{dx} = \frac{d\hat{P}}{dt} = \hat{P}\left(c_2 i s_0 e^{ct} - \alpha^2 D s_0^2 e^{2ct}\right)$$

Integrating over t between $t = 0$ and $t = t'$, we have that

$$\frac{d\hat{P}}{\hat{P}} = dt\left(c_2 i s_0 e^{ct} - \alpha^2 D s_0^2 e^{2ct}\right)$$
$$\log \hat{P}\Big|_{\hat{P}(s_0,t=0)}^{\hat{P}(s,t')} = \frac{c_2 i s_0}{c} e^{ct} - \frac{\alpha^2 D}{2c} s_0^2 e^{2ct}\Big|_{t=0}^{t=t'}.$$

Assuming a Gaussian distribution as a boundary condition with mean μ_0 and variance σ_0^2, the Fourier transform for the boundary is:

$$\hat{P}(s, t=0) = \exp\left\{-\frac{\sigma_0^2}{2} s_0^2 - i s_0 \mu_0\right\}.$$

Then, the solution in frequency space is:

$$\begin{aligned}\hat{P}(s,t) &= \exp\left\{-\frac{\alpha^2 D}{2c} s^2 (1 - e^{-2ct}) - \sigma_0^2 s^2 e^{-2ct} + i s \frac{b_1}{2a_1}(1 - e^{-ct}) - i s e^{-ct}\right\} \\ &= \exp\left\{s^2 f_1(t) - i s f_2(t)\right\}\end{aligned}$$

with $f_1(t) = -\frac{\alpha^2 D}{2c}\left(1 - e^{-2ct}\right) - \frac{\sigma_0^2}{2} e^{-2ct}$ and $f_2(t) = e^{-ct}\mu_0 - \frac{b_1}{2a_1}(1 - e^{-ct})$. Transforming back to the signal domain, we obtain:

$$\sigma^2(t) = -2 f_1(t) = \frac{\alpha^2 D}{c}\left(1 - e^{-2ct}\right) + \sigma_0^2 e^{-2ct}$$
$$\mu(t) = f_2(t) = e^{-ct}\mu_0 - \frac{b_1}{2a_1}(1 - e^{-ct}).$$

References

1. Ortega, P.A.; Braun, D.A. Thermodynamics as a theory of decision-making with information-processing costs. *Proc. R. Soc. A Math. Phys. Eng. Sci.* **2013**, *469*, doi:10.1098/rspa.2012.0683.
2. Wolpert, D.H. Information theory-the bridge connecting bounded rational game theory and statistical physics. In *Complex Engineered Systems*; Springer: New York, NY, USA, 2006; pp. 262–290.
3. Tishby, N.; Polani, D. Information theory of decisions and actions. In *Perception-Action Cycle*; Springer: New York, NY, USA, 2011; pp. 601–636.
4. Wolpert, D.H. The free energy requirements of biological organisms; implications for evolution. *Entropy* **2016**, *18*, 138, doi:10.3390/e18040138.
5. Von Neumann, J.; Morgenstern, O. *Theory of Games and Economic Behavior*; Princeton University Press: Princeton, NJ, USA, 1944.
6. Savage, L.J. *The Foundations of Statistics*; John Wiley and Sons: New York, NY, USA, 1954.
7. Simon, H.A. A behavioural model of rational choice. *Q. J. Econ.* **1955**, *69*, 99–118.
8. Simon, H.A. Rational decision-making in business organizations. *Am. Econ. Rev.* **1979**, *69*, 493–513.
9. Russell, S. Rationality and intelligence. In Proceedings of the 14th International Joint Conference on Artificial Intelligence, Montreal, QC, Canada, 20–25 August 1995; pp. 950–957.
10. Russell, S.J.; Subramanian, D. Provably bounded-optimal agents. *J. Artif. Intell. Res.* **1995**, *2*, 575–609.
11. Howes, A.; Lewis, R.; Vera, A. Rational adaptation under task and processing constraints: Implications for testing theories of cognition and action. *Psychol. Rev.* **2009**, *116*, 717–751.
12. Horvitz, E. *Reasoning under Varying and Uncertain Resource Constraints*; AAAI: Menlo Park, CA, USA, 1988; Volume 88, pp. 111–116.
13. Dean, T. An Analysis of time-dependent planning. In Proceedings of the Seventh AAAI National Conference on Artificial Intelligence, Saint Paul, Minnesota, 21–26 August 1988.
14. Zilberstein, S. Using any time algorithms in intelligent systems. *AI Mag.* **1996**, *17*, 73, doi:10.1609/aimag.v17i3.1232.

