# Thermodynamics as a theory of decision-making with information processing costs

Pedro A. Ortega and Daniel A. Braun

November 27, 2024

**Abstract**

Perfectly rational decision-makers maximize expected utility, but crucially ignore the resource costs incurred when determining optimal actions. Here we propose an information-theoretic formalization of bounded rational decision-making where decision-makers trade off expected utility and information processing costs. Such bounded rational decision-makers can be thought of as thermodynamic machines that undergo physical state changes when they compute. Their behavior is governed by a free energy functional that trades off changes in internal energy—as a proxy for utility—and entropic changes representing computational costs induced by changing states. As a result, the bounded rational decision-making problem can be rephrased in terms of well-known concepts from statistical physics. In the limit when computational costs are ignored, the maximum expected utility principle is recovered. We discuss the relation to *satisficing* decision-making procedures as well as links to existing theoretical frameworks and human decision-making experiments that describe deviations from expected utility theory. Since most of the mathematical machinery can be borrowed from statistical physics, the main contribution is to axiomatically derive and interpret the thermodynamic free energy as a model of bounded rational decision-making.

## 1 Introduction

In everyday life decision-makers often have to make fast and frugal choices [1, 2]. Consider, for example, an antelope that quickly has to choose a direction of flight when faced with a predator. By the time an antelope had considered all possible flight paths to determine the optimal one, it would most probably be already eaten. In general, decision-makers seem to trade off the expected desirability of the consequences of an action against the effort and resources (time, money, food, computational effort, knowledge, opportunity costs, etc.) required for searching the optimum [3, 4].

Classic theories of decision making generally ignore information-processing costs by assuming that decision makers always pick the option with maximum

return—irrespective of the effort or the resources it might take to find or compute the optimal action [5, 6, 7]. Such decision-makers are described as *perfectly rational*. However, being perfectly rational seems to contradict our intuition of real-world decision-making, where information processing constraints play an important role [1]. This has led to an abundant literature on *bounded rationality* [8, 9, 10, 11]. Unlike perfectly rational decision makers, bounded rational decision-makers are subject to information processing constraints, that is they may have limited time and speed to process a limited amount of information.

## 1.1 Thermodynamic Intuition

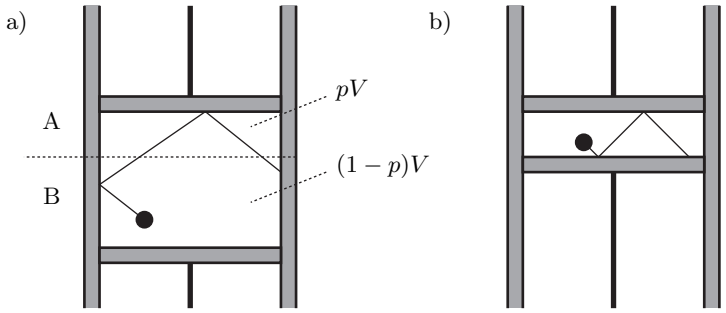


Figure 1: The Molecule-In-A-Box Device. (a) Initially, the molecule moves freely within a space of volume $V$ delimited by two pistons. The compartments A and B correspond to the two logical states of the device. (b) Then, the lower piston pushes the molecule into part A having volume $V' = pV$.

Here we follow a thermodynamic argument [12] that allows measuring resource (or information) costs in physical systems in units of energy. The generality of the argument relies on the fact that ultimately any real agent has to be incarnated in a physical system, as any process of information processing must always be accompanied by a pertinent physical process [13]. In the following we conceive of information processing as changes in information states (i.e. ultimately changes of probability distributions), which consequently implies changes in physical states, such as flipping gates in a transistor, changing voltage on a microchip, or even changing location of a gas particle. Such state changes in physical systems are not for free, that is the do not happen spontaneously. Consequently, if we want to control a physical system into a desirable state we also have to take into consideration that changing from the current state to the desirable state incurs a cost.

According to Landauer's principle, one can postulate a formal correspondence between one unit of information and one unit of energy [14, 15, 16]. Consider representing one bit of information using one of the following logical devices: a molecule that can be located either on the top or the bottom part of

a box; a coin whose face-up side can be either head or tail; a door that can be either open or closed; a train that can be orientated facing either north or south; and so forth. Assume that all these devices are initialized in an undetermined logical state, where the first state has probability $p$ and the second probability $1-p$. Now, imagine you want to set these devices to their first logical state. In the case of the molecule in a box, this means the following. Initially, the molecule is uniformly moving around within a space confined by two pistons as depicted in Figure 1a. Assuming that the initial volume is $V$, the molecule has to be pushed by the lower piston into the upper part of the box having volume $V' = pV$ (Figure 1b). From information theory, we know that the number of bits that we fix by this operation is given by $-\log p$.

To make things concrete, we assume that the device has diathermal walls and is immersed in a heat bath at constant temperature $T$. Since the walls are diathermal, the temperature inside of the box is maintained at the temperature of the heat bath. We model the particle as an ideal gas. When an ideal gas is compressed under isothermal conditions from an initial volume $V$ to a final volume $V'$, then the work is calculated as

$$W = -\int_V^{V'} \frac{NkT}{V}\, dV = NkT \ln \frac{V}{V'}, \tag{1}$$

where $N \geq 0$ is the amount of substance and $k > 0$ is the Boltzmann constant. The minus sign is just a convention to denote work done by the piston rather than by the gas. If we assume $N = 1$ and make use of the fact that $V' = pV$ we get

$$W = kT \ln \frac{V}{pV} = -kT \ln p = -\frac{kT}{\log e} \log p = -\gamma_{\text{mol}} \log p,$$

where the constant $\gamma_{\text{mol}} := \frac{RT}{\log e} > 0$ can be interpreted as the conversion factor between one unit of information and one unit of energy for the molecule-in-a-box device.

How do we compute the information and work for the case of the coin, door and train devices? The important observation is that we can model these cases as if they were like molecule-in-a-box devices, with the difference that their conversion factors between units of information and units of work are different. Hence, the number of bits fixed while these devices are set to the first state is given by $-\log p$, i.e. exactly as in the case of the molecule. However, the work is given by

$$-\gamma_{\text{coin}} \log p, \quad -\gamma_{\text{door}} \log p, \quad \text{and} \quad -\gamma_{\text{train}} \log p$$

respectively, where $\gamma_{\text{coin}}$, $\gamma_{\text{door}}$ and $\gamma_{\text{train}}$ are the associated conversion factors between units of information. Obviously, $\gamma_{\text{mol}} \leq \gamma_{\text{coin}} \leq \gamma_{\text{door}} \leq \gamma_{\text{train}}$. The point is that changes in knowledge states are costly and that these costs are proportional to the information. In the next section, we derive a general expression of information costs in physical systems that make decisions.

# 2 Information-Theoretic Foundations

## 2.1 Resource Costs

We model any observable sequential process, such as a sequence of interactions or a sequence of computation steps, as a filtration on a measure space. To simplify our exposition, we consider only finite measure spaces. Let $(\Omega, \Sigma)$ denote a measurable space, where $\Omega$ denotes the sample space and where $\Sigma$ is a $\sigma$-algebra on $\Omega$. Let $p$ be a conditional probability measure on $(\Omega, \Sigma)$, such that for any two events $A, B \in \Sigma$, $p(A|B)$ denotes the conditional probability of the $A$ given $B$, where the condition $B$ plays the role of the current information state of the process. The sequential realization of a process is modelled as a sequence of conditions $A_1, A_2, \ldots, A_T$ on the sample space $\Omega$, where each new condition $A_t$ refines the current information state $\bigcap_{\tau \leq t} A_\tau$ by excluding the complement $A_t^{\complement}$.

