PROCEEDINGS OF THE ROYAL SOCIETY A

rspa.royalsocietypublishing.org

Research

CrossMark click for updates

Cite this article: Ortega PA, Braun DA. 2013 Thermodynamics as a theory of decision-making with information-processing costs. Proc R Soc A 469: 20120683. http://dx.doi.org/10.1098/rspa.2012.0683

Received: 14 November 2012
Accepted: 6 February 2013

Subject Areas:
computational biology, biomedical engineering, biophysics

Keywords:
decision-making, bounded rationality, information processing

Author for correspondence:
Daniel A. Braun
e-mail: daniel.braun@tuebingen.mpg.de

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

Pedro A. Ortega[1,2] and Daniel A. Braun[1,2]

[1]Max Planck Institute for Biological Cybernetic, and [2]Max Planck Institute for Intelligent Systems, Speemanstrasse 38, Tübingen 72076, Germany

Perfectly rational decision-makers maximize expected utility, but crucially ignore the resource costs incurred when determining optimal actions. Here, we propose a thermodynamically inspired formalization of bounded rational decision-making where information processing is modelled as state changes in thermodynamic systems that can be quantified by differences in free energy. By optimizing a free energy, bounded rational decision-makers trade off expected utility gains and information-processing costs measured by the relative entropy. As a result, the bounded rational decision-making problem can be rephrased in terms of well-known variational principles from statistical physics. In the limit when computational costs are ignored, the maximum expected utility principle is recovered. We discuss links to existing decision-making frameworks and applications to human decision-making experiments that are at odds with expected utility theory. Since most of the mathematical machinery can be borrowed from statistical physics, the main contribution is to re-interpret the formalism of thermodynamic free-energy differences in terms of bounded rational decision-making and to discuss its relationship to human decision-making experiments.

1. Introduction

In everyday life, decision-makers often have to make fast and frugal choices that are constrained by limited resources such as time, money, food, knowledge or computational effort [1–4]. Classic theories of decision-making generally ignore information-processing constraints by assuming that *perfectly rational* 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 [5–7]. Unlike

Royal Society Publishing
Informing the science of the future

perfectly rational decision-makers, *bounded rational* decision-makers are subject to limited information-processing resources. Starting with Simon [8–10], bounded rationality has extensively been studied in psychology, economics, political science, industrial organization, computer science and artificial intelligence research [11–18]. Numerous experiments in behavioural economics have shown that humans are bounded rational and systematically deviate from perfect rationality [19]. Here, we develop a thermodynamic model of bounded rational decision-making that can explain some of these deviations.

Previous attempts to apply thermodynamics and statistical physics to the problem of bounded rational decision-making [20–23] have focused on the Boltzmann distribution, thereby stipulating an analogy between the concepts of energy and utility: just like physical systems tend to pick states with low energy, decision-makers tend to pick states with high utility. Being perfectly rational then corresponds to physical systems with zero temperature, in which all probability mass is concentrated on the lowest energy state. In particular, quantal response equilibrium (QRE) models of bounded rationality typically assume bounded rational players whose choice probabilities are given by the Boltzmann distribution and whose rationality is determined by a temperature parameter [20,21]. Boltzmann-like stochastic choice rules have also been extensively studied in the psychological and econometric literature [24,25], in particular in the form of logit choice models going back to Luce [26], McFadden [27], Meginnis [28] and Fudenberg & Kreps [29]. In the machine learning and reinforcement learning literature [30], Boltzmann-like choice rules are known as *softmax* rules and used for stochastic sampling of actions in the context of the exploration–exploitation dilemma.

In statistical physics, it is well known that the Boltzmann distribution satisfies a variational principle in the *free energy* $F = U - TS$, which instantiates a trade-off between the internal energy U and the entropic cost S [31]. These two terms have been previously related to utility and information-processing costs in QRE models of bounded rational decision-making [20–23]. In this article, we generalize these previous models of bounded rationality based on the duality between information and utility [32–34]: instead of considering absolute free energies F, we consider differences in free energy ΔF between an initial state and a final state corresponding to the situation before and after the deliberation associated with the decision-making process. Considering energy differences rather than absolute energies is not only physically meaningful, but it also accounts for the fact that human decision-makers have been shown to consider *changes* in value rather than absolute value, which is one of the cornerstones of prospect theory [35]. We will show that this seemingly innocuous extension leads to a substantial generalization that allows definition of a certainty-equivalent concept that is closely related to the physical concept of work. The simple Boltzmann distribution is still contained as a special case in the general class of exponential family distributions that satisfy a generalized variational principle in the *free-energy difference*. Intriguingly, this variational principle can be applied to both action and perception. As special cases, it allows not only the derivation of a number of decision-making frameworks—including expected utility theory [5–7]—but also the formulation of a variational principle for (approximate) Bayesian inference, which has recently been suggested to underlie self-organizing systems [36,37].

Below we will first expand the thermodynamic intuition, then relate the thermodynamic quantities to decision-making variables and show how to apply this decision-making framework to actual decision-making experiments.

2. Thermodynamic intuition

In the following, we conceive information processing as changes in information states, assuming that information states can be represented as probability distributions. The archetypical example is the updating of beliefs from observations in the shape of Bayesian inference: the initial information state is given by a prior distribution, which is transformed by applying a likelihood function to a posterior distribution representing the final information state [38]. Importantly, the

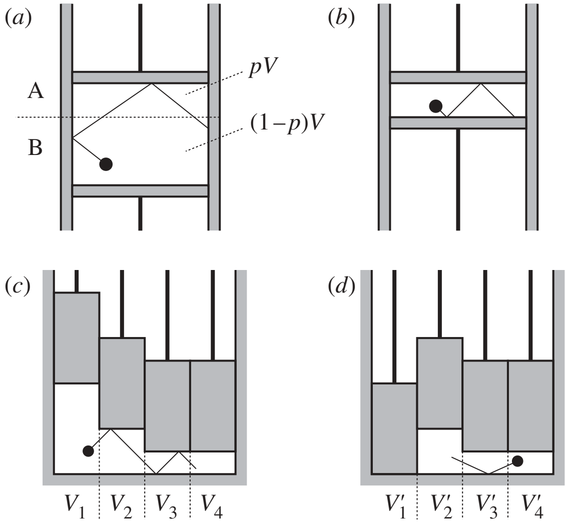

Figure 1. The molecule-in-a-box device. (*a*) Initially, a molecule moves freely within a space of volume *V* delimited by two pistons. The compartments A and B correspond to the two logical states of the device. (*b*) Representing a different distribution with reduced uncertainty over the two states requires work by the pistons. (*c*) A molecule in a box with multiple compartments. (*d*) When changing the individual compartment volumes, the free-energy difference in each compartment can be related to the total work by equation (2.2).

same idea can also be applied to the case of acting. The initial information state then corresponds to a prior or default policy before deliberation takes place, for example, given by a uniform distribution over the set of available actions. Applying a likelihood function corresponds to a deliberation process that concentrates probability mass on desirable actions. The final information state corresponds to the posterior distribution over actions, and includes the delta function over the single best action as a special case.

Bounded rationality in the case of acting comes into play when only a certain amount of change in information state can be afforded. To quantify the cost of changing information states, we can employ a thermodynamic illustration (figure 1*a*,*b*). Assume that we want to physically represent a probability distribution by means of an ideal gas particle in a box of volume V with diathermal walls immersed in a heat bath at constant temperature T. Assume further that we want to update our information state to a new distribution over a reduced volume $V' = cV$. This transformation requires physical work given by the free-energy difference

$$\Delta F = -\int_{V}^{V'} \frac{kT}{V}\,\mathrm{d}V = kT\ln c, \tag{2.1}$$

where $k > 0$ is the Boltzmann constant and kT can be interpreted as the conversion factor between one unit of information and one unit of energy for a molecule-in-a-box device [39–43]. Depending on what physical system we use to represent the distributions, this conversion factor could be higher, thus, making information processing more expensive.