15. Kahneman, D. Maps of bounded rationality: Psychology for behavioural economics. *Am. Econ. Rev.* **2003**, *93*, 1449–1475.
16. Gigerenzer, G.; Goldstein, D.G. Reasoning the fast and frugal way: Models of bounded rationality. *Psychol. Rev.* **1996**, *103*, 650–669.
17. Camerer, C. *Behavioral Game Theory: Experiments in Strategic Interaction*; Princeton University Press: Princeton, NJ, USA, 2003.
18. Todorov, E. Efficient computation of optimal actions. *Proc. Natl. Acad. Sci. USA* **2009**, *106*, 11478–11483.
19. Still, S. An information-theoretic approach to interactive learning. *Europhys. Lett.* **2009**, *85*, 28005, doi:10.1209/0295-5075/85/28005.
20. Ortega, P.; Braun, D. Information, utility and bounded rationality. *Lect. Notes Artif. Intell.* **2011**, *6830*, 269–274.
21. Braun, D.; Ortega, P.; Theodorou, E.; Schaal, S. Path integral control and bounded rationality. In Proceedings of the 2011 IEEE Symposium on Adaptive Dynamic Programming And Reinforcement Learning (ADPRL), Paris, France, 11–15 April 2011; pp. 202–209.
22. Friston, K. The free-energy principle: A unified brain theory? *Nat. Rev. Neurosci.* **2010**, *11*, 127–138.
23. Rubin, J.; Shamir, O.; Tishby, N. Trading value and information in MDPs. *Intell. Syst. Ref. Libr.* **2012**, *28*, 57–74.
24. Still, S.; Sivak, D.A.; Bell, A.J.; Crooks, G.E. Thermodynamics of prediction. *Phys. Rev. Lett.* **2012**, *109*, 120604, doi:10.1103/PhysRevLett.109.120604.
25. Kappen, H.; Gómez, V.; Opper, M. Optimal control as a graphical model inference problem. *Mach. Learn.* **2012**, *1*, 1–11.
26. Vijayakumar, K.R.; Toussaint, M.; Vijayakumar, S. On stochastic optimal control and reinforcement learning by approximate inference. In Proceedings of the Robotics: Science and Systems, Sydney, Australia, 9–13 July 2012.
27. Braun, D.A.; Ortega, P.A. Information-theoretic bounded rationality and ε-optimality. *Entropy* **2014**, *16*, 4662–4676.
28. Luce, R. *Individual Choice Behavior*; Wiley: Oxford, UK, 1959.
29. Meginnis, J. *A new Class of Symmetric Utility Rules for Gambles, Subjective Marginal Probability Functions, and a Generalized Bayes Rule;* Columbia University, Graduate School of Business: New York, NY, USA, 1976; pp. 471–476.
30. McFadden, D. Econometric models for probabilistic choice among products. *J. Bus.* **1980**, *53*, S13–S29.
31. McKelvey, R.D.; Palfrey, T.R. Quantal response equilibria for normal form games. *Games Econ. Behav.* **1995**, *10*, 6–38.
32. Fudenberg, D.; Levine, D. *The Theory of Learning in Games*; MIT Press: Cambridge, MA, USA, 1998.
33. Mattsson, L.G.; Weibull, J.W. Probabilistic choice and procedurally bounded rationality. *Games Econ. Behav.* **2002**, *41*, 61–78.
34. Sims, C.A. Implications of rational inattention. *J. Monetary Econ.* **2003**, *50*, 665–690.
35. Polani, D.; Nehaniv, C.; Martinetz, T.; Kim, J. Relevant information in optimized persistence vs. progeny strategies. In Proceedings of the Tenth International Conference on the Simulation and Synthesis of Living Systems, Bloomington, IN, USA, 3–7 June 2006.
36. Stratonovich, R. On value of information. *Izv. USSR Acad. Sci. Tech. Cybern.* **1965**, *5*, 3–12.
37. Kanaya, F.; Nakagawa, K. On the practical implication of mutual information for statistical decisionmaking. *IEEE Trans. Inf. Theory* **1991**, *37*, 1151–1156.
38. Akamatsu, T. Cyclic flows, markov process and stochastic traffic assignment. *Transp. Res. Part B Methodol.* **1996**, *30*, 369–386.
39. Belavkin, R.V. Information trajectory of optimal learning. In *Dynamics of Information Systems*; Springer: New York, NY, USA, 2010; pp. 29–44.
40. Rieskamp, J. The probabilistic nature of preferential choice. *J. Exp. Psychol. Learn. Mem. Cogn.* **2008**, *34*, 1446–1465.
41. Andrieu, C.; Freitas, N.; Doucet, A.; Jordan, M.I. An introduction to MCMC for machine learning. *Mach. Learn.* **2003**, *50*, 5–43.