We further assume that a transformation of an information state from $B$ to $(A \cap B)$ entails a cost $\rho(A|B)$ that could be measured in dollars, time or any arbitrary scale of effort. Moreover, we assume that this transformation cost is decomposable; that is, if we undergo a knowledge change from $C$ to $(A \cap B \cap C)$, then we should pay the same cost as undergoing a change first from $C$ to $(B \cap C)$ and then from $(B \cap C)$ to $(A \cap B \cap C)$. Finally, the quintessential information-theoretic postulate is that conditional probabilities impose a monotonic order over transformation costs[1]. We can sum up our postulates as follows:

**Definition 1** (Axioms of Transformation Costs)**.** Let $(\Omega, \Sigma)$ be a measurable space and let $p : (\Sigma \times \Sigma) \to [0, 1]$ be a conditional probability measure over $\Sigma$ (i.e. for any $A \in \Sigma$, $p(\cdot|A)$ is a probability measure over $A$). A function $\rho : (\Sigma \times \Sigma) \to \mathbb{R}^+$ is a transformation cost function for $p$ iff it has the following three properties for all events $A, B, C, D \in \Sigma$:

$$\begin{aligned}
&\text{A1. real-valued:} && \exists f, \quad \rho(A|B) = f\big(p(A|B)\big) \in \mathbb{R}, \\
&\text{A2. additive:} && \rho(A \cap B|C) = \rho(A|C) + \rho(B|A \cap C), \\
&\text{A3. monotonic:} && [\rho(A|B) > \rho(C|D)] \quad \Leftrightarrow \quad [p(A|B) \lesseqgtr p(C|D)].
\end{aligned}$$

These three properties enforce a strict correspondence between probabilities and transformation costs [18, 19].

**Theorem 1** (Transformation Costs $\leftrightarrow$ Probabilities)**.** *If $f$ is such that $\rho(A|B) = f(p(A|B))$ for every choice of the probability space $(\Omega, \Sigma, p)$, then $f$ is of the form*

$$f(\cdot) = -\tfrac{1}{\beta} \log(\cdot),$$

*where $\beta$ is a real parameter.*

[1] This intuition is central for optimal coding theory where short codewords are assigned to frequent events and long codewords are assigned to rare events [17]. Therefore, we could regard the codeword length as a valuable resource that we have to bet on events with different probabilities.

That is, the transformation cost $\rho(A|B)$ is *proportional* to the information content $-\log p(A|B)$, where the parameter $\beta$ plays the role of the conversion factor. The logarithmic mapping between probabilities and "costs" is well-known in information theory, and there are many possible ways to derive it [20, 21]. The important observation is that our derivation stems purely from postulates regarding transformation costs.

According to Definition 1, transformation costs measure the relative cost of an event *relative* to a reference event. However, we can also introduce an *absolute* cost measure to single events such that transformation costs are obtained as differences.

**Definition 2** (Potential)**.** Let $\rho$ be a transformation cost function. A set function $\phi : \Sigma \to \mathbb{R}$ is called a *cost potential* for $\rho$ iff for all $A, B \in \Sigma$,

$$\begin{aligned} \phi(\Omega) &:= \phi_0 \\ \phi(A \cap B) &:= \phi(B) + \rho(A|B) \qquad \forall A, B \in \Sigma, \end{aligned}$$

where $\phi_0$ is an arbitrary real value.

One can easily verify that this potential is well defined for all events, and that $\rho(A|B) = \phi(A \cap B) - \phi(B)$. It captures the intuition that starting out from the high-probability event $B$ with potential $\phi(B)$ one has to pay the cost $\rho(A|B)$ to arrive at the low-probability event $A \cap B$ with potential $\phi(A \cap B)$.

In the following, consider a reference set $S \in \Sigma$ having a measurable partition $\mathcal{X}$. Cost potentials have an important recursive structure: the cost potential of an event is uniquely determined by the potential of its constituent events. If $\mathcal{X}$ is a measurable partition of a reference event $S \in \Sigma$, then

$$\phi(S) = -\tfrac{1}{\beta} \log \sum_{x \in \mathcal{X}} e^{-\beta \phi(x)}. \tag{2}$$

Furthermore, the probability of a member $x \in \mathcal{X}$ of the partition relative to $S$ can be expressed as a *Gibbs measure*:

$$p(x|S) = \frac{e^{-\beta\phi(x)}}{e^{-\beta\phi(S)}} = \frac{e^{-\beta\phi(x)}}{\sum_{x \in \mathcal{X}} e^{-\beta\phi(x)}}. \tag{3}$$

In statistical physics it is well-known that the Gibbs measure satisfies a variational principle in the *free energy*, which is defined as

$$F_\beta[q] := \sum_{x \in \mathcal{X}} q(x)\phi(x) + \frac{1}{\beta} \sum_{x \in \mathcal{X}} q(x) \log q(x). \tag{4}$$

More specifically, it is well known that for any probability measure $q$ over the partition $\mathcal{X}$ of $S$,

$$F[q] \geq F[p] = -\frac{1}{\beta} \log \phi(S), \tag{5}$$

where the lower bound is attained by the Gibbs measure $p(x) \propto e^{-\beta\phi(x)}$. Equations (2) to (5) constitute fundamental results that will be generalized and interpreted in the next section.

## 2.2 Gains and Losses

Equipped with the results from the preceding section, we can now proceed to model a bounded rational decision maker. Because transformation costs matter, we model a decision as a transformation of a prior behavior into a final behavior, where we represent the direction of change as a utility criterion.

The Gibbs measure in (3) allows us describing a probability measure $p$ over a partition $\mathcal{X}$ in terms of a cost potential $\phi$ over $\mathcal{X}$. In particular, we see that a decision-maker's a priori behavior or belief described by $p_0(x)$ and $\phi_0(x)$ changes to $p(x)$ and $\phi(x)$ if he is exposed to the gain (or loss) $U(x)$, such that

$$\phi(x) = \phi_0(x) - U(x) \tag{6}$$

and

$$p(x) \propto e^{-\beta\phi_0(x)+\beta U(x)} \propto p_0(x)e^{\beta U(x)} \tag{7}$$

as illustrated in Figure 1. The function $U$ represents either gains or losses and not absolute levels of costs, because it expresses a difference in the potential $U(x) = \phi_0(x) - \phi(x)$. The equilibrium distribution (7) that arises in a change can also be characterized in terms of a variational principle, in a manner analogous to (5).

**Theorem 2** (Negative Free Energy Difference)**.** *Let $p_0(x)$ and $p(x)$ be the Gibbs measures with potentials $\phi_0(x)$ and $\phi(x)$ and resource parameter $\beta$. Let $F_0$ and $F$ be the free energies minimized by $p_0$ and $p$ respectively. Then, the negative free energy difference $-\Delta F = F_0 - F$ is*

$$-\Delta F = \sum_{x\in\mathcal{X}} p(x)U(x) - \frac{1}{\beta}\sum_{x\in\mathcal{X}} p(x)\log\frac{p(x)}{p_0(x)}, \tag{8}$$

*where $U(x) = \phi_0(x) - \phi(x)$.*

Since the difference in the negative free energy $-\Delta F = F - F_0$ has the same dependency on $p$ as the free energy $F$, we can use $-\Delta F$ directly as a variational principle in $p$.

**Corollary 3** (Variational Principle)**.** *The negative free energy difference provides a variational principle for the equilibrium distribution, i.e.*

$$-\Delta F[q] := \sum_{x\in\mathcal{X}} q(x)U(x) - \frac{1}{\beta}\sum_{x\in\mathcal{X}} q(x)\log\frac{q(x)}{p_0(x)}$$

*is maximized by*

$$p(x) = \frac{1}{Z}p_0(x)e^{\beta U(x)}, \quad \textit{where} \quad Z := \sum_{x\in\mathcal{X}} e^{\beta U(x)}.$$

*Furthermore,*

$$\Delta F[q] \leq \Delta F[p] = \tfrac{1}{\beta}\log Z.$$

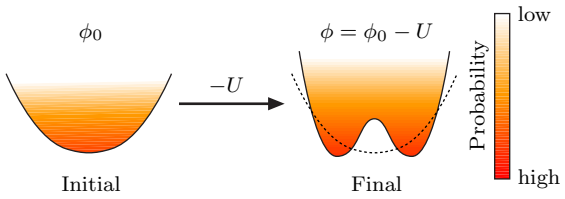


Figure 2: Representing a decision maker as a thermodynamic system, the behavior of the decision-maker exposed to a gain $U$ can be expressed as a change of his initial cost potential $\phi_0$ to a final cost potential $\phi$, where $\phi = \phi_0 - U$. The choice or belief probabilities of the decision-maker change according to (7) from $p_0$ to $p$.