While this toy example illustrates the cost of changing information states, it hides the dependence on the underlying information structure given by the partitioning of possible microstates. Therefore, we consider a modified example by introducing multiple compartments—see figure 1*c*,*d*. Let each compartment x have an initial volume $V(x)$ that is changed to $V'(x)$ such that $\sum_x V(x) = V$ and $\sum_x V'(x) = V'$. The initial probability of being in state x is $p_0(x) = V(x)/V$ and the final probability is $p(x) = V'(x)/V'$. The free-energy difference can then be expressed as

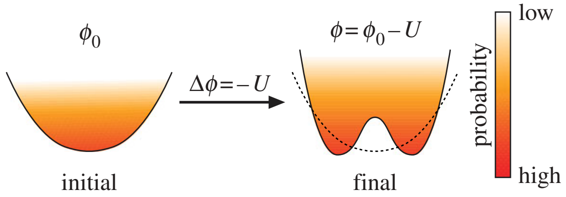

Figure 2. Representing a decision-maker as a thermodynamic system, the behaviour of the decision-maker exposed to a gain $\Delta\phi = -U$ can be expressed as a change of initial cost potential ϕ_0 to a final cost potential ϕ, where $\phi = \phi_0 - U$. The choice or belief probabilities of the decision-maker change according to (3.2) from p_0 to p. (Online version in colour.)

follows:

$$\begin{aligned}\Delta F &= kT \ln \frac{V}{V'}\\ &= kT \sum_x \frac{V'(x)}{V'} \ln \left\{ \frac{V}{V'} \frac{V(x)}{V(x)} \frac{V'(x)}{V'(x)} \right\}\\ &= \sum_x \frac{V'(x)}{V'} \left\{ kT \ln \frac{V(x)}{V'(x)} \right\} + kT \frac{V'(x)}{V'} \ln \left\{ \frac{V'(x)}{V'} \Big/ \frac{V(x)}{V} \right\}\\ &= \mathbb{E}_p[\Delta F(x)] + kT \quad \mathbb{D}_{\mathrm{KL}}(p \,\|\, p_0),\end{aligned} \tag{2.2}$$

where $\Delta F(x)$ represents the free-energy difference of compartment x. Importantly, it can be seen that the free-energy difference consists of two terms: the averaged free-energy contributions of the individual compartments and a cost term that measures the information-theoretic distance between the initial and final information state, which is then converted into units of energy.

(a) The free-energy difference

As illustrated in figure 2, the isothermal transformations discussed in the two examples can be characterized in a general and abstract manner as follows.

- An initial information state is given by a prior distribution $p_0(x)$ over some states $x \in \mathcal{X}$. Each state x has associated to it an initial energy potential $\phi_0(x)$ such that $p_0(x) = (1/Z_0)\,\mathrm{e}^{-\alpha\phi_0(x)}$ with $Z_0 = \sum_{x'} \mathrm{e}^{-\alpha\phi_0(x')}$.
- A transformation is applied by adding a new potential $\Delta\phi(x)$.
- The final information state is represented by a distribution $q(x)$ under the constraint of the combined potential $\phi(x) = \phi_0(x) + \Delta\phi(x)$.

The parameter α indicates the inverse temperature. The free-energy difference between the initial and final information state can be computed as follows:

$$\begin{aligned}\Delta F[q] &= F[q] - F[p_0]\\ &= \sum_{x\in\mathcal{X}} q(x)(\phi_0(x) + \Delta\phi(x)) + \frac{1}{\alpha} \sum_{x\in\mathcal{X}} q(x) \ln q(x) - F[p_0]\\ &= \sum_{x\in\mathcal{X}} q(x) \left(-\frac{1}{\alpha} \ln Z_0 - \frac{1}{\alpha} \ln p_0(x) + \Delta\phi(x) \right) + \frac{1}{\alpha} \sum_{x\in\mathcal{X}} q(x) \ln q(x) - F[p_0]\\ &= \sum_{x\in\mathcal{X}} q(x)\Delta\phi(x) + \frac{1}{\alpha} \sum_{x\in\mathcal{X}} q(x) \ln \frac{q(x)}{p_0(x)},\end{aligned}$$

where we have used the equality $F[p_0] = -(1/\alpha)\ln Z_0$. As described in standard textbooks on statistical physics [31], the free energy $F[q]$ obtains its minimum for the equilibrium distribution $p(x) = (1/Z_\phi)\,\mathrm{e}^{-\alpha\phi(x)}$ with $Z_\phi = \sum_{x'} \mathrm{e}^{-\alpha\phi(x')}$. Since $\Delta F[q]$ and $F[q]$ differ only by a constant, the same minimum is obtained in the free-energy difference. However, the free-energy difference allows us to write the equilibrium distribution in the equivalent form $p(x) = (1/Z)p_0(x)\,\mathrm{e}^{-\alpha\Delta\phi(x)}$ with $Z = \sum_{x'} p_0(x')\,\mathrm{e}^{-\alpha\Delta\phi(x')}$ to separate the two essential ingredients: the prior information state $p_0(x)$ and the transforming potential difference $\Delta\phi(x)$. Accordingly, the free-energy difference can always be thought of as a trade-off between the energy $\mathbb{E}_q[\Delta\phi(x)]$ and the information divergence measured by $\mathbb{D}_{\mathrm{KL}}(q\|p_0)$. In §3, we are going to exploit this trade-off to model bounded rational decision-makers that reason about utility gains $U(x) := -\Delta\phi(x)$.

3. Thermodynamics of decision-making

In decision theory, preferences between alternative outcomes $x \in \mathcal{X}$ are usually represented by a real-valued function U over $\mathcal{X}$ called the utility function. Among other things, this requires that the preferences between any two elements of $\mathcal{X}$ can be established and that these preferences are stable and transitive [5–7]. Given a choice over the whole set $\mathcal{X}$, a perfectly rational decision-maker will consistently choose the best outcome $x^* = \arg\max_x U(x)$, presupposing that such a unique optimum exists. However, if the set $\mathcal{X}$ is very large and the available resources to search this set are very limited, then it might not always be possible to find the best option for a *bounded rational* decision-maker.

As a specific example of such a search, consider drawing balls labelled by numbers U from an urn with replacement such that the drawing process can be described by a probability $\mu(U)$. After drawing once, twice, thrice, etc., we observe a time series $U_1, U_2, U_3, \ldots$ of independent and identically distributed data, and our task is to find the maximum by keeping track of the largest number seen so far, which after m draws is $v := \max\{U_1, U_2, \ldots, U_m\}$. The cumulative distribution function of choosing v after m draws is given by $F_m(v) = F_0(v)^m$, where F_0 is the cumulative distribution function of μ [44]. In the continuous limit, the associated density is given by $p_m(v) = (\mathrm{d}/\mathrm{d}v)F_0(v)^m$. Similar to other evidence accumulation schemes [45], we can then compute the log odds between any two random outcomes v and v' as

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

where $p_m(v)$ can be rewritten as an exponential family distribution $p_m(v) \propto \mu(v)\,\mathrm{e}^{\beta V(v)}$ with energy levels $V(v) = \log F_0(v)$ and the parameter $\beta = m - 1$. As β plays the role of the number of draws m, it might be considered as a measure of the information-processing resource. Accordingly, the more resources β a decision-maker spends, the more he resembles a perfectly rational decision-maker that chooses the maximum without fail, whereas for any finite number of β, some uncertainty about the maximum remains.

