

**Decision-Making under
Bounded Rationality and Model Uncertainty:
an Information-Theoretic Approach**

Dissertation

zur Erlangung des Grades eines
Doktors der Naturwissenschaften

der Mathematisch-Naturwissenschaftlichen Fakultät
und
der Medizinischen Fakultät
der Eberhard-Karls-Universität Tübingen

vorgelegt

von

Jordi Grau Moya
aus Terrassa, Spanien

October 2016

Tag der mündlichen Prüfung: 14. Juni 2017

Dekan der Math.-Nat. Fakultät: Prof. Dr. W. Rosenstiel
Dekan der Medizinischen Fakultät: Prof. Dr. I. B. Autenrieth

1. Berichterstatter: Prof. Dr. Dr. Daniel A. Braun
2. Berichterstatter: Prof. Dr. Martin A. Giese

Prüfungskommission: PD Dr. Axel Lindner

Prof. Dr. Dr. Daniel A. Braun

Prof. Dr. Martin A. Giese

Prof. Dr. Felix A. Wichmann

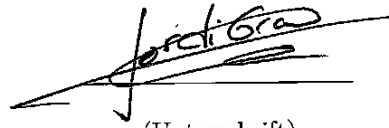
Erklärung

Ich erkläre, dass ich die zur Promotion eingereichte Arbeit mit dem Titel:
“Decision-Making under Bounded Rationality and Model Uncertainty: an Information
Theoretic Approach”

selbständig verfasst, nur die angegebenen Quellen und Hilfsmittel benutzt und wörtlich
oder inhaltlich übernommene Stellen als solche gekennzeichnet habe. Ich versichere an
Eides statt, dass diese Angaben wahr sind und dass ich nichts verschwiegen habe. Mir ist
bekannt, dass die falsche Abgabe einer Versicherung an Eides statt mit Freiheitsstrafe bis
zu drei Jahren oder mit Geldstrafe bestraft wird.

Tübingen, den 26.06.2017

(Datum)



(Unterschrift)

Declaration

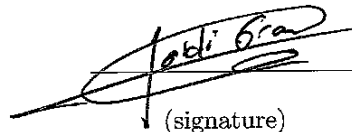
I hereby declare that I have produced the work entitled:

“Decision-Making under Bounded Rationality and Model Uncertainty: an Information
Theoretic Approach”,

submitted for the award of a doctorate, on my own (without external help), have used
only the sources and aids indicated and have marked passages included from other works,
whether verbatim or in content, as such. I swear upon oath that these statements are true
and that I have not concealed anything. I am aware that making a false declaration under
oath is punishable by a term of imprisonment of up to three years or by a fine.

Tübingen, on 26.06.2017

(date)



(signature)

Acknowledgements

Many thanks to my supervisor Daniel that gave me the opportunity to do a PhD in a great town, Tübingen, and in a exciting research environment, the Max Planck Institute. He has been an incredibly supportive supervisor, from whom I have learned every aspect of how to do *excellent* research. Working with him has been a pleasure thanks to his contagious sympathy that has always set a good mood in the lab. Special thanks goes to Pedro for the numerous discussions about virtually everything, for teaching me, and for challenging me with exciting new ideas. He has become a lifelong friend. Big thanks to my colleagues and friends Tim, Felix and Zhen for the countless hours of fun and research.

I want to thank all my friends in Tübingen whom I have shared life experiences through different periods of my PhD adventure: my flatmates and friends, David, Alex, Lea, Rami, Gerard, Andreita, Edu and Lotte, and the rest of the crew, Manuel, Aurelie, Jim, Andrea, Marie, Ainhoa, Nerea, Javier, Laura, Cristina, Alonso, Michel, Carlos, Iñaki and Ainitze. Gracias especiales a Mari por sus sorpresas y alegrías, por su apoyo y confianza, por ser mi compañera de viaje.

Sobretudo doo las gracias a mi familia que siempre me ha apoyado desde la distancia. Gracias papá y especialmente a ti mamá por vuestros valiosos consejos y por hacerme saber que tengo un hogar de por vida. Gracias Sergi por mostrarme tu punto de vista sobre la vida, por tu esfuerzo, por tu apoyo, y por ser mi hermano. Eres el mejor.

Jordi

Abstract

Artificial intelligence research and high computational power have recently led to breakthroughs in solving high-dimensional reinforcement learning and sequential decision-making problems. The foundations of these advances rely on the classical theory of choice under uncertainty, the so-called Subjective Expected Utility (SEU) theory. However, SEU theory assumes two important unrealistic scenarios. First, it disregards computational limitations when making decisions by assuming *perfectly rational* agents i.e. agents with unlimited computational resources. Importantly, humans and artificial agents are *bounded rational*, or equivalently, they suffer from precision and computational limitations. Second, SEU theory assumes that the internal models employed for computation can be fully trusted and that they do not suffer from *model uncertainty*. However, any model of the environment is inherently incorrect and thus it should not be fully trusted. Therefore, humans and artificial agents are indeed subject to model uncertainty.

This thesis consists of an experimental and a theoretical part. On the experimental side, I aimed to explain human sensorimotor behavior with information-theoretic models of bounded rationality and model uncertainty. In particular, we designed three experiments where we expose human subjects to decision-making scenarios involving model uncertainty. We discover that human decision-making behavior can be explained by information-theoretic models that manifest as risk-sensitive and ambiguity-sensitive models. On the theoretical part, we developed a novel planning algorithm for sequential decision-making that accounts for both, information-processing constraints and model uncertainty. Finally, we examined and extended bounded rational models of decision-making under precision and time limitations whose we drew analogies with non-equilibrium thermodynamics. This non-equilibrium thermodynamical point of view allowed to connect decision-making with concepts such as dissipation and time-reversibility, and to discover novel relations connecting equilibrium with non-equilibrium decision-making.

In conclusion, information-theoretic models of decision-making might be the missing cornerstone towards unifying principles of decision-making able to explain complex behavior beyond classic expected-utility models.

Contents

1	Introduction	1
1.1	Bounded Rationality and Limited Computational Resources	3
1.2	Decision-Making under Model Uncertainty	5
1.2.1	Decision-Making under Risk	5
1.2.2	Decision-Making under Ambiguity	7
1.2.3	Neural Correlates of Decision-Making under Risk and Ambiguity	8
1.3	An Information-Theoretic Approach to Decision Theory	10
1.3.1	Bounded Rationality as an Information Constraint	10
1.3.2	Model Uncertainties as Information Constraints	11
1.4	Overview of the Thesis	14
1.4.1	List of Publications and Contributions	18
2	Risk-sensitivity in Bayesian Sensorimotor Integration	21
2.1	Results	23
2.2	Discussion	28
2.3	Materials and Methods	30
2.4	Supplementary Material	33
3	Framing Effects in Decision-Making under Ambiguity	39
3.1	Results	41
3.1.1	An Information-Theoretic Model of Decision-Making under Ambiguity	41
3.1.2	Experiments	44
3.1.3	Experiment 1: Urn Task vs Motor Task	44
3.1.4	Experiment 2: Stimulus versus Motor Framing	50
3.2	Discussion	52
3.3	Materials and Methods	55
3.3.1	Ethics Statement	55
3.3.2	Subjects	55

3.3.3	Materials	55
3.3.4	Information-Theoretic Model Details	56
3.3.5	Experimental Design: Experiment 1 (Urn Task)	57
3.3.6	Experimental Design: Experiment 1 (Motor Task)	58
3.3.7	Experimental Design: Experiment 2	60
3.4	Supporting Information	62
4	The Effect of Model Uncertainty on Cooperation in Sensorimotor Interactions	69
4.1	A Risk-sensitive Model of Interaction	72
4.1.1	Model Uncertainty and Statistical Physics	73
4.2	Simulation Results	75
4.3	Experimental Methods	78
4.3.1	Experimental Design	78
4.3.2	Experimental Apparatus	79
4.3.3	Participants	79
4.4	Results	80
4.5	Discussion	83
5	Planning with Information-Processing Constraints and Model Uncertainty in Markov Decision Processes	89
5.1	Background and Notation	91
5.1.1	Information-Theoretic Constraints for Acting	91
5.1.2	Information-Theoretic Constraints for Model Uncertainty	92
5.2	Model Uncertainty and Bounded Rationality in MDPs	93
5.2.1	Free Energy Iteration Algorithm	95
5.3	Convergence	96
5.4	Experiments: Grid World	99
5.4.1	The Role of the Parameters α and β on Planning	99
5.4.2	Updating the Bayesian Posterior μ with Observations from the Environment	100
5.5	Discussion and Conclusions	103
6	Non-equilibrium Relations for Bounded Rational Decision-making in Changing Environments	105
6.1	Equilibrium Thermodynamics and Decision-Making	107
6.2	Non-equilibrium Thermodynamics and Decision-Making	111
6.2.1	Non-equilibrium Thermodynamics	111