42. Ortega, P.A.; Braun, D.A.; Tishby, N. Monte Carlo methods for exact & efficient solution of the generalized optimality equations. In Proceedings of the 2014 IEEE International Conference on Robotics and Automation (ICRA), Hong Kong, China, 31 May–5 June 2014; pp. 4322–4327.
43. Ortega, P.A.; Braun, D.A. Generalized thompson sampling for sequential decision-making and causal inference. *Complex Adapt. Syst. Model.* **2014**, *2*, 1–23.
44. Crooks, G.E. Nonequilibrium measurements of free energy differences for microscopically reversible Markovian systems. *J. Stat. Phys.* **1998**, *90*, 1481–1487.
45. Jarzynski, C. Equalities and inequalities: Irreversibility and the second law of thermodynamics at the nanoscale. *Annu. Rev. Condens. Matter Phys.* **2011**, *2*, 329–351.
46. Gomez-Marin, A.; Parrondo, J.; van den Broeck, C. Lower bounds on dissipation upon coarse graining. *Phys. Rev. E* **2008**, *78*, 011107, doi:10.1103/PhysRevE.78.011107.
47. Roldán, É. *Irreversibility and Dissipation in Microscopic Systems*; Springer: New York, NY, USA, 2014.
48. Jarzynski, C. Nonequilibrium equality for free energy differences. *Phys. Rev. Lett.* **1997**, *78*, 2690–2693.
49. Grünwald, P. The safe Bayesian. In Proceedings of the International Conference on Algorithmic Learning Theory, Lyon, France, 29–31 October 2012; pp. 169–183.
50. Caticha, A.; Giffin, A. Updating Probabilities. In *Bayesian Inference and Maximum Entropy Methods in Science and Engineering*; AIP Publishing: Melville, NY, USA, 2006; Volume 872, pp. 31–42.
51. Giffin, A.; Caticha, A. Updating Probabilities with Data and Moments. In *Bayesian Inference and Maximum Entropy Methods in Science and Engineering*; AIP Publishing: Melville, NY, USA, 2006; Volume 954, pp. 74–84.
52. Chib, S.; Greenberg, E. Understanding the metropolis-hastings algorithm. *Am. Stat.* **1995**, *49*, 327–335.
53. Gaveau, B.; Schulman, L. A general framework for non-equilibrium phenomena: The master equation and its formal consequences. *Phys. Lett. A* **1997**, *229*, 347–353.
54. Seifert, U. Entropy production along a stochastic trajectory and an integral fluctuation theorem. *Phys. Rev. Lett.* **2005**, *95*, 040602, doi:10.1103/PhysRevLett.95.040602.
55. Ortega, P.A.; Stocker, A.A. Human decision-making under limited time. In *Advances in Neural Information Processing Systems 29*; Lee, D.D., Sugiyama, M., Luxburg, U.V., Guyon, I., Garnett, R., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2016; pp. 100–108.
56. Garcia-Palacios, J. Introduction to the theory of stochastic processes and Brownian motion problems. *arXiv* **2007**, arXiv:cond-mat/0701242.
57. Jarvstad, A.; Hahn, U.; Rushton, S.K.; Warren, P.A. Perceptuo-motor, cognitive, and description-based decision-making seem equally good. *Proc. Natl. Acad. Sci. USA* **2013**, *110*, 16271–16276.
58. Kappen, H.J. Path integrals and symmetry breaking for optimal control theory. *J. Stat. Mech. Theory Exp.* **2005**, *2005*, P11011, doi:10.1088/1742-5468/2005/11/P11011.
59. Gershman, S.J.; Horvitz, E.J.; Tenenbaum, J.B. Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science* **2015**, *349*, 273–278.
60. Parkes, D.C.; Wellman, M.P. Economic reasoning and artificial intelligence. *Science* **2015**, *349*, 267–272.
61. Moreno-Bote, R.; Knill, D.C.; Pouget, A. Bayesian sampling in visual perception. *Proc. Natl. Acad. Sci. USA* **2011**, *108*, 12491–12496.
62. Levy, R.P.; Reali, F.; Griffiths, T.L. Modeling the effects of memory on human online sentence processing with particle filters. In Proceedings of the 23rd Annual Conference on Neural Information Processing Systems Vancouver, BC, Canada, 7–10 December 2009; pp. 937–944.
63. Griffiths, T.L.; Tenenbaum, J.B. Optimal predictions in everyday cognition. *Psychol. Sci.* **2006**, *17*, 767–773.
64. Sanborn, A.N.; Griffiths, T.L.; Navarro, D.J. Rational approximations to rational models: Alternative algorithms for category learning. *Psychol. Rev.* **2010**, *117*, 1144–1167.
65. Fiser, J.; Berkes, P.; Orbán, G.; Lengyel, M. Statistically optimal perception and learning: From behaviour to neural representations. *Trends Cogn. Sci.* **2010**, *14*, 119–130.
66. Lieder, F.; Griffiths, T.; Goodman, N. Burn-in, bias, and the rationality of anchoring. In Proceedings of the 26th Annual Conference on Neural Information Processing Systems, Lake Tahoe, NV, USA, 3–6 December 2012; pp. 2690–2798.
67. Vul, E.; Goodman, N.; Griffiths, T.L.; Tenenbaum, J.B. One and done? Optimal decisions from very few samples. *Cogn. Sci.* **2014**, *38*, 599–637.
68. Lieder, F.; Griffiths, T.L.; Huys, Q.J.M.; Goodman, N.D. The anchoring bias reflects rational use of cognitive resources. *Psychon. Bull. Rev.* **2017**, doi:10.3758/s13423-017-1286-8.