In general, the boundedness parameter β can be thought of as a Lagrange multiplier in a constrained optimization problem. In QRE models [20–23], the Lagrange multiplier is used to constrain the entropy or mean cost, whereas here we apply it to express a constraint on the relative entropy or Kullback–Leibler (KL) divergence. By replacing the thermodynamic energy potential $\Delta\phi(x)$ of §2*a* with the economic utility gain $U(x) = -\Delta\phi(x)$, we can formulate the following variational principle in the *negative free-energy difference*.

Variational principle. Given an initial information state $p_0(x)$, a resource parameter β and a utility gain $U(x)$, the negative free-energy difference $-\Delta F = F_0 - F$ between the initial and final information state

$$-\Delta F[q] := \sum_{x\in\mathcal{X}} q(x)U(x) - \frac{1}{\beta}\sum_{x\in\mathcal{X}} q(x)\log\frac{q(x)}{p_0(x)} \tag{3.1}$$

is maximized by the equilibrium distribution $q(x) = p(x)$, where

$$p(x) = \frac{1}{Z}p_0(x)\,\mathrm{e}^{\beta U(x)} \quad \text{with} \quad Z := \sum_{x\in\mathcal{X}} p_0(x)\,\mathrm{e}^{\beta U(x)}, \tag{3.2}$$

such that

$$-\Delta F[q] \leq -\Delta F[p] = \frac{1}{\beta} \log Z. \tag{3.3}$$

Just like its physical pendant [31], this variational principle can be regarded in two ways: (i) as a minimum (relative) entropy principle where the expected utility gain is fixed or (ii) as a maximum utility gain principle where the (relative) entropy is fixed. The first interpretation provides a principle for estimation in the context of observer modelling under model uncertainty, where utility gains (or losses) can induce deviations from the probabilistic belief p_0. The second interpretation provides a principle for bounded rational decision-making in the case of actor modelling, where the relative entropy constrains the information-processing capacity of the decision-maker. This naturally leads to a trade-off between utility gains and information-processing costs.

(a) Choice and belief probabilities

In line with the twofold interpretation of the variational principle, the distribution (3.2) can be interpreted both as an action or observation probability. In the case of *actions*, p_0 represents the *a priori* choice probability of the agent that is refined to the choice probability p when evaluating the imposed gain (or loss) U. The allowed change in probability depends on the resource parameter β and corresponds to the search 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 because of the presence of anticipated gains (or losses) U that are evaluated by the holder of the belief. This way, the agent can anticipate a collaborative ($\beta > 0$) or adversarial ($\beta < 0$) environment of (assumed) rationality β. For different values of β, the distribution (3.2) has the following limits:

$$\lim_{\beta\to\infty} p(x) = \delta(x - x^*), \quad x^* = \max_x U(x),$$

$$\lim_{\beta\to 0} p(x) = p_0(x)$$

and

$$\lim_{\beta\to -\infty} p(x) = \delta(x - x^*), \quad x^* = \min_x U(x).$$

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 $U(x)$ or minimizes loss $-U(x)$. The limit $\beta \to 0$ is an actor without resources that simply selects his action according to his prior p_0. 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 in a cooperative environment that can be essentially regarded as an extension of the agent, an extremely pessimistic observer ($\beta \to -\infty$) who anticipates only the worst-case scenario, and a risk-neutral Bayesian observer ($\beta \to 0$) who simply relies on the probabilistic model p_0.

(b) The certainty equivalent

In economics, the certainty equivalent measures the value of a risky gamble in terms of a risk-free equivalent such that a decision-maker would be indifferent between the risky and the risk-free option. In statistical physics [31], the free-energy difference $\Delta F = \Delta E - Q = W$ measures the amount of available work W by subtracting the heat Q from the total change in energy ΔE. The crucial physical intuition is that heat is a form of energy that we have uncertainty about, for example, we do not know the exact trajectories of all gas particles at temperature β. This uncertainty implies that we do not have full control over the objects and consequently cannot make use of all the energy [43]. Work, on the other hand, is a pure form of energy uncontaminated by uncertainty and therefore fully transformable and usable. Economically speaking, the physical

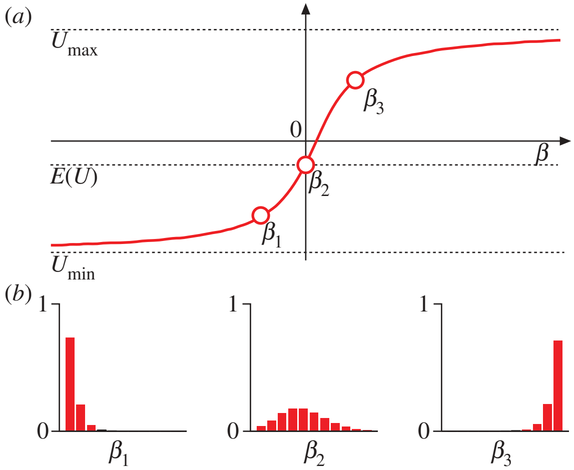

Figure 3. (*a*) Negative free-energy difference $-\Delta F$ versus the resource parameter β. The resource parameter allows modelling decision-makers with bounded resources, either when generating their own actions ($\beta_3 > 0$) or when anticipating an adversarial environment ($\beta_1 < 0$). The negative free-energy difference corresponds to the certainty equivalent. (*b*) Distribution over the outcomes depending on the resource parameter β. For large positive β, the distribution concentrates on the outcome with maximum gain $U_{\max}$. For large negative β, the distribution concentrates on the worst outcome with gain $U_{\min}$. For $\beta_2 = 0$, the outcomes follow the given distribution p_0. (Online version in colour.)

concept of work, and therefore also the difference in free energy, measures the certainty equivalent of an energy difference that is contaminated by uncertainty. In the following, we generalize the use of the free-energy difference to ascribe a certainty-equivalent value to gambles that are characterized by utility gains (or losses) U and an initial information state p_0. As can be seen from (3.3), the free-energy difference between equilibrium distributions is given by the log-partition function, i.e. the logarithm of the normalization constant $Z = \sum_{x'} p_0(x')\, \mathrm{e}^{\beta U(x')}$. For different values of β, the certainty equivalent takes the following limits:

$$\lim_{\beta\to\infty} \frac{1}{\beta} \log Z = \max_x U(x),$$

$$\lim_{\beta\to 0} \frac{1}{\beta} \log Z = \sum_x p_0(x) U(x)$$

and

$$\lim_{\beta\to -\infty} \frac{1}{\beta} \log Z = \min_x U(x).$$

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. For illustration, see figure 3.