## 2.3 Choice & Belief Probabilities

The distribution (7) can be interpreted both as an action or observation probability in the context of bounded rational decision-making. In the case of actions, $p_0$ represents the a priori choice probability of the agent which is refined to the choice probability $p$ when evaluating the imposed gain (or loss) $U$. The associated change in probability depends on the resource parameter $\beta$ and corresponds to the computation that is necessary to evaluate the gains (or losses). In the case of observations, $p_0$ represents the a priori belief of the agent given by a probabilistic model, which is then distorted due to the presence of possible gains (or losses) that are evaluated by the holder of the belief. This way, model uncertainty and risk-aversion can be parameterized by $\beta$.

For different values of $\beta$ the distribution (7) has the following limits

$$\begin{aligned} \lim_{\beta\to\infty} p(x) &= \delta(x - x^*), \qquad x^* = \max_x U(x) \\ \lim_{\beta\to 0} p(x) &= p_0(x) \\ \lim_{\beta\to-\infty} p(x) &= \delta(x - x^*), \qquad x^* = \min_x U(x). \end{aligned}$$

In the case of actions the three limits imply the following: The limit $\beta \to \infty$ corresponds to the perfectly rational actor that infallibly selects the action that maximizes gain (or minimizes loss $-U(x)$. The limit $\beta \to 0$ is an actor without resources that simply selects his action according to his prior. The limit $\beta \to -\infty$ corresponds to an actor that is perfectly "anti-rational" and always selects the action with the worst outcome. In the case of observations the three limits correspond to an extremely optimistic observer ($\beta \to \infty$) who believes only in the best possible outcome, an extremely pessimistic observer ($\beta \to -\infty$) who anticipates only the worst, and a risk-neutral Bayesian observer ($\beta \to 0$) who simply relies on the probabilistic model $p_0$.

## 2.4 The Certainty Equivalent

In statistical physics [22], the free energy difference

$$\Delta A = \Delta E - Q = W$$

measures the amount of available "good energy" (work $W$) by subtracting the "bad energy" (heat $Q$) from the total energy $\Delta E = \mathbf{E}[U]$. The crucial physical intuition is that we have uncertainty about some aspects of the objects that make up the heat energy, for example we do not know the exact trajectories of all gas particles at temperature $\beta$. This uncertainty means that we do not have full control over the objects and cannot extract all the energy as work [12]. Economically speaking, the physical concept of work, and therefore also the difference in free energy, measures the certainty equivalent of a gain (or loss) that is contaminated by uncertainty. In general, we can therefore use the free energy difference to ascribe a certainty equivalent value to choice situations of the form (7). As can be seen from Corollary 3, this value is given by the log partition function, i.e. the logarithm of the normalization constant $Z$. For different values of $\beta$, the certainty equivalent takes the following limits

$$\begin{aligned}
\lim_{\beta \to \infty} \frac{1}{\beta} \log Z &= \max_x U(x) \\
\lim_{\beta \to 0} \frac{1}{\beta} \log Z &= \sum_x p_0(x) U(x) \\
\lim_{\beta \to -\infty} \frac{1}{\beta} \log Z &= \min_x U(x).
\end{aligned}$$

Again, the case $\beta \to \infty$ corresponds to the perfectly rational actor (or the extremely optimistic observer), the case $\beta \to -\infty$ corresponds to the perfectly "anti-rational" actor (or the extremely pessimistic observer) and the case $\beta \to 0$ corresponds to the actor that has no resources (or the risk-neutral observer) such that the best one can expect is the expected gain or loss.

Corollary 3 has two interpretations in statistical physics, either as an instantiation of a *minimum energy principle* or as a *maximum entropy principle* [22]. Accordingly, (7) can either be seen as the distribution that maximizes the entropy given a constraint on the expectation value of $U$ or as the distribution that minimizes the expectation of $-U$ given a constraint on the entropy of $p$. In the context of observer modeling, the first interpretation provides a principle for estimation and the second interpretation provides a principle for bounded rational decision-making in the case of acting, which is a maximum expected gain principle with a relative entropy constraint that bounds the information-processing capacity of the decision-maker. In the relative entropy we recognize the term $\frac{1}{\beta} \log p(x)$ as our transformation costs $\rho$ from Theorem 1 such that we can express the negative free energy difference $-\Delta F$ as

$$-\Delta F = \mathbf{E}[U] - \mathbf{E}[R],$$

where $U(x) = \phi_0(x) - \phi(x)$ represents gains (or losses) and $R(x) = \rho(x) - \rho_0(x)$ represents the extra resource costs required to achieve the gain (or loss) $U$.

We can therefore see how the variational principle of Corollary 3 formalizes a trade-off between expected gains (or losses) and information processing costs.

# 3 Summary of Main Concepts

In decision theory, choices between alternatives are usually formalized as choices between lotteries, where a lottery is formalized as a set $\mathcal{X}$ of possible outcomes, a probability distribution $p_0$ over $\mathcal{X}$, and a real-valued function $U$ over $\mathcal{X}$ called the utility function. In particular expected utility theory predicts that a decision-maker always chooses the lottery with the higher expected utility $\mathbf{E}[U] = \sum_x p_0(x)U(x)$. Here we introduce the notion of a *bounded lottery* as a lottery that is additionally characterized by a *resource parameter* $\beta \in \mathbb{R}$ that captures the resource constraints of the decision-maker.

We have derived a thermodynamic framework for bounded lotteries from simple axioms that measure information processing cost—see also [19]. The most important difference of bounded decision-making compared to perfectly rational decision-making is that the bounded decision-maker will not be able to choose infallibly the best lottery. In fact, the resource constraints lead to stochastic choice behavior which can be characterized by a probability distribution. The decision process then transforms an initial choice probability $p_0$ into a final choice probability $p$ by taking into account the utility gains (or losses) *and* the transformation costs. This transformation process can be formalized as

$$p(x) = \frac{1}{Z} p_0(x) e^{\beta U(x)}, \quad \text{where} \quad Z = \sum_{x'} p_0(x') e^{\beta U(x')}. \tag{9}$$

Accordingly, the choice pattern of the decision-maker is predicted by the probability $p$. Crucially, the probability $p$ extremizes the variational principle

$$\max_p \left\{ \sum_x p(x)U(x) - \frac{1}{\beta} \sum_x p(x) \log \frac{p(x)}{p_0(x)} \right\}. \tag{10}$$

These two terms can be interpreted as determinants of bounded rational decision-making in that they formalize a trade-off between an expected utility gain (first term) and the information processing cost of transforming $p_0$ into $p$ (second term). The certainty equivalent value of a bounded lottery can be obtained by inserting the choice probability $p$ from (9) into (10), yielding

$$V = \frac{1}{\beta} \log \left( \sum_x p_0(x) e^{\beta U(x)} \right), \tag{11}$$

which corresponds to the log partition sum. For different values of $\beta$, the cer-