6.2.2	Non-equilibrium Decision-Making	114
6.3	Application to Exemplary Learning and Planning Systems	120
6.3.1	No-Planning: Dissipation and Information-Processing Rate	120
6.3.2	No-Planning: Dissipation and Learning Hysteresis	123
6.3.3	Planning: Jarzynski and Crooks Relations for Episodic Decision-Making	125
6.3.4	Planning: Jarzynski and Crooks Relations for Continuous Decision-Making	128
6.4	Discussion	131
7	Discussion	137
7.1	Summary	137
7.2	Discussion and Outlook	139
7.3	Conclusions	143
	Bibliography	145

Chapter 1

Introduction

“Certainty of some thing is considered either objectively and in itself and means none other than its real existence at present or in the future; or subjectively, depending on us, and consists in the measure of our knowledge of this existence.”

- Jakob Bernoulli, *The Art of Conjecturing*

Probability has taken two different meanings since its conception (Machina and Viscusi, 2013). First, it has an objective meaning in which it deals with relative frequencies of events in repeated trials. Second, it takes a subjective meaning in which it deals with the decision-maker’s degree of belief about the truth of a proposition. Regardless of its subjective or objective meaning, decision-makers facing a choice problem must quantify the probability of an event and, additionally, its expected utility. The concept of utility replaced monetary value as a measure for quantifying the desirability of an event (Bernoulli, 1954). The development of the modern theory of choice needed not only the concept of utility but also two additional formal advancements. One corresponds to the operationalization of subjective probabilities (De Finetti, 1937; Ramsey, 1926) and the other with formalization of maximum expected utility theory (Von Neumann and Morgenstern, 1944). These developments culminated in Subjective Expected Utility (SEU) theory described in *The Foundations of Statistics* (Savage, 1954) in which subjective probabilities and utilities are defined via preferences over observable acts.

SEU theory is formalized as follows (Savage, 1954; Von Neumann and Morgenstern, 1944). The decision-maker chooses an action $x \in \mathcal{X}$ which yields an observation $y \in \mathcal{Y}$ from the environment. Action-observation pairs can be mapped to a single desirability value through the utility function $U : \mathcal{X} \times \mathcal{Y} \mapsto \mathbb{R}$ which expresses the preferences of the decision-maker. Additionally, the subjective probabilistic model of the environment $p(y|x)$ expresses the decision-maker’s belief of observing y given the choice of x . This probabilistic model allows to compute the expected utility

$$U(x) = \sum_y p(y|x)U(x, y) \tag{1.1}$$

of choice x . The expected utility summarizes the choice's worth with a single numerical value, also known as the *certainty equivalent*. Accordingly, a perfectly rational decision-maker chooses the optimal action x^* that maximizes the expected utility

$$x^* = \operatorname{argmax}_x \sum_{y \in \mathcal{Y}} p(y|x)U(x, y). \quad (1.2)$$

Assuming a stochastic strategy $p(x)$ and optimizing in this new space of policies $\mathcal{P} \ni p$, the previous maximization problem is rewritten as

$$p^*(x) = \operatorname{argmax}_{p(x)} \sum_{x \in \mathcal{X}} p(x) \sum_{y \in \mathcal{Y}} p(y|x)U(x, y) \quad (1.3)$$

where $p^*(x)$ is the optimal stochastic policy. The solution to this problem is given by $p^*(x) = \delta(x - x^*)$ (assuming a unique maximum in U). So, in principle, maximizing SEU is straightforward; we only need to compute the expected utility of every action x and put all the probability mass on the action with highest expected utility x^* .

In the context of sensorimotor decisions where noise is inherent to the decision-making process (Faisal et al., 2008), the theoretical framework to evaluate the optimality of motor behavior is optimal feedback control theory which follows the same principles as expected utility theory (Diedrichsen et al., 2010). In fact, SEU has explained successfully human sensorimotor behavior under reaching tasks with monetary payoffs (Trommershäuser et al., 2003b; Trommershäuser et al., 2008) or energy expenditure (C. M. Harris and Daniel M Wolpert, 1998; Todorov and Jordan, 2002). However, in the context of economic decision-making it has been systematically demonstrated that SEU theory fails as a descriptive theory. It is unable to describe human behavior in diverse situations mainly due to framing effects (Daniel Kahneman and Amos Tversky, 1984; Amos Tversky and Daniel Kahneman, 1981), biased estimation of probabilities (Allais, 1953; Daniel Kahneman and Amos Tversky, 1979; Amos Tversky and Daniel Kahneman, 1975) or limited computational resources (Camerer, 2003; Gigerenzer and Selten, 2002; Marewski et al., 2010). As a normative theory, SEU lacks a built-in description of computational resources which are important to consider in all decision-making scenarios subject to time or precision limitations. Additionally when requiring performance guarantees, it also needs accurate probabilistic models of the environment to compute expected utilities, which virtually never happens.

In this thesis we consider the following fundamental problems regarding the normative and descriptive drawbacks of the theory:

- **Bounded Rationality:** Decision-makers have limited computational resources and cannot evaluate expected utilities for all possible actions.
- **Model-Uncertainty:** Decision-makers cannot fully trust their model of the environment $p(y|x)$ because it might be incorrect or uncertain.

In some situations these problems are not big enough to be noticeable. For example, humans dealing with simple decisions without uncertainty (no risk), enough time to make a decision,

and small number of options may not be subject to bounded rationality or model uncertainty. Similarly, artificial decision-making systems that have to make decisions with limited resources might be optimal in simple scenarios where the model of the environment is known and where the computational complexity is low. Importantly, in more realistic situations where decision-problems might be very complex, decision-makers are subject to the previous two impediments. First, the dynamics of the environment are not known and therefore decision-makers must deal with inaccurate models, thus being subject to model uncertainty. Second, their computational resources are scarce and they must balance the value of the decision with its computational cost, thus being subject to bounded rationality. In the following sections, bounded rationality and model uncertainty is described in more detail.

1.1 Bounded Rationality and Limited Computational Resources

SEU theory describes how *perfect rational* decision-makers ought to act regardless of their computational limitations. However, maximizing expected utility is a costly operation that requires decision-makers to search through the whole decision space. More precisely, expected utility optimality can only be achieved by evaluating all elements of the sets \mathcal{X} and \mathcal{Y} —compare Equation (1.3). Importantly, for too large sets this computation becomes intractable. Decision-makers with limited computational resources are unable to perform these intractable computations and are said to be *bounded rational* (Simon, 1955; Simon, 1979). The main point of bounded rationality is that artificial systems and humans should not act according to the best possible action (given by SEU) which requires expensive computations but act in a *satisficing way* given their bounded computational resources. Following such a strategy decision-makers can achieve close to perfect performance but with far fewer computational resources.

Exemplary bounded rational decision-makers are without any doubt human decision-makers. Even though the extraordinary complexity and information processing capabilities of the human brain, we are subject to computational limitations. There are mainly three bottlenecks that limit our capacity to perceive and act (Marois and Ivanoff, 2005). First, we are limited at a perceptual level because of the time it takes to consolidate visual stimulus in short-term memory. These limitations have been studied under the attentional blink paradigm (AB). Second, the visual short-term memory can only hold a limited number of “objects”. Finally, the third bottleneck arises when acting upon a perceived stimulus in the form of a delayed response for a subsequent stimulus—the so-called “psychological refractory period” (PRP) (Welford, 1952). Limited capacity in the visual short-term memory has been found to be localized in the posterior parietal and occipital cortex, whereas AB and PRP are localized in fronto-parietal networks. Both neural regions converge in a common area, the lateral frontal cortex, which acts as an information-processing bottleneck (Buschman et al., 2011) that forces competition between the selection of different actions (Cisek, 2007). In order to perform action selection the brain needs to represent information and compute optimal choices. In fact, such useful representations and computations map directly to decision-theoretic concepts. In

particular, the parietal cortex has been shown to perform state estimation, the basal ganglia computes learning costs and rewards, the premotor cortex implements optimal control policies (Shadmehr and Krakauer, 2008), and the medial prefrontal cortex is in charge of the number of depth levels in strategic reasoning (Coricelli and Nagel, 2009). At the behavioral side, bounded rational behavior has been studied since the conception of decision theory. Early experiments correlated increasing reaction times with increasing number of available options (Hick, 1952; Hyman, 1953), and in more recent experiments it has been found that humans perform approximate Bayesian inference (Moreno-Bote et al., 2011), use limited samples (Vul et al., 2014) and bias choice behavior due to limited computational resources (Lieder et al., 2012). All this evidence not only states the conspicuous idea that we have limited computational resources, but it also shades light where the limitations and bottlenecks of our brains are. Knowing the complexities of how the human brain computes might be relevant to design artificial agents that perform computations in a different way in order to diminish such limitations.