(c) Sequential decision-making

The formalism for bounded rational decision-making can also be extended to sequential decision-making where a vector of random variables $x_{\le T} = x_1 x_2 \ldots x_T$ has to be determined by consecutively drawing from distributions $p(x_t \,|\, x_{<t})$ that depend on the history $x_{<t} = x_1 \ldots x_{t-1}$ [46]. Each history $x_{<t}$ defines a decision node with respect to the variable x_t and can be characterized by an initial information state $p_0(x_t \,|\, x_{<t})$, a utility gain $U(x_t \,|\, x_{<t})$ and a Lagrange multiplier $\beta(x_{<t})$. The latter allows assigning node-specific information-processing

resources. The negative free-energy difference $-\Delta F = F_0 - F$ between the initial information state $p_0(x_{\leq T}) = \prod_{t=1}^{T} p_0(x_t \mid x_{<t})$ and the final information state $q(x_{\leq T})$ is then

$$-\Delta F[q] = \sum_{x_{\leq T}} q(x_{\leq T}) \sum_{t=1}^{T} \left\{ U(x_t \mid x_{<t}) - \frac{1}{\beta(x_{<t})} \log \frac{q(x_t \mid x_{<t})}{p_0(x_t \mid x_{<t})} \right\}. \tag{3.4}$$

This negative free-energy difference has a nested structure where the latest time step forms the innermost variational problem that can be solved first, and all other variational problems of the previous time steps can be solved recursively by working backwards in time. This leads to a recursive solution for the equilibrium distribution,

$$p(x_t \mid x_{<t}) = \frac{1}{Z(x_{<t})} p_0(x_t \mid x_{<t}) \exp\left\{ \beta(x_{<t}) \left[U(x_t \mid x_{<t}) + \frac{1}{\beta(x_{\leq t})} \log Z(x_{\leq t}) \right] \right\},$$

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

$$Z(x_{<t}) = \sum_{x_t} p_0(x_t \mid x_{<t}) \exp\left\{ \beta(x_{<t}) \left[U(x_t \mid x_{<t}) + \frac{1}{\beta(x_{\leq t})} \log Z(x_{\leq t}) \right] \right\}, \tag{3.5}$$

with the certainty equivalent

$$V(x_{<t}) = \frac{1}{\beta(x_{<t})} \log \left\{ \sum_{x_t} p_0(x_t | x_{<t}) \exp\{\beta(x_{<t})[U(x_t | x_{<t}) + V(x_{\leq t})]\} \right\}, \tag{3.6}$$

such that $V(\cdot) = (1/\beta(\cdot)) \log Z(\cdot)$. Depending on how the resource parameters $\beta(x_{<t})$ are set, a range of different decision-making schemes can be recovered.

- *KL control.* For state-dependent loss functions $U(x_t)$, Markov probabilities $p_0(x_t \mid x_{t-1})$ and uniform $\beta(x_{<t}) = \beta$, the KL-control framework [47–50] can be recovered, where (3.6) simplifies to a recursion equivalent to z-iteration [49,51].
- *Optimal control.* For $\beta(x_{<t}) \rightarrow \infty$ at all action nodes and $\beta(x_{<t}) \rightarrow 0$ at all observation nodes, (3.6) simplifies to the Bellman optimality equations [52] of the perfectly rational decision-maker in a stochastic environment.
- *Risk-sensitive control.* For $\beta(x_{<t}) \rightarrow \infty$ at all action nodes and $\beta(x_{<t}) = \beta \neq 0$ at all observation nodes, the framework of risk-sensitive control [53] can be recovered, where a decision-maker allows for model uncertainty in his observational beliefs.
- *Robust control.* For $\beta(x_{<t}) \rightarrow \infty$ at all action nodes and $\beta(x_{<t}) \rightarrow -\infty$ at all observation nodes, robust control and minimax games can be recovered, where in either case the decision-maker makes worst-case assumptions about his environment [54,55].

Robust control and minimax problems have long been known to be related to risk-sensitive control [56,57]. Risk-sensitive KL control was previously studied in [58]. See [33,51] for more details on the relationships between these frameworks.

4. Application to decision-making experiments

In the following, we will discuss the thermodynamic model of bounded rational decision-making in relation to four different kinds of decision-making experiments on humans that have shown systematic deviations from expected utility theory: (i) the Ellsberg experiment, (ii) the Allais experiment, (iii) perceptual decision-making experiments and (iv) risk-sensitive sensorimotor integration experiments. The first two experiments come from behavioural economics, whereas the last two experiments are taken from behavioural neuroscience.

(a) Application to Ellsberg's paradox

In his infamous experiment, Ellsberg could show that humans systematically violate expected utility theory because they are averse to *ambiguity*, that is, model uncertainty as opposed to known *risk* uncertainty [59]. Ambiguity aversion can be modelled by variational preferences [60] that can be straightforwardly incorporated by a bounded rational decision-maker. Consider, for example, an urn that contains 90 balls, 30 of which are red, and the remaining 60 balls are black or yellow in an unknown proportion. While the uncertainty with respect to red constitutes a known risk, we have ambiguity regarding the proportion of black and yellow, as there are many possible distributions. We are now presented with the following gambles:

gamble 1:	receive \$100 if red is drawn
gamble 2:	receive \$100 if black is drawn
gamble 3:	receive \$100 if red or yellow is drawn
gamble 4:	receive \$100 if black or yellow is drawn.

Most people prefer 1 over 2 and 4 over 3, a pattern that is inconsistent with expected utility theory because there is no single probability λ for the proportion of black balls that can explain both choices. This can be seen as follows. In the following, we assume $U(\$0) = 0$ to simplify the argument, although it would, in general, suffice to assume $U(\$100) > U(\$0)$. Preferring 1 over 2 implies $EU_1 > EU_2$, that is, in our case

$$\tfrac{1}{3}U(\$100) > \lambda U(\$100),$$

which implies $\lambda < \frac{1}{3}$. Preferring 4 over 3, however, implies $EU_4 > EU_3$, that is,

$$(\tfrac{2}{3} - \lambda + \lambda)U(\$100) > \tfrac{1}{3}U(\$100) + (\tfrac{2}{3} - \lambda)U(\$100),$$

which implies $\lambda > \frac{1}{3}$, and therefore leads to a contradiction.

The choice pattern is, however, consistent with a decision-maker that is averse to ambiguity. Such an ambiguity averse decision-maker can be constructed by assuming an adversarial environment with rationality $\beta < 0$. The crucial observation is that this leads to a distortion of our beliefs about the hidden variable λ that takes on different values for all possible scenarios—that is all the different proportions of black and yellow. We can then construct the certainty equivalent CE_i of gamble i as follows:

$$\mathrm{CE}_i = \frac{1}{\beta}\log\int_0^{2/3} \mathrm{d}\lambda\, p_0(\lambda)\, \mathrm{e}^{\beta\sum_{x\in\{R,B,Y\}} p_0(x\,|\,\lambda)U(x)},$$

where $p_0(x = R\,|\,\lambda) = \frac{1}{3}$, $p_0(x = B\,|\,\lambda) = \lambda$ and $p_0(x = Y\,|\,\lambda) = \frac{2}{3} - \lambda$. Assuming a uniform prior probability density for $p_0(\lambda) = \frac{3}{2}$ and assuming $U(\$0) = 0$, one can easily check that $\mathrm{CE}_1 > \mathrm{CE}_2$ and $\mathrm{CE}_3 < \mathrm{CE}_4$.

(b) Application to Allais' paradox

Allais' experiment reveals another systematic violation of expected utility theory, which stipulates that the addition or removal of common consequences to two lotteries may not reverse a decision-maker's preferences [61]. Yet, such reversals are frequently observed in gambles like the following [62,63]:

gamble 1:	$0.33 \cdot \$2500$,	$0.66 \cdot \$2400$,	$0.01 \cdot \$0$
gamble 2:	—	$1.00 \cdot \$2400$	—
gamble 3:	$0.33 \cdot \$2500$,	—	$0.67 \cdot \$0$
gamble 4:	—	$0.34 \cdot \$2400$,	$0.66 \cdot \$0$.

Most people prefer 2 over 1 and 3 over 4, even though gambles 1 and 2 can be obtained from gambles 3 and 4 by adding the common consequence of $0.66 \cdot \$2400$. This can be seen as follows.