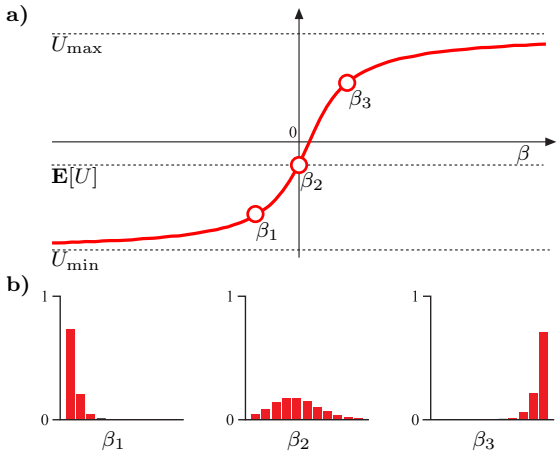


Figure 3: a) Negative free energy difference $\Delta F$ versus the resource parameter $\beta$. The resource parameter allows modeling decision-makers with bounded resources, either when generating their own actions ($\beta > 0$) or when anticipating their environment ($\beta < 0$). The negative free energy difference corresponds to the certainty equivalent. b) Distribution over the outcomes depending on the resource parameter $\beta$. For large positive $\beta$ the distribution concentrates on the outcome with maximum gain $U_{\max}$. For large negative $\beta$ the distribution concentrates on the worst outcome with gain $U_{\min}$. For $\beta = 0$ the outcomes follow the given distribution $p_0$.

tainty equivalent takes the following limits

$$\begin{aligned}
\lim_{\beta\to\infty} \frac{1}{\beta} \log Z &= \max_x U(x) \\
\lim_{\beta\to 0} \frac{1}{\beta} \log Z &= \sum_x p_0(x) U(x) \\
\lim_{\beta\to-\infty} \frac{1}{\beta} \log Z &= \min_x U(x).
\end{aligned}$$

The case $\beta \to \infty$ corresponds to the perfectly rational actor (or the extremely optimistic observer), the case $\beta \to -\infty$ corresponds to the perfectly "anti-rational" actor (or the extremely pessimistic observer) and the case $\beta \to 0$ corresponds to the actor that has no resources (or the risk-neutral observer) such that the best one can expect is the expected gain or loss. For illustration see Figure 2.

# 4 Bounded Rationality and Satisficing

Herbert Simon [23] proposed in the 50s that bounded rational decision-makers do not commit to an unlimited optimization by searching for the absolute best option. Rather, they follow a strategy of *satisficing*, i.e. they settle for an option that is *good enough* in some sense. Since then, it has been debated whether *satisficing* decision-makers can be described as bounded rational decision-makers that act optimally under resource constraints or whether optimization is the wrong concept altogether [11]. If decision-makers did indeed explicitly attempt to solve such a constrained optimization problem, this would lead to an infinite regress and the paradoxical situation that a bounded rational decision-maker would have to solve a more complex (i.e. constrained) optimization problem than a perfectly rational decision-maker.

To resolve this paradox, the bounded rational decision maker must not be able to reason about his constraints. He just searches randomly for the best option, until his resources run out. An observer will then be able to assign a probability distribution to the decision-maker's choices and investigate how this probability distribution changes depending on the available resources. Consider, for example, an anytime algorithm that will compute a solution more and more precisely the more time it has at its disposal. As one does not want to wait forever for an answer, the anytime computation will be interrupted at some point where one assumes that the answer is going to be good enough. This concept of satisficing can be used to interpret Equation 7 which describes the choice rule of a bounded rational decision-maker.

Consider the problem of picking the largest number in a sequence $U_0, U_1, U_2, \ldots$ of i.i.d. data, where each $U_i \in \mathcal{U}$ is drawn from a source with probability distribution $\mu$. This could be, for instance, an urn with numbered balls that we draw with replacement and we always keep track of the largest number seen so far. After $m$ draws the largest number will be given by

$$v := \max\{U_1, U_2, \ldots, U_m\}.$$

Naturally, the larger the number of draws, the higher the chances of observing a large number. The cumulative distribution function of choosing $v$ after $m$ draws is given by

$$F_m(v) = F_0(v)^m, \tag{12}$$

where $F_0$ is the cumulative distribution function of $\mu$ [24]. If we only cared about finding the maximum with absolute certainty then we would need to draw an infinite amount of times. However, a bounded rational decision-maker would stop after a certain time, when he feels that the benefit of further exploration does not justify the effort of further drawings. Thus, the number of draws in this example can be regarded as a resource and the numbers on the balls can be regarded as utilities. The behavior of the bounded rational decision-maker is then stochastic even though he acts perfectly deterministically, in the sense that he chooses option $v$ with probability (12) given the resource constraint $m$. According to (12), the more resources a decision-maker spends, the more he

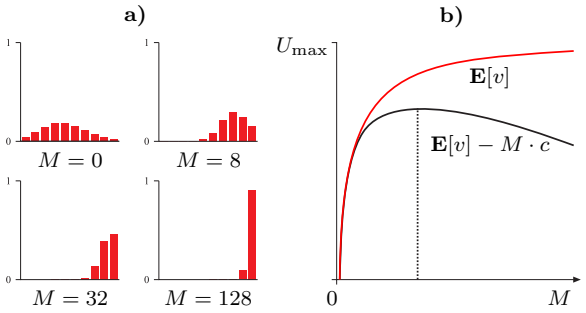


Figure 4: a) Distributions over the maximum for various sample sizes $(M + 1)$. The distribution $\mu$ over the ten values $v$ in $\mathcal{U} = \{1, 2, 3, \ldots, 10\}$ follows a truncated Poisson distribution with parameter $\lambda = 5$, as can be seen in the plot for $M = 0$. The distribution approaches a delta function over $v = 10$ for increasing values of $M$. b) The expected maximum $v$ versus sample size $(M+1)$. The incremental gain of the expected maximum is marginally decreasing as the sample size increases (red). If the sampling process is associated with a cost—e.g. $c = 0.02$ per sample in the figure—, then the penalized expected maximum (black) reaches a unique maximum for a finite sample size—the optimal sample size is $M = 35$ in the figure.

resembles a perfectly rational decision-maker that chooses the maximum number (Figure 1a), since the expected utility increases monotonically with the amount of resources spent (Figure 1b). Importantly, however, note that the marginal increase in the expected utility diminishes with larger effort—hence larger and larger effort pays out less and less in the end. Below we formalize this trade-off.

Here we show that the boundedness parameter $\beta$ plays an analogous role to the number of draws $m$. In the limit of a continuous cumulative function $F_0$, the density after $m$ draws is given by $p_m(v) = \frac{d}{dv} F_0(v)^m$. We can now compute the log odds for two random outcomes $v$ and $v'$, which results in

$$\log \frac{p_m(v)}{p_m(v')} = (m-1) \log \frac{F_0(v)}{F_0(v')} + \log \frac{\mu(v)}{\mu(v')},$$

where $F_0(v)$ is again the cumulative of $\mu$. If we require the probabilities $p_m(v)$ to be representable by a distribution of the exponential family such that $p_m(v) = \frac{\mu(v) \exp(\alpha U(v))}{\int dv' \mu(v') \exp(\alpha U(v'))}$, we see that the log odds have the following relation

$$\log \frac{p_m(v)}{p_m(v')} = \alpha \left( U(v) - U(v') \right) + \log \frac{\mu(v)}{\mu(v')}.$$

We see that $\alpha$ and $m$ play the role of the number of samples or computations. In general, the following theorem can be shown to hold.

**Theorem 4.** *Let $\mathcal{X}$ be a finite set. Let $Q$ and $M$ be strictly positive probability distributions over $\mathcal{X}$. Let $\alpha$ be a positive integer. Define $M_\alpha$ as the probability distribution over the maximum of $\alpha$ samples from $M$. Then, there are strictly positive constants $\delta$ and $\xi$ depending only on $M$ such that for all $\alpha$,*

$$\left| \frac{Q(x)e^{\alpha U(x)}}{\sum_{x'} Q(x')e^{\alpha U(x')}} - M_\alpha(x) \right| \leq e^{-(\alpha-\xi)\delta}.$$

Consequently, one can interpret the inverse temperature as a resource parameter that determines how many samples are drawn to estimate the maximum. Note that the distribution $M$ is arbitrary as long as it has the same support as $Q$. This interpretation can be extended to a negative $\alpha$, by noting that $\alpha U(x) = (-\alpha)(-U(x))$, i.e. instead of the maximum we take the minimum of $-\alpha$ samples.