There are widely different approaches to model bounded rationality (Horvitz, 1988; Zilberstein, 2008) such as, for example, heuristics, meta-reasoning or bounded optimality. In particular, heuristic search uses domain knowledge to guide the search process in a decision tree (Gigerenzer and Goldstein, 1996). Generally, these methods do not provide formal guarantees that are necessary to have a formal definition of bounded rationality. In that sense, heuristic approaches might be useful to solve some decision-problems under limited resources but do not provide a formal solution to the bounded rationality problem. A different approach that is more formal is bounded optimality which consists in searching for the maximally successful program that can be computed in a particular machine (S. J. Russell and Subramanian, 1995). Although formally correct, bounded optimality approaches have limited use in practical problems due to their difficulty of finding such programs. Another approach is meta-reasoning in which the agent reasons about its own computational resources (Costantini, 2002). Instead of simply solving the hard problem of maximizing the expected utility, the agent finds a strategy that maximizes utility and also penalizes strategies that incur high computational costs. In principle, this should lead to finding strategies with low computational costs, however, this new meta-level of optimization is also subject to information-processing costs. We can clearly see that this leads to an infinite regression where every time the agent instantiates a new meta-level it generates new costs (Ortega et al., 2015). Thus meta-reasoning, even though it might be useful to solve some problems it is only a partial solution of the bounded rationality problem when disregarding the computational costs at the highest level of the hierarchy.

In this thesis, an existing model of bounded rationality based on information-theoretic quantification of computational costs is adopted (Ortega and Braun, 2013). The main point is that the decision-maker is not allowed to reason about its own computational limitations and instead interrupts computation according to some internal principle. In this way, computational limitations are only noticeable from an observer’s perspective. These kind of bounded rational computations are also named anytime algorithms (Zilberstein, 1996). The theoretical foundations of the adopted approach is described in Section 1.3.1.

1.2 Decision-Making under Model Uncertainty

Decision-makers are confronted with two sources of uncertainty (Knight, 1921). On one hand they face irreducible uncertainty due to the stochastic nature of the environment, for example, when throwing a dice. On the other hand, they face unknown uncertainty when the probabilities of events are not known, for example, the winner in a one-time horse race. In economics, irreducible environmental uncertainty is called risk, whereas the uncertainty regarding probabilities is called ambiguity. Adopting this nomenclature, a *risk-sensitive* decision-maker takes into account not only the expected utility value $\sum_y p(y|x)U(x, y)$, but also higher order moments of U arising from $p(y|x)$. For example, given two options with the same expected utility, a risk-averse decision-maker would choose the one with less variance. In contrast to a risk-sensitive decision-maker, an *ambiguity-sensitive* decision-maker might have a model $p(y|x)$ but it is an untrusted or misspecified model and it shall not be used completely to compute expectations. In general, a decision-maker is risk-sensitive or ambiguity-sensitive depending on the nature of the options of the decision problem. An example of a risky option corresponds to the flip of a fair coin, and an example of an ambiguous option corresponds to the flip of a coin with unknown bias. This highlights the fact ambiguity can vanish when observing multiple tosses of the coin with unknown bias. In other words, the decision-maker becomes more certain about the underlying bias and ultimately uncertainty is reduced to zero. When this happens there is only risk and no ambiguity. In the following two sections I explain further the origins of risk and ambiguity and their implications in decision-making processes.

1.2.1 Decision-Making under Risk

Risk and utility have been born together with the scientific discipline of decision-making under uncertainty around three hundreds years ago (Bernoulli, 1954). In fact, the concept of utility came to life thanks to the St. Petersburg paradox closely related to risk-sensitivity. The paradox goes as follows. We toss a coin repeatedly until the outcome is heads at the n -th toss (where $n \in \mathbb{N}$). The monetary payoff depends on the number of tosses $V(n) = 2^n$. The expected payoff \bar{V} for such a game is infinite

$$\bar{V} = \sum_{n=1}^{\infty} p(n)V(n) = \sum_{n=1}^{\infty} \left(\frac{1}{2}\right)^n 2^n = \infty,$$

where $p(n)$ is the probability of getting heads after n toss. Even though the expectation is infinite, it is clear that nobody would bet an infinite amount of money to play this game. The paradox was solved when Daniel Bernoulli (Bernoulli, 1954)¹ in 1738 recognized that the true value that a person would assign to this game is not simply a monetary expectation but a “moral expectation”, nowadays named expected utility. In particular, (Bernoulli, 1954) suggested a logarithmic function as a utility function to model diminishing utility gains for

¹The citation is a translation from 1954 of the original paper published in 1738.

equal monetary gains as a person gets richer. This induces concavity in the utility function which makes the expectation in the St. Petersburg game finite and thus solves the paradox.

Modeling subjective utilities as marginally diminishing with increasing wealth in order to obtain risk-sensitive behavior, seems to suggest that the curvature of the utility function models risk-sensitivity. In particular, Arrow and Pratt identify risk aversion with the curvature of the utility function in what is called the absolute risk aversion measure (John W Pratt, 1964). This measure is convenient because it is invariant to affine transformations. However, the drawback of defining risk-sensitivity by the curvature of the utility function is that it requires a one-dimensional continuous differentiable function. Therefore, although SEU theory allows to compare options or actions as dissimilar objects, this measure of risk does not. There exist other measures of risk which do not suffer from this drawback. The primary alternative is the mean-variance decision-model (Markowitz, 1952) where the decision-maker does not only care about the expected utility $\mathbb{E}[U]$ of option x , but also about the variance $\text{VAR}[U]$ in the following way $\mathbb{E}[U] - \alpha \text{VAR}[U]$. Thus in mean-variance models the risk associated to a random outcome is measured by the variance which can be computed in both, real valued and categorical random variables. The parameter α captures the risk-attitude of decision-makers. For $\alpha = 0$, $\alpha > 0$ and $\alpha < 0$ we obtain risk-neutral, risk-averse, and risk-seeking decision-makers. Although mean-variance models are interesting as a way to model risk-sensitivity they do not take into account all the other higher order moments of the random utility gains. For that, other models have been proposed that use exponential functions that generate all moments (Whittle, 1981).

In an experimental context, one of the most intriguing violations of expected utility which cannot be explained by the aforementioned risk models is the Allais' paradox (Allais, 1953). The paradox essentially takes two lotteries, A and B , and adds a common consequence to both (e.g. adding a 66% of obtaining a certain amount of money) in order to obtain the new lotteries C and D , respectively. Under SEU, the decision-maker should prefer C if it first preferred A , or D if it first preferred B . However, in many experiments, humans choose a reversed pattern, that is they switch from preferring A in the first choice, to preferring D in the second, or switch from B to C . This suggests a violation of the independence axiom, central to SEU theory (Ray, 1973). Allais' paradox spurred the development of one of the most successful theories of choice, Prospect Theory (Daniel Kahneman and Amos Tversky, 1979).

In contrast to economic decision-making, sensorimotor decisions have been shown to be consistent with risk-neutral SEU theory (Braun et al., 2011a; Diedrichsen et al., 2010; C. M. Harris and Daniel M Wolpert, 1998; Todorov and Jordan, 2002; Trommershäuser et al., 2003b; Trommershäuser et al., 2008). However, few studies have reported instances of risk-sensitivity in sensorimotor control, for example, when testing the mean-variance risk model (Nagengast et al., 2010; Nagengast et al., 2011b). Recently, Wu et. al (Wu et al., 2009) have compared economic decision-making with sensorimotor decision-making in an Allais' type scenario, and they explained the observed behavior with decision-making models of risk (Prospect Theory). Intriguingly, they have reported a switch of risk-attitude when framing from pen-and-paper to

sensorimotor decision-making contexts. This is an interesting result because framing effects—so important in shaping human behavior (Daniel Kahneman and Amos Tversky, 1984; Amos Tversky and Daniel Kahneman, 1981)—might not only be detectable in pen-and-paper psychological tasks, but also at a lower level in simple perception-action sensorimotor tasks.