We will again assume that $U(\$0) = 0$ to simplify the argument. Preferring 3 over 4 implies $EU_3 > EU_4$, that is,

$$0.33U(\$2500) > 0.34U(\$2400).$$

However, preferring 2 over 1 implies $EU_2 > EU_1$, that is,

$$(0.34 + 0.66)U(\$2400) > 0.33U(\$2500) + 0.66U(\$2400),$$

which implies $0.34U(\$2400) > 0.33U(\$2500)$, and leads to a contradiction.

Usually, this reversal is explained by distortions of the probabilities in cumulative prospect theory [35,64]. An alternative explanation of the reversal has been proposed in [65] by introducing a non-vanishing weighting function $a(x)$ in a *generalized quasi-linear mean model* [65–68] such that the certainty equivalent has the form $M_a = \phi^{-1}(\int \mathrm{d}x p(x) a(x) \phi(x) / \int \mathrm{d}x p(x) a(x))$. Such a weighting function could also be constructed for the negative free-energy certainty equivalent by defining a generalized difference $-\Delta\tilde{F}$,

$$\begin{aligned} -\Delta\tilde{F} &= \frac{1}{\beta} \log \left(\frac{\sum_x p_0(x) \mathrm{e}^{\beta V(x)}\, \mathrm{e}^{\beta U(x)}}{\sum_{x'} p_0(x')\, \mathrm{e}^{\beta V(x')}} \right) \\ &= \frac{1}{\beta} \log Z_{V+U} - \frac{1}{\beta} \log Z_V, \end{aligned} \tag{4.1}$$

where $a(x) = \mathrm{e}^{\beta V(x)}$ introduces a bias. In the example, by setting $U(\$2500) = 1$, $U(\$0) = 0$ and $V(\$2500) = V(\$0) = 0$, the reversal can be achieved by setting $U(\$2400) > \frac{1}{\beta} \log \left(\frac{0.33}{0.34} \mathrm{e}^{\beta} + \frac{0.01}{0.34} \right)$ and $V(\$2400) > \frac{1}{\beta} \log \left(\frac{0.66/0.34(1-\mathrm{e}^{\beta})}{(\mathrm{e}^{\beta} + 0.67/0.33 - 1/0.33\, \mathrm{e}^{\beta U(\$2400)})} \right)$ assuming again $\beta < 0$. The reversal cannot be explained without the biasing term $a(x)$, unless one assumes a change of β between the two choices. The biasing term can be regarded both as biasing the probabilities p_0 and as distorting the utility gains U. However, like in prospect theory, there is no explanation why this biasing term should occur from a normative point of view.

(c) Application to perceptual decision-making

In perceptual decision-making experiments, subjects typically have to judge on very short time scales whether one stimulus is bigger or brighter than another, whether a stimulus is moving up or down, left or right, and so forth. One of the most widespread perceptual decision-making experiments are the random-dot motion experiments [69], where a cloud of points move on a screen. While a certain percentage of points moves coherently either up or down, the remainder of points move in random directions. The task is to decide whether the coherent movement is up or down. Depending on the level of coherence, this decision can be easy or difficult, which is reflected in subjects' reaction times—a proxy for their computational resources.

Choice probabilities and reaction times in random-dot motion tasks are usually modelled by diffusion-to-bound models [45]. In the basic model, there is a binary choice between option A and option B associated with values U_A and U_B, respectively, such that the difference in value is given by $\mu = U_A - U_B$. Crucially, the difference in value is not known *a priori* and at each moment of time, there is only a noisy sample of μ available. By observing many samples, however, more evidence can be accumulated, which is modelled by a diffusion process

$$\mathrm{d}x = \mu\, \mathrm{d}t + \sigma\, \mathrm{d}w, \quad x(0) = 0, \quad |x| \le \theta,$$

where μ corresponds to the drift and σ to the diffusion constant. The process starts at zero and ends when it hits one of two bounds θ or $-\theta$, where each bound corresponds to one of the two options. This leads to a speed–accuracy trade-off: the further away the bounds, the more probable it is to reach the bound that corresponds to the option with a higher value, but the longer it will take, as many little diffusion steps will be needed to reach a distant bound.

If we assume a prior probability p_0 for choosing A—and hence $1 - p_0$ for the prior choice probability of B—we can incorporate this information by setting the bound to

$\theta_A = (\sigma^2/2\mu)\log(p_0/(1-p_0)) + (\sigma^2/2)\beta$. This bound can either have a positive or negative sign, which means that choosing A could correspond to crossing the upper ($\theta_A > 0$) or lower ($\theta_A < 0$) bound. The probability of deciding for option A is then

$$P(A) = \frac{1}{1 + \mathrm{e}^{-(2\mu/\sigma^2)\theta}} = \frac{p_0\,\mathrm{e}^{\beta U_A}}{p_0\,\mathrm{e}^{\beta U_A} + (1-p_0)\,\mathrm{e}^{\beta U_B}}$$

and $P(B) = 1 - P(A)$, which corresponds exactly to the formulae suggested to describe bounded rational decision-making. It can also be shown that for large β, the average duration $\langle T \rangle$ of the diffusion is approximately proportional to β,

$$\begin{aligned} \langle T \rangle &= \frac{\theta}{\mu}\tanh\left(\theta\frac{\mu}{\sigma^2}\right) \\ &\propto |\beta| \quad \text{for } |\beta| \gg \left|\frac{1}{\mu}\right|\left[2 + \left|\log\left(\frac{p_0}{1-p_0}\right)\right|\right]. \end{aligned}$$

Consequently, the resource parameter β can also be seen as a proxy of the average computation time.

(d) Application to sensorimotor control

Sensorimotor behaviours can often be described by optimality principles that take into account performance criteria such as energy requirements, endpoint accuracy, smoothness of movement trajectories and other task-related criteria [70,71]. Therefore, choosing how to move an effector can also be considered as a decision-making problem [72]. While many previous studies have investigated the optimization of expected movement costs to describe behaviour, a number of recent studies has found risk-sensitive deviations from expected utility theory in sensorimotor control [73–77].

As already outlined in §3*c*, risk-sensitive decision-makers optimize a *stress function* of the form $J_\beta(u) = -(2/\beta)\ln\mathbb{E}[\mathrm{e}^{-(1/2)\beta C(u)}] \approx \mathbb{E}[C(u)] - \frac{1}{4}\mathbb{VAR}[C(u)]$, where the utility is expressed as a cost $C(u)$ that depends on a control command u [53]. Risk-sensitive decision-makers do not simply maximize the expectation of the utility, but also consider higher-order cumulants. A risk-averse decision-maker ($\beta < 0$) discounts variability, whereas a risk-seeking decision-maker ($\beta > 0$) adds value to the expected utility in the face of variability. In terms of bounded rationality, one could regard the environment also as a bounded rational agent of rationality β that can be either adversarial or collaborative. In fact, a collaborative environment of rationality β is mathematically equivalent to a decision-maker with rationality β and can therefore be regarded as an extension of the decision-maker. If the real environment is partially unknown, risk-sensitivity can be used as a tool to consider model uncertainty. Boundedness in this case is the lack of information about the adequacy of the model. Thus, risk sensitivity can bias the beliefs about the environment optimistically (collaborative environment) or pessimistically (adversarial environment).

Both the effect of sensitivity to variance in the utility and the distortion of beliefs as a consequence of model uncertainty have been recently reported. In [74], subjects had to choose between hitting different targets whose size could be varied. Succeeding or failing to hit a target required a second movement with varying effort. By these two variations, 'motor lotteries' with different degrees of mean and variance in the motor effort could be created. For equal means, subjects were sensitive to this variance. In [77], subjects had to integrate prior information about the position of a target with noisy feedback information. The beliefs about the target position were indicated with a robot handle that required different effort levels for indicating different beliefs. Crucially, in the absence of uncertainty about the target position, these effort levels did not affect behaviour. However, the more uncertainty subjects had about the target position, the more they tended to deviate from the Bayes optimal belief—a signature of model uncertainty.