69. Lieder, F.; Griffiths, T.L.; Huys, Q.J.M.; Goodman, N.D. Empirical evidence for resource-rational anchoring and adjustment. *Psychono. Bull. Rev.* **2017**, doi:10.3758/s13423-017-1288-6.
70. Genewein, T.; Leibfried, F.; Grau-Moya, J.; Braun, D.A. Bounded rationality, abstraction, and hierarchical decision-making: An information-theoretic optimality principle. *Front. Robot. AI* **2015**, *2*, 27, doi:10.5281/zenodo.32410.
71. Still, S.; Precup, D. An information-theoretic approach to curiosity-driven reinforcement learning. *Theory Biosci.* **2012**, *131*, 139–148.
72. Ortega, P.A.; Braun, D.A. A minimum relative entropy principle for learning and acting. *J. Artif. Intell. Res.* **2010**, *38*, 475–511.
73. Theodorou, E.; Buchli, J.; Schaal, S. A generalized path integral control approach to reinforcement learning. *J. Mach. Learn. Res.* **2010**, *9999*, 3137–3181.
74. Peters, J.; Mülling, K.; Altün, Y. Relative Entropy Policy Search. In Proceedings of the Twenty-Fourth National Conference on Artificial Intelligence, Atlanta, GA, USA, 11–15 July 2010; pp. 1607–1612
75. Grau-Moya, J.; Leibfried, F.; Genewein, T.; Braun, D.A. Planning with information-processing constraints and model uncertainty in markov decision processes. *arXiv* **2016**, arXiv:1604.02080.
76. Hirono, Y.; Hidaka, Y. Jarzynski-type equalities in gambling: Role of information in capital growth. *arXiv* **2015**, arXiv:1505.06216.
77. Mlodinow, L.; Brun, T.A. Relation between the psychological and thermodynamic arrows of time. *Phys. Rev. E* **2014**, *89*, 052102.
78. Roldán, É.; Neri, I.; Dörpinghaus, M.; Meyr, H.; Jülicher, F. Decision making in the arrow of time. *Phys. Rev. Lett.* **2015**, *115*, 250602, doi:10.1103/PhysRevLett.115.250602.
79. Grau-Moya, J.; Braun, D.A. Bounded rational decision-making in changing environments. *arXiv* **2013**, arXiv:1312.6726
80. Grau-Moya, J.; Hez, E.; Pezzulo, G.; Braun, D. The effect of model uncertainty on cooperation in sensorimotor interactions. *J. R. Soc. Interface* **2013**, *10*, 20130554, doi:10.1098/rsif.2013.0554.
81. Hayakawa, T.; Aoyagi, T. Learning in neural networks based on a generalized fluctuation theorem. *Phys. Rev. E* **2015**, *92*, 052710, doi:10.1103/PhysRevE.92.052710.
82. Chatelain, C. A temperature-extended Jarzynski relation: Application to the numerical calculation of surface tension. *J. Stat. Mech. Theory Exp.* **2007**, *2007*, P04011, doi:10.1088/1742-5468/2007/04/P04011.
83. Gong, Z.; Quan, H.T. Jarzynski equality, Crooks fluctuation theorem, and the fluctuation theorems of heat for arbitrary initial states. *Phys. Rev. E* **2015**, *92*, 012131, doi:10.1103/PhysRevE.92.012131.
84. Risken, H. Fokker-planck equation. In *The Fokker-Planck Equation*; Springer: New York, NY, USA, 1984; pp. 63–95.