1.2.2 Decision-Making under Ambiguity

Although Knight, 1921 did already distinguish between the terms “risk” and “uncertainty” (also known as risk and ambiguity), he did so referring to the absence or existence of *objective* probabilities not *subjective* probabilities (Machina and Viscusi, 2013). It was not until the work of Ellsberg, 1961 and his paradox that human subjects were found to be inconsistent with Savage’s axioms of SEU theory (Slovic and Amos Tversky, 1974). The paradox specifically tackles the situation where one does not know the probability of a certain event, thus in mathematical terms, the decision-maker might have a model $p(y|x)$ but it is untrusted. In particular, Ellsberg (Ellsberg, 1961) presents to human subjects a decision problem where they have to choose one of two boxes. One of them is filled with a known 50% proportion of black and white balls, and the other is filled with an unknown proportion of black and white balls. From the chosen box, a ball is going to be randomly drawn with distinct reward associated to its color e.g. a reward of \$5 if it is white and \$0 if it is black. Therefore, the probabilities $p(y|x)$ of observing $y = \text{“white”}$ or $y = \text{“black”}$ are known for the first box and unknown for the second. When asking human participants to choose a box, they prefer, in 70% of the cases (Ellsberg, 1961), the box with known probabilities; paradoxically, they maintained this preference even when swapping the prize to the other color! Importantly, there is no single subjective belief about the proportion of the unknown box that can explain this choice behavior and thus SEU theory seems to be violated. These results provide evidence of the short-comings of SEU theory to explain human decision-making under ambiguity.

Two-player games. Ambiguity also plays a role in two-player game-theoretic scenarios where one player chooses x and the other chooses y . In these scenarios, the strategy of the second player $p(y|x)$ (in sequential games) is hidden—from the first player’s point of view. Due to the uncertainty about the hidden strategy, decision-makers might want to choose their actions robustly, expecting that the opponent is going to choose the worst possible action y . More formally, a robust decision-maker facing an adversarial player chooses according to a “maxmin” strategy which can be expressed as

$$x^* = \operatorname{argmax}_x \min_{p(y|x)} \sum_y p(y|x) U(x, y).$$

Additionally, in the case of a friendly opponent that is expected to choose the best possible observation, a “maxmax” strategy can be employed, which is formally written as

$$x^* = \operatorname{argmax}_x \max_{p(y|x)} \sum_y p(y|x) U(x, y).$$

However, when facing a bounded rational adversarial opponent the choice strategy is doubtfully optimal from the perspective of the first player. Similarly, a teammate with few computational resources might not be able to choose the best strategy. How can we model these intermediate levels of pessimism and optimism? The following section 1.3.2.2 explains how to model such kind of intermediate model trust and intermediate optimism with a free energy formulation.

1.2.3 Neural Correlates of Decision-Making under Risk and Ambiguity

In the following I review the findings in the literature connecting the computational models of decision-making with neural correlates in the brain. In particular, I focus on the neural signatures relevant to the topics of this thesis, such as where and how the brain encodes value, uncertainty, risk and ambiguity.

Representation of utility and uncertainty. Decision-theory relies on two fundamental pillars: probability theory and utility theory. Therefore, in order to make decisions, the brain must represent both, probabilities and utilities (or value). The most important brain region that is known to encode value is the *orbitofrontal cortex* (OFC) (Bartra et al., 2013) which not only encodes value in a variety of contexts but also consistently with behavioral theories of decision-making under risk (Rangel and Hare, 2010). In particular, specific rewards have been shown to be represented in a continuous scale in the OFC which specializes in stimulus value coding, and also in the anterior cingulate cortex (ACC) which specializes in action cost and value coding (Grabenhorst and Rolls, 2011; Rangel and Hare, 2010; Shenhav et al., 2013). Not only that, blood oxygenation level dependent (BOLD) signals in the ventromedial prefrontal cortex (VMPFC)—a bigger region that includes the OFC—scale with the subjective value of the available options at time of choice and respond when rewards are received (Bartra et al., 2013), thus suggesting a common system for reward prediction and acquisition.

A key aspect of the decision-making process is the evaluation of probabilities. For that, the brain must have a way to represent uncertainties or probabilities associated with every possible outcome when selecting an action or when doing inference. Previously thought a unique human capability, representing uncertainty is a fundamental component of sensorimotor processing also present in other animals including smaller rodents (Kepecs et al., 2008). An abundant amount of research has been done in unveiling how uncertainty is represented in the brain—see (Pouget et al., 2013) for an extensive review. In particular, in the last two decades it has been proposed that neural activity encodes functions over latent variables. These functions are either probability distributions or log probabilities. For example, the response of a neuron calibrated to detect a certain stimulus is proportional to either the probability or the log of the probability of that stimulus being present (Anastasio et al., 2000; Koehler et al., 1999). Either representation is appropriate for different computations such as adding or multiplying probabilities. Adding probabilities is easy when using a probability code, whereas multiplying probabilities is easy when using a log probability code. Using

either code, functions can be represented by a sum of basis functions. For example, the log probability of a latent variable h given a vector of neural responses \mathbf{r} can be represented as $\log p(h|\mathbf{r}) = \sum_i r_i f_i(h) + k$ where f_i are the basis functions and k is a normalization constant. This way of representing functions is useful because multiplications of likelihoods can be performed by just adding neural responses. Likelihood multiplications are the basis for performing inference, for example, when combining different sources of information coming from different sensory input in order to estimate a latent variable. These probabilistic approaches to decode neural computations assume that ultimately the brain mimics an optimal inference machine, also known as the Bayesian brain hypothesis (Knill and Pouget, 2004). In the end, the objective of these theoretical frameworks is to integrate them with theories of decision-making at a neural level with the goal of explaining, not only how we represent uncertainty but how do we use it for action (Bach and Dolan, 2012).

Representation and processing of risk and ambiguity. Importantly, the value associated with actions also depends on their corresponding riskiness and ambiguity. There have been breakthroughs in showing how and where the brain encodes and processes risk and ambiguity (Hsu et al., 2005; Huettel et al., 2006). A bigger volume of investigations showed neural correlates for decision-making under risk compared to decision-making under ambiguity (Platt and Huettel, 2008). However, here we give an succinct overview of the research regarding both kinds of decision-making scenarios and their corresponding neural substrates.

In essence, option valuation under risk is modulated by the magnitude of both, the expected value and the variance. Numerous neuroimaging studies examined neural activity under conditions of risk. Options with higher expected values induced higher activation in distinct regions of the striatum and options with higher uncertainty elicited increasing activations in the lateral orbitofrontal cortex (l-OFC) (Hsu et al., 2005; O’Neill and Schultz, 2010; Tobler et al., 2007). Interestingly, these studies also found that higher activations in the l-OFC correlated with risk-averse behavior, and that activations in medial areas correlated with risk-seeking behavior. Additionally, they found that value coding in the prefrontal cortex (PFC) correlates differentially with uncertainty. This suggest separate prefrontal regions being involved in processing of risk-attitudes. Bold oxygen level-dependent (BOLD) correlates of value and risk were found in regions of the ventral striatum and anterior cingulate, respectively, and the inferior frontal gyros activity was associated to low risk and safe options. Interestingly, these correlations allowed to accurately decode the behavioral response (Christopoulos et al., 2009). Risk prediction is another important aspect when dealing with sequential decision-making tasks and it has been shown that the insula governs the mechanisms for risk prediction error (Preuschoff et al., 2008). In summary, all of these studies reinforce the view that valuation of decisions under risk is processed in the striatum and the OFC, and that risk-sensitivity is encoded in distinct regions in the PFC.