5. Discussion

In the proposed thermodynamic interpretation of bounded rationality, decision-makers can be thought of as thermodynamic systems abstractly represented by probability distributions. When information processing takes place, these distributions change. Physically, one can imagine the change in distribution as a consequence of imposing a new energy potential. The expected difference in the potential corresponds to a utility gain in economic choice. However, changing states is also costly. In thermodynamic systems, the KL divergence provides a natural measure for the costs that arise due to the changes in the probability distributions. The resulting trade-off between utility gains and resource costs provides a variational principle for bounded rational decision-making in the shape of a negative free-energy difference. The adequacy of this framework can be demonstrated in a number of decision-making experiments from behavioural economics and neuroscience.

The variational principle in the negative free-energy difference generalizes previous studies in which QRE models of bounded rationality assume bounded rational players whose choice probabilities are given by the Boltzmann distribution [20–23], and whose temperature parameters can be interpreted as a strategic choice of preferences [78]. The QRE model can be obtained as a special case of the model presented here where all prior probabilities are assumed to be uniform. However, unlike in the QRE model, these prior probabilities have to be explicitly stated when considering the difference in free energy rather than the free energy itself. These prior probabilities are crucial when defining the certainty equivalent in terms of a finite log-partition sum that ranges from minimum to maximum and includes the expected utility as a limit case. As the certainty equivalent corresponds to physical work, this also allows relating bounded rational decision-making to thermodynamic processes. The distinction of a prior policy and a utility is fundamental to the notion of bounded rationality proposed in this article, which ultimately allows extending the explanatory power of bounded rationality beyond the realm of QRE models.

(a) Bounded rationality

Simon [8] proposed in the 1950s 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 a wrong concept altogether [16]. 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 (3.2) in terms of an anytime search, as illustrated with the example of finding the maximum by drawing from an urn with replacement.

(b) Information theory in control and game theory

As already outlined in §3*c*, a number of studies have recently suggested the use of the relative entropy as a cost function for control [48–50,79]. In [48,49], the transition probabilities of a Markov decision process are controlled directly, and the control costs are given by the KL divergence of the manipulated state transition probabilities with respect to a baseline distribution that describes the *passive dynamics*. In the sequential decision-making setup proposed in §3*c*, this *KL control* corresponds to the case where all random variables are action variables with boundedness parameter β. The stochasticity in this case, however, is not thought to arise from environmental passive dynamics like in the KL-control literature, 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 [47,80], but essentially the same ideas can be applied [51]. This has inspired the formulation of optimal control problems as inference problems [50,81].

Previously, Saridis [82] 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 [83], to use information theory to approximate joint strategies in games with bounded rational players [22] and to the problem of optimal search [84,85], where the utility losses correspond directly to search effort. Recently, Tishby & Polani [86] applied information-theoretic reasoning to understand information flow in the action–observation cycle. The contribution of our study is to extend previous models of bounded rationality by exploiting a thermodynamically motivated variational principle that trades off utility gains and information-processing costs and to apply this principle to human decision-making experiments. In the future, it will be interesting to relate the information-processing resource costs of bounded rational agents to more traditional notions of resource costs in computer science such as space and time requirements of algorithms [87].

(c) Variational preferences

In the economic literature, the KL divergence has appeared in the context of multiplier preference models that can deal with model uncertainty [55]. In particular, it has been proposed that a bound on the KL 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 neighbourhood of p_0. In variational preference models [60], this is generalized to models of the form

$$f \succeq g \iff \min_p \left(\int u(f)\, \mathrm{d}p + c(p) \right) \geq \min_p \left(\int u(g)\, \mathrm{d}p + c(p) \right),$$

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. Owing to this stochasticity of the choice behaviour 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.

(d) Stochastic choice

Stochastic choice rules have extensively been studied in the psychological and econometric literature, in particular logit choice models based on the Boltzmann distribution [24,25]. The literature on Boltzmann distributions for decision-making goes back to Luce [26], extending through McFadden [27], Meginnis [28] and Fudenberg & Kreps [29]. Luce [26] has studied stochastic choice rules of the form $p(x_i) \sim w_i / \sum_j w_j$, which includes the Boltzmann distribution and the softmax rule known in the reinforcement learning literature [30]. McFadden [27] 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 [88–90], 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 the above choice rules is that stochastic choice behaviour 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 β.

(e) Variational Bayes and free-energy principle

It is well known that (approximate) Bayesian inference satisfies a variational principle in the free energy [91]. Given a prior $p_0(h)$ over a latent variable h, a likelihood model $p(y \mid h)$ explaining observation y, and a distribution $q(h)$ to approximate the posterior $p(h \mid y)$, the free-energy difference

$$\Delta F[q] = -\sum_h q(h)\log p(y \mid h) + \sum_h q(h)\log\frac{q(h)}{p_0(h)}$$

is extremized for the Bayesian posterior $q(h) = p(h \mid y)$ under the particular choice of utility $U(h) = -\log p(y \mid h)$, which minimizes informational surprise. Variational Bayes methods often rewrite the free energy as

$$\Delta F[q] = -\log p(y) + \sum_h q(h)\log\frac{q(h)}{p(h \mid y)},$$

which provides a bound on the evidence $-\log p(y)$. Thus, approximate Bayesian inference can be achieved by minimizing free-energy differences, where the boundedness consists of being restricted to model class q. This variational Bayes approach has recently also been proposed as a theoretical framework to understand brain function [36,37], where perception is modelled as variational Bayesian inference over hidden causes of observations. But in this case, the likelihood model $p(y \mid h, a)$ also depends on the chosen action a. According to [36,37], action and perception then consist of choosing a and q, respectively, so as to minimize free energy, thereby minimizing surprise.

6. Conclusion

Thermodynamics provides a framework for bounded rationality that can be both descriptive and prescriptive. It can be descriptive in the sense that it describes behaviour that is clearly sub-optimal from the point of view of a perfectly rational decision-maker with infinite resources. It can be prescriptive in the sense that it prescribes how a bounded rational actor should behave optimally given resource constraints formalized by β. Ultimately, it might even be possible to connect computational processes of bounded rational decision-makers with real physical processes, for example, by relating the generated entropy to energy requirements [40]. Finally, the thermodynamic model of bounded rationality suggests a notion of intelligence that is closely related to the process of evolution. It was already mentioned that bounded rational controllers of the form (3.2) share their structure with Bayes' rule, which in turn shares its structural form with discrete replicator dynamics that model evolutionary processes [92]. In such evolutionary processes, a population of samples 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, favouring 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 time scale, similar random processes might underlie (bounded) intelligent behaviour in individuals on an ontogenetic time scale.

This study was supported by the DFG, Emmy Noether grant no. BR4164/1-1.