# 5 Sequential Decision-Making

In the case of sequential decision-making the assumption of uniform temperatures has to be relaxed—the proofs of the following theorems can be found in [25]. In general, we can then dedicate different amounts of computational resources to each node of a decision tree. However, this requires a translation between a tree with a single temperature and to a tree with different temperatures. This translation can be achieved using the following theorem

**Theorem 5.** *Let $P$ be the equilibrium distribution for a given inverse temperature $\alpha$, utility function $U$ and reference distribution $Q$. If the temperature changes to $\beta$ while keeping $P$ and $Q$ fixed, then the utility function changes to*

$$V(x) = U(x) - \left(\tfrac{1}{\alpha} - \tfrac{1}{\beta}\right) \log \frac{P(x)}{Q(x)}.$$

If we now define the reward as the change in utility of two subsequent nodes, then the rewards of the resulting decision tree are given by

$$\begin{aligned} R(x_t|x_{<t}) &:= \left[V(x_{\leq t}) - V(x_{<t})\right] \\ &= \left[U(x_{\leq t}) - U(x_{<t})\right] - \left(\tfrac{1}{\alpha} - \tfrac{1}{\beta(x_{<t})}\right) \log \frac{P(x_t|x_{<t})}{Q(x_t|x_{<t})}. \end{aligned}$$

This allows introducing a collection of node-specific (not necessarily time-specific) inverse temperatures $\beta(x_{<t})$, allowing for a greater degree of flexibility in the representation of information costs. The next theorem states the connection between the free energy and the general decision tree formulation.

**Theorem 6.** *The free energy of the whole trajectory can be rewritten in terms*

*of rewards:*

$$F_\alpha[P] = \sum_{x_{\leq T}} P(x_{\leq T}) \Big\{ U(x_{\leq T}) - \frac{1}{\alpha} \log \frac{P(x_{\leq T})}{Q(x_{\leq T})} \Big\}$$
$$= U(\varepsilon) + \sum_{x_{\leq T}} P(x_{\leq T}) \sum_{t=1}^{T} \Big\{ R(x_t|x_{<t}) - \frac{1}{\beta(x_{<t})} \log \frac{P(x_t|x_{<t})}{Q(x_t|x_{<t})} \Big\}. \quad (13)$$

This translation allows applying the free energy principle to each node with a different resource parameter $\beta(x_{<t})$. By writing out the sum in (13), one realizes that this free energy has a nested structure where the latest time step forms the innermost variational problem and all other variational problems of the previous time steps can be solved recursively by working backwards in time. This then leads to the following solution:

**Theorem 7.** *The solution to the free energy in terms of rewards is given by*

$$P(x_t|x_{<t}) = \frac{1}{Z(x_{<t})} Q(x_t|x_{<t}) \exp\Big\{ \beta(x_{<t}) \big[ R(x_t|x_{<t}) + \frac{1}{\beta(x_{\leq t})} \log Z(x_{\leq t}) \big] \Big\},$$

*where* $Z(x_{\leq T}) = 1$ *and where for all* $t < T$

$$Z(x_{<t}) = \sum_{x_t} Q(x_t|x_{<t}) \exp\Big\{ \beta(x_{<t}) \big[ R(x_t|x_{<t}) + \frac{1}{\beta(x_{\leq t})} \log Z(x_{\leq t}) \big] \Big\}.$$

# 6 Limit Cases of Bounded Rational Control

As described in the previous section, the belief and action probabilities of an agent in a sequential decision-making setup can be determined by recursion of the log-partition function

$$V(x_{<t}) = \frac{1}{\beta(x_{<t})} \log \Big\{ \sum_{x_t} Q(x_t|x_{<t}) \exp\Big\{ \beta(x_{<t}) \big[ R(x_t|x_{<t}) + V(x_{\leq t}) \big] \Big\} \Big\}, \quad (14)$$

where we have introduced $V(x_{\leq t}) = \frac{1}{\beta(x_{\leq t})} \log Z(x_{\leq t})$. If $x_t$ is an action variable then $Q(x_t|x_{<t})$ reflects the prior policy and the agent's rationality $\beta(x_{<t})$ determines in how far the value $R(x_t|x_{<t}) + V(x_{\leq t})$ can be optimized by the agent. If $x_t$ is an observation variable then $Q(x_t|x_{<t})$ reflects the agent's prior belief and the rationality of the environment $\beta(x_{<t})$ indicates how much one should deviate from the prior belief considering the possible values $R(x_t|x_{<t}) + V(x_{\leq t})$. Depending on $\beta(x_{<t})$, different decision-making schemes can be recovered—compare Figure 3.

1. **KL control.** When assuming a history-independent loss function $r(x_t)$, Markov probabilities $p_0(x_t|x_{t-1})$ and $\beta(x_{<t}) = \beta$ for all $x_{<t}$, Equation (14) simplifies to a recursion that is equivalent to $z$-iteration which has previously been suggested in [26, 27] to approximately solve MDPs by means

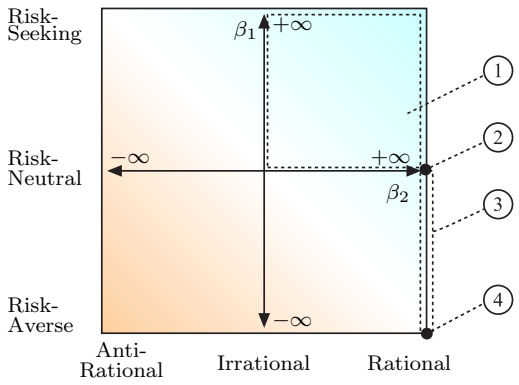


Figure 5: Schematic illustration of how resource parameters can model a range of decision-making schemes: (1)- risk-seeking, bounded rational; (2) risk-neutral, perfectly rational; (3) risk-averse, perfectly rational; and (4) robust, perfectly rational.

of linear algebra—see [28, 29] for details of this equivalence relation. In [26, 27] the transition probabilities of the MDP are controlled directly and the control costs are given by the Kullback-Leiber divergence of the manipulated state transition probabilities with respect to a baseline distribution that describes the passive dynamics. In our framework, this kind of KL control corresponds to the special case where all random variables are action variables and the agent has boundedness parameter $\beta$. The stochasticity in this case, however, is not thought to arise from environmental passive dynamics, but rather is a direct consequence of bounded rational control in a (possibly) deterministic environment. The continuous case of KL control relies on the formalism of path integrals [30, 31], but essentially the same relation to bounded rationality can be established—see [28] for details.

2. **Optimal stochastic control.** When assuming $\beta(x_{<t}) \to \infty$ for all action variables and $\beta(x_{<t}) \to 0$ for all observation variables, we approach the limit of the perfectly rational decision-maker in a stochastic environment. In this limit, the log-partition function converges to the expected utility and the decision-maker acts deterministically so as to maximize the expected utility. For action variables, recursion (14) becomes the well-known *Bellman Optimality Equation* [32]—see [28, 29] for details.

3. **Risk-sensitive control.** Risk-sensitive control [33] corresponds to a decision-maker with $\beta(x_{<t}) \to \infty$ for all action variables and $\beta(x_{<t}) \neq 0$ for observation variables. Risk-sensitivity in the context of continuous KL control has been previously proposed in [34]. Mean-variance deci-

sion criteria used in finance can be equally derived [35]. Risk-sensitive decision-makers do not simply maximize the expectation of the utility, but also consider higher-order cumulants by optimizing a *stress function* given by the log partition sum. A risk-averse decision-maker ($\beta(x_{<t}) < 0$), for example, discounts variability off the expected utility. In contrast, risk-seeking decision-makers ($\beta(x_{<t}) > 0$) add value to the expected utility in the face of variability. Risk-sensitivity biases the beliefs about the environment optimistically (collaborative environment) or pessimistically (adversarial environment). Alternatively, one could regard a collaborative environment also as a bounded rational controller that can choose its own observation—that is the environment behaves like an extension of the agent with partial control. Importantly, the stress function is typically assumed in risk-sensitive control schemes in the literature, whereas here it falls out naturally—see [29] for more details.