Ambiguity refers to the lack of knowledge about the probabilities of random outcomes. Thus if one learns the probability distribution of the random process, then ambiguity vanishes and the decision-problem only comprises risk-uncertainty. For this reason, the majority of the

neuroimaging studies about ambiguity include also instances of risk. In particular, it has been found that decision-making under ambiguity activates more the amygdala and the OFC than under risk and that patients with lesions in the OFC show no risk or ambiguity aversion (Hsu et al., 2005). Another area relevant to the processing of risk and ambiguity is the PFC. In this area, activity representing subjective value has been shown to be higher for ambiguity compared to risk (Huettel et al., 2006) whereas the posterior parietal cortex (PPC) showed higher activation for risk compared to ambiguity. These findings suggest distinct mechanisms in the processing of risk and ambiguity. Importantly, the subjective valuation under risk and ambiguity is governed in both cases by a common system, the striatum and the medial prefrontal cortex (mPFC) (I. Levy et al., 2010). In general, activation patterns in decision-making under risk and all degrees of ambiguity have been shown to be correlated in a “fronto-parietal-striatal” network. Interestingly, higher activation was found in the PFC for partial ambiguity conditions, suggesting that this neural region does not simply track the degree of ambiguity but possibly the difficulty of the decision-process (Bach et al., 2009; Lopez Paniagua and Seger, 2013). In general, it seems that the same area, the OFC, is involved in valuation of both risky and ambiguous options and that it produces higher activation for higher uncertainty.

1.3 An Information-Theoretic Approach to Decision Theory

The methods presented in the previous sections for the modeling of bounded rationality and model uncertainty, seem diverse and disconnected. However, when considering information-theoretic approaches they can be unified by common theoretical principles. In the following, I show how adding information-theoretic constraints to the SEU maximization problem leads to a seamless integration of bounded rationality and model uncertainty.

1.3.1 Bounded Rationality as an Information Constraint

In the recent past there has been an increasing interest in modeling decision-making with limited information processing capabilities from an information theoretic point of view (Braun and Ortega, 2014; Braun et al., 2011b; Friston, 2010; Kappen et al., 2012; Ortega and Braun, 2011; Ortega and Braun, 2013; Ortega et al., 2015; Rubin et al., 2012; Still, 2009; Still et al., 2012; Tishby and Polani, 2011; Todorov, 2009; Vijayakumar et al., 2012). In the following, I present an information-theoretic bounded rationality framework based on (Ortega et al., 2015).

A bounded rational decision-maker that cannot select the optimal action x^* due to limited information-processing capabilities can be characterized by the following constrained maximization problem

$$\max_{\pi} \sum_x \pi(x) U(x) \quad \text{s.t.} \quad D_{\text{KL}}(\pi || \rho) \leq K \quad (1.4)$$

where $D_{\text{KL}}(\pi || \rho) = \sum_x \pi(x) \log \frac{\pi(x)}{\rho(x)}$ is the relative entropy that measures the information

‘distance’² between the prior reference distribution ρ over actions and the posterior policy π . This constrained maximization problem trades off maximization of utility with computational cost measured by the relative entropy. Conversely, the same problem (1.4) can be rewritten as an unconstrained optimization problem with the method of Lagrange multipliers, which gives

$$\max_{\pi} \sum_x \pi(x)U(x) - \frac{1}{\alpha} D_{\text{KL}}(\pi||\rho). \quad (1.5)$$

The solution can be found by taking the functional derivative, equating to zero, and solving for π . This solution is written as

$$\pi^*(x) = \frac{\rho(x)e^{\alpha U(x)}}{\sum_x \rho(x)e^{\alpha U(x)}} \quad (1.6)$$

where the parameter α controls the computational resources of the decision-maker. For $\alpha \rightarrow 0$ it models a decision-maker with no resources that chooses according to its prior strategy ρ and for $\alpha \rightarrow \infty$ we recover a SEU decision-maker that chooses the best action $x^* = \operatorname{argmax}_x U(x)$. For finite α , we have a bounded rational decision-maker that trades off utility and computational resources. Note that when the prior is a uniform distribution the choice strategy takes the form of a soft-max rule.

Although the information-theoretic description is appealing and theoretically convenient, the bridge from information resources to algorithmic resources is not clear and has to be studied case by case. However, it can be shown that a decision-maker using a rejection-sampling scheme (MacKay, 1998) needs less samples from ρ to obtain a sample from the posterior policy π when the boundedness parameter α is reduced (Ortega and Braun, 2014). This highlights an important connection between information-theoretic computational resources measured by the relative entropy and algorithmic computational resources measured by the number of samples. Even though the employed information-theoretic terms that quantify computations, at first sight, seem disconnected from real computations, they are not.

1.3.2 Model Uncertainties as Information Constraints

In the following section I adopt an information-theoretic approach to model both risk-sensitivity and ambiguity sensitivity in a similar way that we modeled bounded rationality.

1.3.2.1 Risk-Sensitivity as an Information Constraint

In SEU theory, the certainty equivalent $F(x)$ assigned to option x is the expected utility value

$$F(x) = \sum_y p(y|x)U(x, y), \quad (1.7)$$

while in the mean-variance models the certainty equivalent is computed as

$$F(x) = \sum_y p(y|x)U(x, y) + \beta \mathbb{V}\mathbb{A}\mathbb{R}[U] \quad (1.8)$$

²In fact is not a distance because is not symmetric, but for our purposes it effectively acts as a distance.

where $\text{VAR}[U] = \sum_y (U(x, y) - \bar{U})^2 p(y|x)$ and $\bar{U} = \sum_y p(y|x)U(x, y)$ is the mean. A risk-averse ($\beta < 0$) or risk-seeking ($\beta > 0$) decision-maker facing, for example, two options x_1 , and x_2 , chooses the one with the highest certainty equivalent.

The information-theoretic used here is a similar model to the mean-variance models of risk-sensitivity, it falls within the class of entropic risk measures (Föllmer and Schied, 2011), and it has been used before in the control literature (Whittle, 1981). Specifically, we model risk-sensitivity as an optimization problem that trades off utility and information. In contrast to information-theoretic bounded rationality, in order to express risk-sensitivity we limit the informational distance between a biased model $q(y|x)$ and a true model of the environment $p(y|x)$. Formally, to obtain the certainty equivalent for a risk-seeking decision-maker ($\beta > 0$), the following optimization problem needs to be solved

$$F(x) = \max_{q(y|x)} \sum_y q(y|x)U(x, y) - \frac{1}{\beta} D_{\text{KL}}(q||p). \quad (1.9)$$

In contrast, for a risk-averse decision-maker ($\beta < 0$), the following problem needs to be solved

$$F(x) = \max_{q(y|x)} \sum_y q(y|x)U(x, y) - \frac{1}{\beta} D_{\text{KL}}(q||p). \quad (1.10)$$

Importantly, the solution for both problems is the same and it is written as

$$q^*(y|x) = \frac{p(y|x)e^{\alpha U(x,y)}}{\sum_y p(y|x)e^{\beta U(x,y)}}. \quad (1.11)$$

Similarly, the certainty equivalent is the same for both problems and it is rewritten as

$$F(x) = \frac{1}{\beta} \log \sum_y p(y|x)e^{\beta U(x,y)}. \quad (1.12)$$

Importantly, under this formulation β can be conveniently set to a positive or negative number keeping the same mathematical form for the expression of the certainty equivalent. Note that with this approach, risk-sensitivity is expressed not by the curvature of the utility function, but directly by the free energy of choice x for a given temperature β . The connection of the free energy model for risk-sensitivity and the mean-variance model can clearly be seen when doing the second order Taylor approximation of the free energy. Specifically,

$$F(x) = \frac{1}{\beta} \log \sum_y p(y|x)e^{\beta U(x,y)} \approx \sum_y p(y|x)U(x, y) + \frac{\beta}{2} \text{VAR}[U] \quad (1.13)$$

when $\beta \text{VAR}[U]$ is small (Whittle, 1981). Therefore, in this view, the utility function only focuses on expressing the desirability of the distinct options alone and does not adopt the role of characterizing risk-sensitivity. Instead, the free energy is responsible for the risk-sensitive valuation. We can recover different risk-sensitive decision-makers depending on the value of β . In particular, for $\beta \rightarrow 0$ we recover a risk-neutral agent where the certainty equivalent takes the form of a plain expected utility $F(x) = \sum_y q(y|x)U(x, y)$. For $\beta > 0$, we recover

a risk-seeking agent that assigns a higher value to the certainty equivalent compared to the risk-neutral agent. In the limit of $\beta \rightarrow \infty$, we recover an infinitely risk-seeking agent with certainty equivalent $F(x) = \max_y U(x, y)$. On the contrary, for $\beta < 0$, we recover a risk-averse agent that assigns a lower value to the certainty equivalent compared to a risk-neutral agent. In the limit of $\beta \rightarrow -\infty$, the certainty equivalent is equal to the lowest utility value $F(x) = \min_y U(x, y)$.