References

1. Gigerenzer G, Todd PM, ABC Research Group. 1999 *Simple heuristics that make us smart*. New York, NY: Oxford University Press.
2. Gladwell M. 2005 *Blink: the power of thinking without thinking*. New York, NY: Little, Brown and Company.
3. Niv Y, Daw ND, Joel D, Dayan P. 2007 Tonic dopamine: opportunity costs and the control of response vigor. *Psychopharmacology* **191**, 507–520. (doi:10.1007/s00213-006-0502-4)
4. Daw ND. 2012 Model-based reinforcement learning as cognitive search: neurocomputational theories. In *Cognitive search: evolution algorithms and the brain*. Boston, MA: MIT Press.
5. Von Neumann J, Morgenstern O. 1944 *Theory of games and economic behavior*. Princeton, NJ: Princeton University Press.
6. Savage LJ. 1954 *The foundations of statistics*. New York, NY: Wiley.
7. Fishburn P. 1982 *The foundations of expected utility*. Dordrecht, The Netherlands: D. Reidel Publishing.
8. Simon HA. 1956 Rational choice and the structure of the environment. *Psychol. Rev.* **63**, 129–38. (doi:10.1037/h0042769)
9. Simon H. 1972 Theories of bounded rationality. In *Decision and organization* (eds CB Radner, R Radner), pp. 161–176. Amsterdam, The Netherlands: North-Holland.
10. Simon H. 1984 *Models of bounded rationality*. Cambridge, MA: MIT Press.
11. Lipman B. 1995 Information processing and bounded rationality: a survey. *Canad. J. Econ.* **28**, 42–67. (doi:10.2307/136022)
12. Russell SJ. 1995 Rationality and intelligence. In *Proc. 14th Int. Joint Conf. on Artificial Intelligence* (ed. C Mellish), pp. 950–957. San Francisco, CA: Morgan Kaufmann.
13. Russell SJ, Subramanian D. 1995 Provably bounded-optimal agents. *J. Artif. Intell. Res.* **3**, 575–609.
14. Aumann RJ. 1997 Rationality and bounded rationality. *Games Econ. Behav.* **21**, 2–14. (doi:10.1006/game.1997.0585)
15. Rubinstein A. 1998 *Modeling bounded rationality*. Cambridge, MA: MIT Press.
16. Gigerenzer G, Selten R. 2001 *Bounded rationality: the adaptive toolbox*. Cambridge, MA: MIT Press.
17. Kahneman D. 2003 Maps of bounded rationality: psychology for behavioral economics. *Am. Econ. Rev.* **93**, 1449–1475. (doi:10.1257/000282803322655392)
18. Spiegler R. 2011 *Bounded rationality and industrial organization*. Oxford, UK: Oxford University Press.
19. Camerer C. 2003 *Behavioral game theory: experiments in strategic interaction*. Princeton, NJ: Princeton University Press.
20. McKelvey RD, Palfrey TR. 1995 Quantal response equilibria for normal form games. *Games Econ. Behav.* **10**, 6–38. (doi:10.1006/game.1995.1023)
21. Mckelvey R, Palfrey TR. 1998 Quantal response equilibria for extensive form games. *Exp. Econ.* **1**, 9–41. (doi:10.1007/BF01426213)
22. Wolpert DH. 2004 Information theory—the bridge connecting bounded rational game theory and statistical physics. In *Complex engineering systems* (eds D Braha, Y Bar-Yam). Cambridge, MA: Perseus Books.
23. Wolpert DH, Harré M, Bertschinger N, Olbrich E, Jost J. 2012 Hysteresis effects of changing parameters of noncooperative games. *Phys. Rev. E* **85**, 036102. (doi:10.1103/PhysRevE.85.036102)
24. Luce RD. 2000 *Utility of gains and losses: measurement-theoretical and experimental approaches*. Mahwah, NJ: Erlbaum.
25. Train KE. 2009 *Discrete choice methods with simulation*, 2nd edn. Cambridge, UK: Cambridge University Press.
26. Luce RD. 1959 *Individual choice behavior*. Oxford, UK: Wiley.

27. McFadden D. 1974 Conditional logit analysis of qualitative choice behavior. In *Frontiers in econometrics* (ed. P Zarembka), pp. 105–142. New York, NY: Academic Press.
28. Meginnis JR. 1976 A new class of symmetric utility rules for gambles, subjective marginal probability functions, and a generalized Bayes' rule. In *Proc. American Statistical Association, Business and Economic Statistics Section*, pp. 471–476.
29. Fudenberg D, Kreps D. 1993 Learning mixed equilibria. *Games Econ. Behav.* **5**, 320–367. (doi:10.1006/game.1993.1021)
30. Sutton RS, Barto AG. 1998 *Reinforcement learning: an introduction*. Cambridge, MA: MIT Press.
31. Callen HB. 1985 *Thermodynamics and an introduction to thermostatistics*. New York, NY: Wiley.
32. Ortega PA, Braun DA. 2010 A conversion between utility and information. In *Proc. 3rd Conf. on Artificial General Intelligence*, *Lugano, Switzerland*, pp. 115–120. Atlantis Press.
33. Ortega PA, Braun DA. 2011 *Information, utility and bounded rationality*. Lecture Notes on Artificial Intelligence, vol. 6830, pp. 269–274. Berlin, Germany: Springer.
34. Ortega PA. 2011 *A unified framework for resource-bounded autonomous agents interacting with unknown environments*. Department of Engineering, University of Cambridge, UK.
35. Kahneman D, Tversky A. 1979 Prospect theory: an analysis of decision under risk. *Econometrica* **47**, 263–291. (doi:10.2307/1914185)
36. Friston K. 2009 The free-energy principle: a rough guide to the brain? *Trends Cogn. Sci.* **13**, 293–301. (doi:10.1016/j.tics.2009.04.005)
37. Friston K. 2010 The free-energy principle: a unified brain theory? *Nat. Rev. Neurosci.* **11**, 127–138. (doi:10.1038/nrn2787)
38. Bishop CM. 2006 *Pattern recognition and machine learning.* Berlin, Germany: Springer.
39. Landauer R. 1961 Irreversibility and heat generation in the computing process. *IBM J. Res. Dev.* **5**, 183–191. (doi:10.1147/rd.53.0183)
40. Tribus M, McIrvine EC. 1971 Energy and information. *Scient. Am.* **225**, 179–188. (doi:10.1038/scientificamerican0971-179)
41. Bennett CH. 1973 Logical reversibility of computation. *IBM J. Res. Dev.* **17**, 525–532. (doi:10.1147/rd.176.0525)
42. Bennett CH. 1982 The thermodynamics of computation—a review. *Int. J. Theoret. Phys.* **21**, 905–940. (doi:10.1007/BF02084158)
43. Feynman RP. 1996 *The Feynman lectures on computation.* Reading, MA: Addison-Wesley.
44. Gumbel EJ. 1958 *Statistics of extremes.* New York, NY: Columbia University Press.
45. Bogacz R, Brown E, Moehlis J, Holmes P, Cohen JD. 2006 The physics of optimal decision making: a formal analysis of models of performance in two-alternative forced-choice tasks. *Psychol. Rev.* **113**, 700–765. (doi:10.1037/0033-295X.113.4.700)
46. Ortega PA, Braun DA. 2012 Free energy and the generalized optimality equations for sequential decision making. In *European Workshop for Reinforcement Learning*, *Edinburgh, UK*.
47. Kappen HJ. 2005 A linear theory for control of non-linear stochastic systems. *Phys. Rev. Lett.* **95**, 200201. (doi:10.1103/PhysRevLett.95.200201)
48. Todorov E. 2006 Linearly solvable Markov decision problems. In *Advances in neural information processing systems*, *Vancouver, Canada*, vol. 19, pp. 1369–1376.
49. Todorov E. 2009 Efficient computation of optimal actions. *Proc. Natl Acad. Sci. USA* **106**, 11 478–11 483. (doi:10.1073/pnas.0710743106)
50. Kappen HJ, Gómez V, Opper M. 2012 Optimal control as a graphical model inference problem. *Mach. Learn.* **87**, 159–182. (doi:10.1007/S10994-012-5278-7)
51. Braun DA, Ortega PA. 2011 Path integral control and bounded rationality. In *IEEE Symp. on Adaptive Dynamic Programming and Reinforcement Learning*, *Paris, France*, pp. 202–209.
52. Bellman RE. 1957 *Dynamic programming.* Princeton, NJ: Princeton University Press.
53. Whittle P. 1990 *Risk-sensitive optimal control*. New York, NY: Wiley.
54. Başar T, Bernhard P. 1991 *H-infinity optimal control and related minimax-design problems: a dynamic game approach*. Boston, MA: Birkhäuser.
55. Hansen LP, Sargent TJ. 2008 *Robustness.* Princeton, NJ: Princeton University Press.
56. Jacobson D. 1973 Optimal stochastic linear systems with exponential performance criteria and their relation to deterministic differential games. *IEEE Trans. Automat. Control* **18**, 124–131. (doi:10.1109/TAC.1973.1100265)
57. Glover K, Boyle J. 1988 State-space formulae for all stabilizing controllers that satisfy an H-norm bound and relations to risk sensitivity. *Syst. Control Lett.* **11**, 167–172. (doi:10.1016/0167-6911(88)90055-2)