4. **Robust control.** When assuming $\beta(x_{<t}) \to \infty$ for all action variables and $\beta(x_{<t}) \to -\infty$ for all observation variables, we approach the limit of the robust decision-maker in an unknown environment. When $\beta(x_{<t}) \to -\infty$, the decision-maker makes a worst case assumption about the environment, namely that it is strictly adversarial and perfectly rational. This leads to the well-known game-theoretic minimax problem. Minimax problems have been used to reformulate robust control problems that allow controllers to cope with model uncertainties [36, 37]. Robust control problems have long been known to be related to risk-sensitive control [38, 39]. Here we derived both control types from the same variational principle—see [29] for more details.

# 7 Discussion

In the proposed thermodynamic interpretation of bounded rationality, agents with limited resources search for a maximum over a set by randomly drawing elements from this set. This random search leads to a utility function that is marginally decreasing when more search effort is allocated. When such agents pay a search cost, the bounded rational optimum is to abort the search as soon as the marginal returns are equal to the search cost. The resulting trade-off between utility maximization and resource costs can be quantified by the Kullback-Leibler divergence with respect to an initial policy or belief. This initial probability distribution corresponds to the initial state of a thermodynamic system that changes when a new potential is imposed. The difference in the potential corresponds to utility gains or losses in economic choice. The difference in the free energy corresponds to physical work and the economic certainty-equivalent. Thus, gains or losses that are associated with uncertainty are effectively devalued or overvalued, depending on the sign of the resource parameter. This way risk-sensitivity, robustness to model uncertainty and game-theoretic minimax-strategies can arise naturally.

**Bounded rationality.** Starting with Simon [23], bounded rationality has been extensively studied in psychology, economics, political science, industrial organization, computer science and artificial intelligence research—see for example [40, 41, 42, 43, 10, 11, 44, 45]. Additionally, numerous experiments in behavioral economics have shown that humans systematically violate perfect rationality, that is they are bounded rational [46]. Probably the most closely related approach to bounded rationality with respect to the present article is quantal response equilibrium (QRE) game theory [47, 48, 49, 50]. QRE models assume bounded rational players whose choice probabilities are given by the Boltzmann distribution and whose rationality is determined by a temperature parameter. Interactions of such bounded rational players can lead to game-theoretic solutions that deviate from the Nash equilibrium. The QRE model is a special case of the model presented here where all prior probabilities are assumed to be uniform. These prior probabilities are crucial when defining the certainty-equivalent that ranges from minimum to maximum via the expected utility. As the certainty-equivalent corresponds to physical work, this also allows to relate bounded rational decision-making to thermodynamic processes. The distinction of a prior policy and a utility that is optimized to some extent is fundamental to the notion of bounded rationality proposed in this paper and therefore also affords a qualitative advance of the bounded rationality model in QRE models.

**Information theory in control and game theory.** As already discussed, a number of papers have suggested the use of the relative entropy as a cost function for control [26, 27, 51, 52]. Previously, Saridis [53] has framed optimal and adaptive control as entropy minimization problems. Statistical physics has also served as an inspiration to a number of other studies, for example, to an information-theoretic approach to interactive learning [54], to use information theory to approximate joint strategies in games with bounded rational players [55] and to the problem of optimal search [56, 57], where the utility losses correspond directly to search effort. Recently, Tishby and Polani [58] have shown how to apply information theory to understand information flow in the action-observation cycle. The contribution of our study is to devise information-theoretic axioms to quantify search costs in bounded optimization problems. This allows for a unified treatment of control and game-theoretic problems, as well as estimation and learning problems for both perfectly rational and bounded rational agents. In the future it will be interesting to relate the thermodynamic resource costs of bounded rational agents to more traditional notions of resource costs in computer science like space and time requirements of algorithms [59].

**Variational Preferences.** In the economic literature the Kullback-Leibler divergence has appeared in the context of multiplier preference models that can deal with model uncertainty [37]. Especially, it has been proposed that a bound on the Kullback-Leibler divergence could be used to indicate how much of a deviation from a proposed model $p_0$ is allowed when computing robust decision

strategies that work under a range of models in the neighborhood of $p_0$. In variational preference models [60] this is generalized to models of the form

$$f \succeq g \Longleftrightarrow \min_p \left( \int u(f)dp + c(p) \right) \geq \min_p \left( \int u(g)dp + c(p) \right),$$

where $c(p)$ can be interpreted as an ambiguity index that can explain effects of ambiguity aversion. The thermodynamic certainty-equivalent of work—computed as the log-partition sum—also falls within this preference model. However, an important difference is that the choice in a thermodynamic system is not deterministic with respect to the certainty-equivalent, but stochastic following a generalized Boltzmann distribution. Due to this stochasticity of the choice behavior itself, the thermodynamic model can be linked to both bounded rationality and model uncertainty, whereas variational preference models have so far concentrated on explaining effects of ambiguity aversion and model uncertainty.

**Ellsberg's and Allais' paradox.** Two of the most famous deviations from expected utility theory that have been consistently observed in human decision-making are the paradoxa of Ellsberg [61] and Allais [62]. While the first paradox has encouraged a large literature dealing with model uncertainty [37], the latter paradox has led to the development of prospect theory [63, 64]. Ellsberg could show that human choice in the face of ambiguity differs from decision-making under risk where precise probability models are available. Humans typically tend to avoid ambiguous options, rather than choosing the option with higher expected utility. The observed ambiguity aversion can be modeled straightforwardly by a bounded rational decision-maker by allowing some degree of minimaxing in the spirit of a risk-sensitive controller—see Supplementary Material for details. Allais could show that humans frequently reverse their preferences in choice tasks that may not lead to preference reversals according to expected utility theory. These reversals typically occur for different levels of riskiness of the same choices. The explanation of the Allais paradox within the framework of bounded rationality is not as straightforward as the Ellsberg paradox, but may involve context-dependent changes of the boundedness parameter or biases in the decision-making process that lead to a *generalized quasi-linear mean model* [65, 66, 67, 68], which provides an alternative account of preference reversals of the Allais type without violating the principle of stochastic dominance—see Supplementary Material for more details.

**Stochastic Choice.** Stochastic choice rules have been extensively studied in the psychological and econometric literature, in particular logit choice models based on the Boltzmann distribution [69, 70]. The literature on Boltzmann distributions for decision-making goes back to Luce [71], extending through McFadden [72], Meginnis [73], Fudenberg [74] and Wolpert [55, 75, 50]. Luce [71] has studied stochastic choice rules of the form $p(x_i) \sim \frac{w_i}{\sum_j w_j}$, which includes the Boltzmann distribution and the "softmax"-rule known in the reinforcement learning literature [76]. McFadden [72] has shown that such distributions can

arise, for example, when utilities are contaminated with additive noise following an extreme value distribution. While stochastic choice models are generally accepted to account for human choices better than their deterministic counterparts [77, 78, 79], they have also been strongly criticized, especially for a property known as *independence of irrelevant alternatives* (IIA). Similar to the independence axiom in expected utility theory, IIA implies that the ratio of two choice probabilities does not depend on the presence of a third irrelevant alternative in the choice set. What distinguishes the free energy equations from above choice rules is that stochastic choice behavior is described by a generalized exponential family distribution of the form $p(x) \sim p_0(x) \exp(\beta U(x))$. Changing the choice set might in general also change the prior $p_0(x)$, but more importantly it might also change the resource parameter $\beta$.