1.3.2.2 Ambiguity-Sensitivity as an Information Constraint

As outlined in Section 1.2.2, ambiguity refers to the lack of knowledge about the model of the world $p(y|x)$. The standard way to express this uncertainty is to assign, for example, a Bayesian model μ over the possible environment models p in the form of $\mu(p)$. Using a parametric model of the environment $p_\theta(y|x)$ (parametrized by θ), we can express the uncertainty about the parameters directly as $\mu(\theta)$. In this way a decision-maker that receives data D updates his probabilistic model by means of Bayes' rule $\mu(\theta|D) = \frac{1}{Z} \mu(\theta) p(D|\theta)$, where $p(D|\theta)$ is the likelihood model. We express the dependency between option x and the Bayesian model $\mu(\theta|D, x)$ by conditioning on x . Under this view, being ambiguous about the probabilities of the true model of the environment $p(y|x)$ translates directly in the amount of uncertainty about the parameters θ by means of the Bayesian model μ . A robust decision-maker that requires utility guaranties against these uncertainties would consider a restricted set of models within the neighborhood of μ . Here we employ a model of ambiguity that falls within the class of methods that use a restricted set of permissible models (Iyengar, 2005; Nilim and El Ghaoui, 2005). In economics, similar class of models have been proposed that fall within the class of variational preference models (Maccheroni et al., 2006) in particular the multiplier preference model (Hansen and Sargent, 2008). Specifically, to model ambiguity-seeking behavior under model uncertainty we allow for optimistic deviations by framing the problem as a free energy optimization problem

$$\psi^*(\theta|x) = \operatorname{argmax}_{\psi} \sum_{\theta} \psi(\theta|x) \sum_y p_{\theta}(y|x) U(x, y) - \frac{1}{\gamma} D_{\text{KL}}(\psi(\theta|x) || \mu(\theta|D, x)) \quad (1.14)$$

with ψ as an argument, whereas to model for ambiguity-averse behavior we allow for pessimistic deviations,

$$\psi^*(\theta|x) = \operatorname{argmin}_{\psi} \sum_{\theta} \psi(\theta|x) \sum_y p_{\theta}(y|x) U(x, y) - \frac{1}{\gamma} D_{\text{KL}}(\psi(\theta|x) || \mu(\theta|D, x)). \quad (1.15)$$

Importantly, the optimal biased model in both optimization problems is

$$\psi^*(\theta|x) = \frac{1}{Z} \mu(\theta|D, x) e^{\gamma \sum_y p_{\theta}(y|x) U(x, y)}. \quad (1.16)$$

and the certainty equivalent of option x is

$$F(x) = \frac{1}{\gamma} \log \int \mu(\theta|D, x) e^{\gamma \sum_y p_{\theta}(y|x) U(x, y)} d\theta \quad (1.17)$$

assuming that the space of θ is continuous.

Similarly to the case of risk-sensitivity, we can recover a number of different decision-makers when changing the parameter γ . For example, when $\gamma \rightarrow \infty$, we recover an infinitely optimistic agent that disregards the current knowledge $\mu(\theta|D)$ about the parameters and assigns a certainty equivalent to option x as $F(x) = \max_{\theta} \sum_y p_{\theta}(y|x)U(x, y)$. In contrast, for $\gamma \rightarrow -\infty$ we recover an infinitely robust agent that assigns a certainty equivalent with $F(x) = \min_{\theta} \sum_y p_{\theta}(y|x)U(x, y)$. For $\gamma \rightarrow 0$, the decision-maker expresses full trust in its probabilistic model $\mu(\theta|D)$ and therefore assigns a certainty equivalent that is the expectation $F(x) = \int \mu(\theta|D) \sum_y p_{\theta}(y|x)U(x, y)d\theta$. Importantly, when learning about the environment, the Bayesian model $\mu(\theta|D, x)$ becomes more peaked around the true parameters θ^* . This makes the biased model ψ^* closer to μ because such deviations become more costly in terms of relative entropy. This means that when the decision-maker learns about the environment, ambiguity vanishes as expected.

1.4 Overview of the Thesis

In this thesis I am interested in experimentally testing the validity of the previously presented models of information-theoretic decision-making in explaining human sensorimotor behavior. Additionally, I also aimed to advance the theoretical part of this framework by designing novel planning and decision-making models that take into account bounded rationality and model uncertainty. In particular, I will try to answer the following experimental and theoretical questions.

Experiments in Human Decision-Making This experimental part deals with the general question: Are humans making choices according to the proposed information-theoretic models of bounded rationality, risk and ambiguity in laboratory experiments? (Chapter 2, 3 and 4). Additionally,

- Is the human sensorimotor system subject to risk-sensitivity in an estimation task? (Chapter 2)
- Do the same humans have different ambiguity attitudes in different situations? What are the important factors that determine these ambiguity attitudes?(Chapter 3)
- In two-player games, how are cooperative solutions affected by model uncertainty coming from different players? (Chapter 4)

Theoretical Advancements of Information-Theoretic Approaches to Decision-Making

In this part we focused on the following questions.

- How can we extend the theory in a sequential decision-making scenario to take into account bounded rationality and model uncertainty simultaneously when planning into the future? (Chapter 5)

- Given that the free energy is a concept from statistical physics, how does bounded rational decision-making relate to non-equilibrium statistical physics? Can we make novel predictions when importing concepts from physics to decision-making? (Chapter 6)

In the following we summarize the remaining chapters of the thesis.

Chapter 2: Risk-Sensitivity in Bayesian Sensorimotor Integration. Previously, it has been shown that the nervous system employs probabilistic models during sensorimotor learning (K. P. Körding and Daniel M Wolpert, 2004; Daniel M Wolpert et al., 1995). In particular, in (K. P. Körding and Daniel M Wolpert, 2004) the authors show in a sensorimotor task that human subjects internally represent sources of information about latent variables stemming from prior knowledge and visual feedback, and that they combine this information consistent to risk-neutral Bayesian inference. However, the way they designed their experimental setup did not allow subjects to exhibit risk-sensitive deviations. In our first experiment we develop a modified estimation task with added sensorimotor costs in order to test for risk-sensitive behavior. We show that the nervous system is not only consistent with Bayesian inference but it is risk-sensitive to the underlying utility function. In particular, the behavioral data recorded in this new experiment is consistent with our predictions from the information-theoretic model of risk-sensitivity (see Section 1.3.2.1) where the decision-making process depends on an interplay between both uncertainty and utility. This chapter is based entirely on the publication:

- Grau-Moya, J., Ortega, P. A., & Braun, D. A. (2012). **Risk-sensitivity in bayesian sensorimotor integration.** PLoS Computatinal Biology, 8(9), e1002698.

Chapter 3: Framing Effects in Decision-Making under Ambiguity. While many studies have shown how human decision-making under ambiguity deviates from expected utility theory in pen-and-paper tasks (mainly from researchers following the work of Ellsberg (Ellsberg, 1961)) less research has been devoted towards the understanding of human sensorimotor behavior under ambiguity. Similarly, while there is a study that shows how humans have different risk-attitudes between tasks under different framing such as pen-and-paper and sensorimotor framing (Wu et al., 2009), there has been no report in the literature about these differences under ambiguity. Here we are interested in understanding human behavior under ambiguity and under different framing conditions including sensorimotor framing and visual framing. To do so we designed an experiment with two tasks. One is a modified version of the original Ellsberg’s task involving choices between urns filled with balls. The other is a translated version of Ellsberg’s task in a sensorimotor domain where uncertainty arises due to the intrinsic stochasticity of the sensorimotor system. Not only do we find that humans exhibit different ambiguity-attitudes depending on the visual framing and not due to a sensorimotor framing, but we also suggest that the reported results in (Wu et al., 2009) might stem from visual framing and not from a sensorimotor framing. This chapter is based entirely on the publication:

- Grau-Moya, J., Ortega, P. A., & Braun, D. A. (2016). **Decision-Making under Ambiguity Is Modulated by Visual Framing, but Not by Motor vs. Non-Motor Context. Experiments and an Information-Theoretic Ambiguity Model.** *PLoS one*, 11(4), e0153179.