58. van den Broek JL, Wiegerinck WAJJ, Kappen HJ. 2010 Risk-sensitive path integral control. In *UAI*, Catalina Island, CA, vol. 6, pp. 1–8.
59. Ellsberg D. 1961 Risk, ambiguity and the savage axioms. *Q. J. Econ.* **75**, 643–669. (doi:10.2307/1884324)
60. Maccheroni F, Marinacci M, Rustichini A. 2006 Ambiguity aversion, robustness, and the variational representation of preferences. *Econometrica* **74**, 1447–1498. (doi:10.1111/j.1468-0262.2006.00716.x)
61. Allais M. 1953 Le comportment de l'homme rationnel devant la risque: critique des postulats et axiomes de l'ecole Americaine. *Econometrica* **21**, 503–546. (doi:10.2307/1907921)
62. Kahnemann D, Tversky A. 2002 *Choices, values and frames.* Cambridge, MA: Cambridge University Press.
63. Wu S, Delgado MR, Maloney LT. 2009 Economic decision-making compared with an equivalent motor task. *Proc. Natl Acad. Sci. USA* **106**, 6088–6093. (doi:10.1073/pnas.0900102106)
64. Tversky A, Kahneman D. 1992 Advances in prospect theory: cumulative representation of uncertainty. *J. Risk Uncertain.* **5**, 297–323. (doi:10.1007/BF00122574)
65. Hong CS. 1983 A generalization of the quasilinear mean with application to the measurement of income inequality and decision theory resolving the Allais paradox. *Econometrica* **51**, 1065–1092. (doi:10.2307/1912052)
66. Nagumo M. 1930 Über eine Klasse der Mittelwerte. *Japan J. Math.* **7**, 71–79.
67. Kolmogorov A. 1930 Sur la notion de la moyenne. *Rend. Accad. Lincei* **12**, 388–391.
68. de Finetti B. 1931 Sul concetto di media. *Giornale dell' istituto italiano degli attuari* **2**, 369–396.
69. Britten K, Shadlen MN, Newsome WT, Movshon JA. 1992 The analysis of visual motion: a comparison of neuronal and psychophysical performance. *J. Neurosci.* **12**, 4745–4767.
70. Trommershauser J, Maloney LT, Landy MS. 2003 Statistical decision theory and trade-offs in the control of motor response. *Spat. Vis.* **16**, 255–275. (doi:10.1163/156856803322467527)
71. Todorov E. 2004 Optimality principles in sensorimotor control. *Nat. Neurosci.* **7**, 907–915. (doi:10.1038/nn1309)
72. Wolpert DM, Landy MS. 2012 Motor control is decision-making. *Curr. Opin. Neurobiol.* **22**, 996–1003. (doi:10.1016/j.conb.2012.05.003)
73. Nagengast AJ, Braun DA, Wolpert DM. 2010 Risk-sensitive optimal feedback control accounts for sensorimotor behavior under uncertainty. *PLoS Comput. Biol.* **6**, e1000857. (doi:10.1371/journal.pcbi.1000857)
74. Nagengast AJ, Braun DA, Wolpert DM. 2011 Risk-sensitivity and the mean-variance trade-off: decision making in sensorimotor control. *Proc. R. Soc. B* **278**, 2325–2332. (doi:10.1098/rspb.2010.2518)
75. Nagengast AJ, Braun DA, Wolpert DM. 2011 Risk sensitivity in a motor task with speed–accuracy trade-off. *J. Neurophysiol.* **105**, 2668–2674. (doi:10.1152/jn.00804.2010)
76. Braun DA, Nagengast AJ, Wolpert DM. 2011 Risk-sensitivity in sensorimotor control. *Front. Hum. Neurosci.* **5**, 1. (doi:10.3389/fnhum.2011.00001)
77. Grau-Moya J, Ortega PA, Braun DA. 2012 Risk-sensitivity in bayesian sensorimotor integration. *PLoS Comput. Biol.* **8**, e1002698. (doi:10.1371/journal.pcbi.1002698)
78. Wolpert D, Jamison J, Newth D, Harré M. 2011 Strategic choice of preferences: the persona model. *BE J. Theoret. Econ.* **11**, 18.
79. Peters J, Mülling K, Altun Y. 2010 Relative entropy policy search. In *24th AAAI Conf. Artificial Intelligence, Atlanta, GA*.
80. Theodorou E, Buchli J, Schaal S. 2010 A generalized path integral approach to reinforcement learning. *J. Mach. Learn. Res.* **11**, 3137–3181.
81. Rawlik K, Toussaint M, Vijayakumar S. 2012 On stochastic optimal control and reinforcement learning by approximate inference. In *Proc. Robotics: Science and Systems, Sydney, Australia*.
82. Saridis G. 1988 Entropy formulation for optimal and adaptive control. *IEEE Trans. Automat. Control* **33**, 713–721. (doi:10.1109/9.1287)
83. Still S. 2009 An information-theoretic approach to interactive learning. *Europhys. Lett.* **85**, 28005. (doi:10.1209/0295-5075/85/28005)
84. Stone LD. 1998 *Theory of optimal search.* New York, NY: Academic Press.
85. Jaynes ET. 1985 Entropy and search theory. In *Maximum entropy and Bayesian methods in inverse problems* (eds CR Smith, WT Grandy). Dordrecht, The Netherlands: D. Reidel.
86. Tishby N, Polani D. 2011 Information theory of decisions and actions. In *Perception-reason-action cycle: models, algorithms and systems* (ed. VT Hussain). Berlin, Germany: Springer.

87. Vitanyi PMB. 2005 Time, space, and energy in reversible computing. In *Proc. 2nd ACM Conf. on Computing Frontiers*, *Ischia, Italy*, pp. 435–444.
88. Rieskamp J. 2008 The probabilistic nature of preferential choice. *J. Exp. Psychol.: Learn. Mem. Cogn.* **34**, 1446–1465. (doi:10.1037/a0013646)
89. Gläscher J, Daw N, Dayan P, O'Doherty JP. 2010 States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron* **66**, 585–595. (doi:10.1016/j.neuron.2010.04.016)
90. McDannald MA, Takahashi YK, Lopatina N, Pietras BW, Jones JL, Schoenbaum G. 2012 Model-based learning and the contribution of the orbitofrontal cortex to the model-free world. *Eur. J. Neurosci.* **35**, 991–996. (doi:10.1111/j.1460-9568.2011.07982.x)
91. MacKay DJC. 2003 *Information theory, inference, and learning algorithms.* Cambridge, UK: Cambridge University Press.
92. Shahlizi CR. 2009 Dynamics of Bayesian updating with dependent data and misspecified models. *Electron. J. Statist.* **3**, 1039–1074. (doi:10.1214/09-EJS485)