**Diffusion-to-bound models.** Diffusion-to-bound models typically model the process of binary decision-making as a random walk process that terminates once it hits one of two given decision bounds [80]. Each time step of the random walk provides noisy evidence towards one of the two options. This implements a natural speed-accuracy trade-off: the further away the bounds the more reliable the decision will be, as the noise can be averaged out, but also the longer one has to wait. The resulting choice probabilities are identical to the choice probabilities of a bounded rational decision-maker if we relate the decision bound of the random walk with the boundedness parameter in 7—see Supplementary Material for details. The boundedness parameter can then also be shown to be proportional to the time required for the decision-making process. Decision-to-bound models have been widely used in behavioral psychology and neuroscience to explain probabilistic choice and reaction times in psychometric experiments—see [81] for a review. Decision-makers that apply the decision-to-bound model may be regarded as bounded rational decision-makers from a normative point of view.

**Free Energy Principle.** A central property of closed thermodynamic systems is that they minimize free energy. A free energy principle based on the variational Bayes approach has recently also been proposed as a theoretical framework to understand brain function [82, 83]. In this framework generative models of the form $p(y|h, a)$ explain how hidden causes $h$ in the environment and actions $a$ produce observations $y$. The brain uses an approximative distribution $Q(h; a)$ to determine the hidden causes. The free energy

$$F = -\int dh\, Q(h; a) \ln P(y, h|a) - \int dh\, Q(h; a) \ln Q(h; a)$$

measures how well the brain is doing with this approximation. According to [82, 83], action and perception consist in choosing $a$ and $Q$ respectively so as to minimize this free energy. In light of the thermodynamic view of free energy, maximizing the likelihood $-\ln P(y, h|a)$—or minimizing surprise—is a particular choice of potential function $\phi$, where the boundedness consists in being

restricted to model class $Q$ instead of having full disposal of $p(y|h,a)$. More generally, variational Bayes methods that use particular classes of distributions to approximate the posterior could thus be regarded as a form of bounded inference within this picture.

# 8 Conclusion

Thermodynamics provides a framework for bounded rationality that can be both descriptive and prescriptive. It is descriptive in the sense that it describes behavior that is clearly sub-optimal from the point of view of a perfect rational decision-maker with infinite resources. It is prescriptive in the sense that it prescribes how a bounded rational actor should behave optimally given resource constraints formalized by $\beta$. As we have argued in this paper, bounded rational decision-making provides an overarching principle in both senses in economics, engineering, artificial intelligence, psychology and neuroscience.

A thermodynamic model of bounded rational decision-making has also two advantages over traditional decision theory of perfect rationality. First, it allows connecting computational processes with real physical processes, for example how much entropy they generate and how much energy they require [13]. Second, it suggests a notion of intelligence that is closely related to the process of evolution. It is straightforward to see that bounded rational controllers of the form (9) share their structure with Bayes' rule, where we identify the prior $p_0(x)$, the likelihood model $e^{\beta U(x)}$ and the posterior $p(x)$, normalized by the partition function, thus, establishing a close link between inference and control [84]. Furthermore, both bounded rational controllers and Bayes' rule share their structural form with discrete replicator dynamics that model evolutionary processes [85], where samples (a population) are pushed through a fitness function (likelihood, gain function) that biases the distribution of the population, thereby transforming a distribution $p_0$ to a new distribution $p$. In this picture different hypotheses $x$ compete for probability mass over subsequent iterations, favoring those $x$ that have a lower-than-average cost. Just like the evolutionary random processes of variation and selection created intelligent organisms on a phylogenetic timescale, similar random processes might underlie (bounded) intelligent behavior in individuals on an ontogenetic timescale.

# 9 Acknowledgments

This study was supported by the DFG, Emmy Noether grant BR4164/1-1.

# References

[1] Gigerenzer G, Todd PM, ABC Research Group. Simple Heuristics That Make Us Smart. New York: Oxford University Press; 1999.

[2] Gladwell M. Blink: the power of thinking without thinking. New York: Little, Brown and Company; 2005.

[3] Niv Y, Daw ND, Joel D, Dayan P. Tonic dopamine: opportunity costs and the control of response vigor. Psychopharmacology (Berl). 2007 Apr;191(3):507–520.

[4] Daw ND. 'Model-based reinforcement learning as cognitive search: neurocomputational theories'. In: Cognitive search: evolution algorithms and the brain. Boston: MIT Press; 2012. .

[5] Von Neumann J, Morgenstern O. Theory of Games and Economic Behavior. Princeton: Princeton University Press; 1944.

[6] Savage LJ. The Foundations of Statistics. New York: John Wiley and Sons; 1954.

[7] Fishburn P. The Foundations of Expected Utility. Dordrecht: D. Reidel Publishing; 1982.

[8] Simon H. Theories of Bounded Rationality. In: Radner CB, Radner R, editors. Decision and Organization. Amsterdam: North Holland Publ.; 1972. p. 161–176.

[9] Simon H. Models of Bounded Rationality. Cambridge, MA: MIT Press; 1984.

[10] Rubinstein A. Modeling Bounded Rationality. Cambridge, MA: MIT Press; 1999.

[11] Gigerenzer G, Selten R. In: Bounded rationality: the adaptive toolbox. Cambridge, MA: MIT Press; 2001. .

[12] Feynman RP. The Feynman Lectures on Computation. Addison-Wesley; 1996.

[13] Tribus M, McIrvine EC. Energy and Information. Scientific American. 1971;225:179–188.

[14] Landauer R. Irreversibility and Heat Generation in the Computing Process. IBM Journal of Research and Development. 1961;5(3):183–191.

[15] Bennett CH. Logical Reversibility of Computation. IBM Journal of Research and Development. 1973;17(6):525–532.

[16] Bennett CH. The thermodynamics of computationa review. International Journal of Theoretical Physics. 1982;21(12):905–940.

[17] MacKay DJC. Information Theory, Inference, and Learning Algorithms. Cambridge University Press; 2003.

[18] Ortega PA, Braun DA. A conversion between utility and information. In: Proceedings of the third conference on artificial general intelligence. Atlantis Press; 2010. p. 115–120.

[19] Ortega PA. A Unified Framework for Resource-Bounded Autonomous Agents Interacting with Unknown Environments. Department of Engineering, University of Cambridge, UK; 2011.

[20] Shannon CE. A mathematical theory of communication. Bell System Technical Journal. 1948 Jul and Oct;27:379–423 and 623–656.

[21] Csiszár I. Axiomatic Characterizations of Information Measures. Entropy. 2008;10:261–273.

[22] Callen HB. Thermodynamics and an introduction to thermostatistics. New York: John Wiley & Sons; 1985.

[23] Simon HA. Rational choice and the structure of the environment. Psychological Review. 1956;63(2):129–38.

[24] Gumbel EJ. Statistics of Extremes. New York: Columbia University Press; 1958.

[25] Ortega PA, A BD. Free Energy and the Generalized Optimality Equations for Sequential Decision Making. arXiv:12053997v1 (presented at the European Workshop for Reinforcement Learning). 2012;.

[26] Todorov E. Linearly solvable Markov decision problems. In: Advances in Neural Information Processing Systems. vol. 19; 2006. p. 1369–1376.

[27] Todorov E. Efficient computation of optimal actions. Proceedings of the National Academy of Sciences USA. 2009;106:11478–11483.

[28] Braun DA, Ortega PA. Path integral control and bounded rationality. In: IEEE Symposium on adaptive dynamic programming and reinforcement learning; 2011. p. 202–209.

[29] Ortega PA, Braun DA. Information, utility and bounded rationality. In: Lecture notes on artificial intelligence. vol. 6830; 2011. p. 269–274.

[30] Kappen HJ. A linear theory for control of non-linear stochastic systems. Physical Review Letters. 2005;95:200201.

[31] Theodorou E, Buchli J, Schaal S. A generalized path integral approach to reinforcement learning. Journal of Machine Learning Research. 2010;11:3137–3181.

[32] Bellman RE. Dynamic Programming. Princeton, NJ: Princeton University Press; 1957.