Chapter 4: The Effect of Model Uncertainty on Cooperation. In the previous two chapters we have tested human behavior under model uncertainty and found it to be consistent with our information-theoretic models. However, to the best of our knowledge, studies regarding model uncertainty in two-player games have not been reported in the literature. In such games it is common that players have model uncertainty given that they are unaware of the probabilities of the opponent’s strategy. When simulating choice behavior of two players using our information-theoretic models we obtain interesting predictions—such as a coupling between choice behavior and the opponent’s ambiguity-sensitivity parameter—that can be tested in an experiment with human subjects. In particular, we designed an experiment where humans played a simple cooperation game, the “stag hunt game”, where the opponent was a computer player. This way we were able to manipulate the computer’s ambiguity-parameter γ (see Section 1.3.2.2) and test human choice behavior under different values of the opponent’s ambiguity-parameter γ . We show that human choice behavior strongly depends on γ and that is consistent with the information-theoretic model under ambiguity that differs from standard models in the literature such as fictitious play. In particular, we show that the opponent’s ambiguity-sensitivity plays a crucial role in driving the interaction into cooperative or non-cooperative behavior. These findings could be important in the designing of robots that need to interact with humans as has already been shown in (Medina et al., 2015). This chapter is based entirely on the publication:

- Grau-Moya, J., Hez, E., Pezzulo, G., & Braun, D. A. (2013). **The effect of model uncertainty on cooperation in sensorimotor interactions.** *Journal of The Royal Society Interface*, 10(87), 20130554.

Chapter 5: Planning with Model Uncertainty and Bounded Rationality. All our experimental findings so far were based on single-step decision-making processes and did not have a sequential nature. In fact, the theoretical developments of sequential decision-making under both, bounded rationality and model uncertainty, have not been reported in the literature so far. Here we are interested in designing a model that considers model uncertainty and bounded rationality in a sequential fashion. For that we ground our theoretical formulation in Markov Decision Processes (MDP). An MDP is a tuple $(\mathcal{S}, \mathcal{A}, T, R, \gamma)$ where \mathcal{S} is a finite set of states, \mathcal{A} is a finite set of actions, $T : \mathcal{S} \times \mathcal{A} \rightarrow \prod(\mathcal{S})$ is a state transition function $T(s', a, s)$ of ending in state s' when taking action a from state s , $R : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow \mathbb{R}$ is a reward function and $\gamma \in (0, 1)$ is a discount factor. In this setting, the goal of the agent is to

maximize the discounted expected return

$$V_\pi(s_0) = \lim_{N \rightarrow \infty} \mathbb{E} \sum_{t=0}^{N-1} \gamma^t R(s_t, a_t, s_{t+1}) \quad (1.18)$$

where the expectation \mathbb{E} is over trajectories of action-state pairs $s_0, a_0, s_1, a_1 \dots$ with probability $\pi(a_0|s_0)T(s_1|s_0, a_0)\pi(a_1|s_1) \dots$ given a starting initial state s_0 .

There are two ways to solve an MDP, by planning—known as *dynamic programming*—or by learning, through interaction with the environment—for example by *temporal-difference* methods (Sutton and Barto, 1998). In this chapter we will construct a new optimization problem which takes into account the information-theoretic cost in the policy (bounded rationality) and in the model of the environment (model uncertainty). We solve this problem in a similar fashion to dynamic programming with a generalized version of value iteration algorithm and we prove its convergence to a unique solution. This chapter is based entirely on the publication:

- Grau-Moya, J., Leibfried, F., Genewein, T. & Braun, D.A. (2016). **Planning with Information-Processing Constraints and Model Uncertainty in Markov Decision Processes.** ECML/PKDD Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer International Publishing, 2016.

Chapter 6: Nonequilibrium Bounded Rational Decision-Making. The model of bounded rationality exposed in Section 1.3.1 has a clear analogy with equilibrium thermodynamics. In particular, the trade-off between utility and information (entropy) that characterizes the model is also a cornerstone of classic thermodynamics that specifically deals with a similar trade-off between energy and entropy. *Entropy*³ is closely related to the evolutionary description of systems. In thermodynamics, entropy relates and specifies statistically the direction of time of physical processes, or in other words, specifies what process is irreversible and what process is reversible. Reversible processes are idealizations of the laws of nature that are invariant to the time inversion. In this case the entropy increase in the universe is zero. However, by the second law of thermodynamics, irreversible processes are characterized by an increase of entropy (or disorder) in the universe. Irreversibility is ubiquitous in nature, fried eggs do not spontaneously return inside their shells and solutions do not spontaneously separate from the solvent.

When considering that the bounded rational posterior distributions can only be achieved after a certain amount of time we can clearly establish relationships with non-equilibrium thermodynamics. These relationships are interesting because they connect suboptimal utility gains due to limited time with the thermodynamic concept of dissipation (or, equivalently, entropy increase in the universe). We also provide several decision-making examples to illustrate our formalism. Statistical thermodynamics has been a very promising source of inspiration to solve a variety of problems in different fields such as artificial intelligence, machine learning

³Defined by Julius Clausius in 1885, *entropy* in greek means evolution or transformation

and neuroscience. Here we show that it can also be useful for decision-making research. This chapter is based entirely on the publication under review at the moment of writing:

- Grau-Moya, J., Krüeger, M., & Braun, D.A. (2016). **Non-equilibrium relations for bounded rational decision-making in changing environments.** *under review.*

1.4.1 List of Publications and Contributions

In the following I expose an exhaustive list of publications while undergoing the research in this thesis. I only state the contributions of the author and co-authors in the relevant publications presented here.

First author peer-reviewed publications exposed in this thesis.

- Grau-Moya, J., Ortega, P. A., & Braun, D. A. (2012). **Risk-sensitivity in bayesian sensorimotor integration.** PLoS Computational Biology, 8(9), e1002698. ([Grau-Moya et al., 2012](#)).

Conceived and designed the experiments: JGM DAB. Performed the experiments: JGM. Analyzed the data: JGM. Contributed reagents/materials/analysis tools: PAO. Wrote the paper: JGM DAB.

- Grau-Moya, J., Hez, E., Pezzulo, G., & Braun, D. A. (2013). **The effect of model uncertainty on cooperation in sensorimotor interactions.** Journal of The Royal Society Interface, 10(87), 20130554. ([Grau-Moya et al., 2013](#)).

JGM and DAB conceived the idea, JGM and EH performed the experiments and analyzed the data, JGM, GP and DAB wrote the paper.

- Grau-Moya, J., Ortega, P. A., & Braun, D. A. (2016). **Decision-Making under Ambiguity Is Modulated by Visual Framing, but Not by Motor vs. Non-Motor Context. Experiments and an Information-Theoretic Ambiguity Model.** PloS one, 11(4), e0153179. ([Grau-Moya et al., 2016b](#)).

Conceived and designed the experiments: JGM DAB. Performed the experiments: JGM DAB. Analyzed the data: JGM. Contributed reagents/materials/analysis tools: PAO. Wrote the paper: JGM DAB.

- Grau-Moya, J., Leibfried, F., Genewein, T. & Braun, D.A. (2016). **Planning with Information-Processing Constraints and Model Uncertainty in Markov Decision Processes.** ECML/PKDD Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer International Publishing, 2016. ([Grau-Moya et al., 2016a](#)).

JGM conceived the idea, JGM derived main equations and algorithm, FL analyzed convergence, JGM and TG performed experiments, JGM DAB and FL wrote the paper.

- Grau-Moya, J., Krüeger, M., & Braun, D.A. (2016). **Non-equilibrium relations for bounded rational decision-making in changing environments.** *under review.*

JGM and DAB conceived the ideas, JGM derived main equations and performed simulations, MK contributed to critical thinking and discussions, JGM MK and DAB wrote the paper.

Chapter 2

Risk-sensitivity in Bayesian Sensorimotor Integration

This chapter is a reproduction of the already published work in ([Grau-Moya et al., 2012](#)).

Abstract

Information processing in the nervous system during sensorimotor tasks with inherent uncertainty has been shown to be consistent with Bayesian integration. Bayesian estimators are, however, risk-neutral in the sense that they are unbiased and weigh all possibilities purely based on their prior expectations and sensory evidence. In contrast, risk-sensitive estimators are sensitive to model uncertainty and bias their beliefs by taking into account the costs and benefits that their predictions might entail. Here we test for risk-sensitivity in a sensorimotor integration task where subjects exhibit Bayesian information integration. When introducing a cost associated with the estimation process, we found that subjects biased their estimates such that costly events were deemed less probable when uncertainty was high. This result is in accordance with a process of risk-sensitive estimation that allows for deviations from Bayesian probabilities in the face of high uncertainty. Our results suggest that both Bayesian integration and risk-sensitivity are important factors to understand sensorimotor integration in a quantitative fashion.