[33] Whittle P. Risk-sensitive optimal control. New York: John Wiley and Sons; 1990.

[34] van den Broek JL, Wiegerinck WAJJ, Kappen HJ. Risk-sensitive path integral control. In: UAI. vol. 6; 2010. p. 1–8.

[35] Markowitz H. Portfolio Selection. The Journal of Finance. 1952;7:77–91.

[36] Başar T, Bernhard P. H-infinity optimal control and related minimax-design problems: a dynamic game approach. Boston: Birkhäuser; 1991.

[37] Hansen LP, Sargent TJ. Robustness. Princeton: Princeton University Press; 2008.

[38] Jacobson D. Optimal stochastic linear systems with exponential performance criteria and their relation to deterministic differential games. IEEE T Automat Contr AC. 1973;18:124–131.

[39] Glover K, Boyle J. State-space formulae for all stabilizing controllers that satisfy an H-norm bound and relations to risk sensitivity. Syst Control Lett. 1988;11:167–172.

[40] Lipman B. Information Processing and Bounded Rationality: A Survey. Canadian Journal of Economics. 1995;28(1):42–67.

[41] Russell SJ. Rationality and Intelligence. In: Mellish C, editor. Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence. San Francisco: Morgan Kaufmann; 1995. p. 950–957.

[42] Russell SJ, Subramanian D. Provably bounded-optimal agents. Journal of Artificial Intelligence Research. 1995;3:575–609.

[43] Aumann RJ. Rationality and Bounded Rationality. Games and Economic Behavior. 1997 October;21(1-2):2–14.

[44] Kahneman D. Maps of Bounded Rationality: Psychology for Behavioral Economics. American Economic Review. 2003 December;93(5):1449–1475.

[45] Spiegler R. Bounded Rationality and Industrial Organization. Oxford: Oxford University Press; 2011.

[46] Camerer C. Behavioral Game Theory: Experiments in Strategic Interaction. Princeton: Princeton University Press; 2003.

[47] D MR, R PT. Quantal Response Equilibria for Normal Form Games. Games and Economic Behavior. 1995 July;10(1):6–38.

[48] Mckelvey R, Palfrey TR. Quantal Response Equilibria for Extensive Form Games. Experimental Economics. 1998;1:9–41.

[49] Anderson SP, Goeree JK, Holt CA. The logit equilibrium: a perspective on intuitive behavioral anomalies. Southern Economic Journal. 2002;69:21–47.

[50] Wolpert DH, Harre M, Bertschinger N, Olbrich E, Jost J. Hysteresis effects of changing parameters of noncooperative games. Physical Review E. 2012;85:036102.

[51] Peters J, Mülling K, Altun Y. Relative entropy policy search. In: AAAI; 2010. .

[52] Kappen HJ, Gómez V, Opper M. Optimal control as a graphical model inference problem. Machine Learning. 2012;1:1–11.

[53] Saridis G. Entropy formulation for optimal and adaptive control. IEEE Transactions on Automatic Control. 1988;33:713–721.

[54] Still S. An information-theoretic approach to interactive learning. Europhysics Letters. 2009;85:28005.

[55] Wolpert DH. Information theory - the bridge connecting bounded rational game theory and statistical physics. In: Braha D, Bar-Yam Y, editors. Complex Engineering Systems. Perseus Books; 2004. .

[56] Stone LD. Theory of optimal search. New York: Academic Press; 1998.

[57] Jaynes ET. Entropy and search theory. In: Smith CR, Grandy WT, editors. Maximum entropy and Bayesian methods in inverse problems. Dordrecht: Reidel; 1985. .

[58] Tishby N, Polani D. Information Theory of Decisions and Actions. In: Vassilis T Hussain, editor. Perception-reason-action cycle: Models, algorithms and systems. Berlin: Springer; 2011. .

[59] Vitanyi PMB. Time, space, and energy in reversible computing. In: Proceedings of the 2nd ACM conference on Computing frontiers; 2005. p. 435–444.

[60] Maccheroni F, Marinacci M, Rustichini A. Ambiguity aversion, robustness, and the variational representation of preferences. Econometrica. 2006;74:1447–1498.

[61] Ellsberg D. Risk, Ambiguity and the Savage Axioms. The Quaterly Journal of Economics. 1961;75:643–669.

[62] Allais M. Le comportment de l'homme rationnel devant la risque: critique des postulats et axiomes de l'ecole Americaine. Econometrica. 1953;21:503–546.

[63] Kahneman D, Tversky A. Prospect Theory: An Analysis of Decision under Risk. Econometrica. 1979;47:263–291.

[64] Tversky A, Kahneman D. Advances in prospect theory: Cumulative representation of uncertainty". Journal of Risk and Uncertainty. 1992;5:297–323.

[65] Nagumo M. Über eine Klasse der Mittelwerte. Japan Journal of Mathematics. 1930;7:71–79.

[66] Kolmogorov A. Sur la notion de la moyenne. Rendiconti accademia dei lincei. 1930;12:388–391.

[67] de Finetti B. Sul concetto di media. Giornale dell' istituto italiano degli attuari. 1931;2:369–396.

[68] Hong CS. A generalization of the quasilinear mean with application to the measurement of income inequality and decision theory resolving the Allais paradox. Econometrica. 1983;51:1065–1092.

[69] Luce RD. Utility of gains and losses: measurement-theoretical and experimental approaches. Mahwah, NJ: Erlbaum; 2000.

[70] Train KE. Discrete Choice Methods with Simulation. 2nd ed. Cambridge: Cambridge University Press; 2009.

[71] Luce RD. Individual choice behavior. Oxford: Wiley; 1959.

[72] McFadden D. Conditional logit analysis of qualitative choice behavior. In: Zarembka P, editor. Frontiers in econometrics. New York: Academic Press; 1974. .

[73] Meginnis JR. A new class of symmetric utility rules for gambles, subjective marginal probability functions, and a generalized Bayes rule. Proceedings of the American Statistical Association, Business and Economic Statistics Section. 1976;p. 471–476.

[74] Fudenberg D, Kreps D. Learning mixed equilibria. Games and Economic Behavior. 1993;5:320–367.

[75] Lee R, Wolpert DH. Game-Theoretic Modeling of Human Behavior in Mid-Air Collisions. In: T Guy MK, Wolpert DH, editors. Decision-Making with Imperfect Decision Makers. Springer; 2011. .

[76] Sutton RS, Barto AG. Reinforcement Learning: An Introduction. Cambridge, MA: MIT Press; 1998.

[77] Rieskamp J. The probabilistic nature of preferential choice. Journal of Experimental Psychology: Learning, Memory, and Cognition. 2008;34:1446–1465.

[78] Glscher J, Daw N, Dayan P, O'Doherty JP. States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. Neuron. 2010;66(4):585–95.

[79] McDannald MA, Takahashi YK, Lopatina N, Pietras BW, Jones JL, Schoenbaum G. Model-based learning and the contribution of the orbitofrontal cortex to the model-free world. Eur J Neurosci. 2012;35(7):991–6.

[80] Busemeyer JR, Diederich A. Survey of decision field theory. Mathematical Social Sciences. 2002;43(3):345–370.

[81] Bogacz R, Brown E, Moehlis J, Holmes P, Cohen JD. The physics of optimal decision making: a formal analysis of models of performance in two-alternative forced-choice tasks. Psychological Review. 2006;113:700–765.

[82] Friston K. The free-energy principle: a rough guide to the brain? Trends in Cognitive Science. 2009;13:293–301.

[83] Friston K. The free-energy principle: a unified brain theory? Nature Review Neuroscience. 2010;11:127–138.

[84] Ortega PA, Braun DA. A minimum relative entropy principle for learning and acting. Journal of Artificial Intelligence Research. 2010;38:475–511.

[85] Shahlizi CR. Dynamics of bayesian updating with dependent data and misspecified models. Electronic Journal of Statistics. 2009;3:1039–1074.