Introduction

Biological organisms have evolved to succeed in environments with considerable uncertainty (Faisal et al., 2008). One important way of dealing with uncertainty is to develop models of the environment and to form beliefs for prediction. Bayesian statistics provides a powerful and unifying framework to deal with uncertainty not only in the cognitive domain, but also in sensorimotor tasks (Doya et al., 2007). Previous studies have shown that sensorimotor integration in uncertain environments is consistent with Bayesian integration by weighing prior expectations and sensory evidence according to their reliability (Knill and Pouget, 2004; K. P. Körding and Daniel M Wolpert, 2004; K. P. Körding and Daniel M Wolpert, 2006). In particular, it has been shown that the nervous system is able to extract the statistics of variable environments and to incorporate this information by modifying prior beliefs during the process of learning (Turnham et al., 2011). The same formalism can also be used to describe the weighing of information stemming from different sensory modalities with different reliability, for example, when integrating visual and haptic information. A number of previous studies have shown that such multi-modal integration in sensorimotor tasks is also in quantitative agreement with Bayesian statistics (Beers et al., 1999; Ernst and Banks, 2002; Girshick and Banks, 2009).

More generally, internal models are thought to play an important role during sensorimotor processing, for example, to predict sensory consequences of one's actions and to estimate the state of body parts from noisy sensory feedback (Kawato, 1999; Tin and Poon, 2005; Daniel M Wolpert et al., 1995). For example, it has been shown that such estimation is consistent with Kalman filtering, a particular form of Bayesian updating, when subjects had to point to where they believed their hand was after making reaching movements in the dark (Daniel M Wolpert et al., 1995). As a generalization of this, Bayesian updating is also used as a module for estimation in optimal feedback control models (Diedrichsen et al., 2010; Todorov, 2004; Todorov, 2005; Todorov and Jordan, 2002) that have successfully explained a wide range of motor behaviors such as variability pattern (Todorov and Jordan, 2002), the response to of bimanual movements to perturbations (Diedrichsen, 2007; Diedrichsen and Dowling, 2009), adaptation to novel tasks (Braun et al., 2009a; Chen-Harris et al., 2008; Izawa et al., 2008) and complex object manipulation (Nagengast et al., 2009).

While Bayesian beliefs are often thought to reflect actual frequencies of repeatable events, distortions of beliefs are also a widely observed phenomenon—for example, in the case of “wishful thinking” when people overestimate their own abilities (Gregersen, 1996; A. J. L. Harris and Hahn, 2011; Kruger, 1999; Kruger and Dunning, 1999; Start, 1963). Here we are interested in whether similar distortions also occur in non-cognitive domains such as basic sensorimotor behavior. In particular, we are interested in risk-sensitive models of information processing, since they capture model uncertainty and allow for deviations from risk-neutral Bayesian statistics (Hansen and Sargent, 2008). Intuitively, model uncertainty implies that the probabilistic Bayesian model is only trusted to some extent and that probabilities can either be biased towards the worst case outcome or towards the best case outcome, which

corresponds to the risk-attitudes of being risk-averse or risk-seeking.

Bayesian estimators are in general risk-neutral in the sense that they are unbiased and weigh all possibilities purely based on their prior expectations and sensory evidence. In contrast, a risk-sensitive estimator also considers costs or benefits of the beliefs (Ramezani and Marcus, 2005; Whittle, 1981; Whittle, 1990). Consider, for example, a goal keeper that tries to catch a ball flying towards the edge of the goal. Not only will he combine his prior beliefs about velocity, direction, etc. with his sensory evidence, but he will also consider the fact that there are quite different costs depending on which side of the goalpost the ball will most likely end up. In other real-life situations the implications of risk-sensitive estimation could even be more serious, for example when considering evidence for low-probability events like the possibility of a rare disease given some symptoms or the possibility of an aeroplane or a space rocket crashing given a malfunction signal from a noisy detector (Thrun et al., 2002).

Recently, risk-sensitivity has been shown to be an important determinant of motor behavior (Braun et al., 2011a; Nagengast et al., 2010; Nagengast et al., 2011a; Nagengast et al., 2011b). The main finding of these studies was that subjects choose their motor commands not only to optimize the expectation value of some performance criterion, but that they are also sensitive to the variability of the achieved performance measure, which can lead to increased control gains (Nagengast et al., 2010), increased (or decreased) hitting velocities (Nagengast et al., 2011a) and acceptance of decreased mean effort (Nagengast et al., 2011b) in environments where performance is highly variable. However, there is an important aspect of risk-sensitivity that these previous studies have not considered: risk-sensitivity does not only affect the control process, but also the estimation process in uncertain environments with latent task variables that are not directly observable (Whittle, 1981). In uncertain environments with latent variables risk-sensitivity leads to effects of model uncertainty, whereby Bayesian probabilities are biased by the costs that are involved in the control process (Hansen and Sargent, 2008). Crucially, none of the previous studies on risk-sensitivity contained any latent variables. To investigate the effects of risk-sensitivity on the estimation process, we therefore designed a sensorimotor experiment that not only contained a latent variable that needed to be estimated, but we also introduced a cost that was associated with the latent variable. This way we could test whether subjects would bias their beliefs about the latent variable in dependence of the imposed cost function.

2.1 Results

Subjects had to hit a target halfway in a reaching movement to a goal bar by controlling a cursor representing their hand position in a virtual reality set-up (Fig. 2.1). In each trial the lateral position of the target was randomly drawn from a Gaussian distribution. However, the reliability of the visual feedback of the target position was manipulated, such that each trial belonged to one of three feedback conditions: σ_0 , σ_1 or σ_∞ . In the σ_0 -condition the target position was displayed clearly and throughout the trial, corresponding to full information and (practically) zero uncertainty. In the σ_1 -condition only blurry feedback was provided

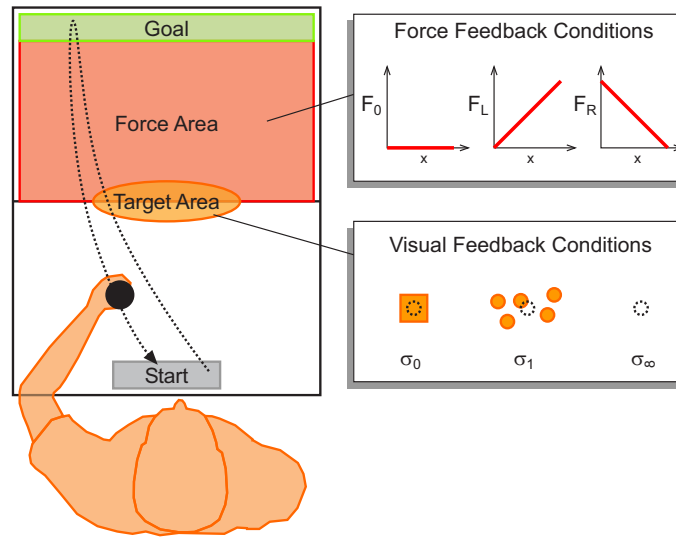


Figure 2.1. Experimental Setup. Subjects move from a start bar to a goal bar and have to hit a target halfway in the reaching movement. In each trial the lateral position of the target was randomly drawn from a Gaussian distribution. The reliability of the visual feedback of the target position was manipulated, such that each trial belonged to one of three feedback conditions: σ_0 , σ_1 or σ_∞ . Furthermore, we imposed three different force functions (F_0 , F_L and F_R) in the force area, where the force depended on subjects’ belief about the target position as they indicated it by their forward movement.

by displaying a short flash of a Gaussian cloud centered around the target. In the σ_∞ -condition no feedback was provided. Naturally, the probability of hitting the target decreased with increasing feedback uncertainty—compare Supplementary Figure 2.4. In this setup, the lateral target position constitutes a latent variable that needs to be estimated in every trial from noisy feedback. The aim is to study subjects’ beliefs about this latent variable and to study the susceptibility of their beliefs to risk-sensitive distortions.

Previous studies have shown that human sensorimotor integration of feedback information with varying degrees of reliability can be understood by Bayesian models (K. P. Körding and Daniel M Wolpert, 2004). In particular, it has been shown that subjects rely more on their prior information when the quality of their sensory feedback gets worse. This can be seen in Figure 2.2 which shows a typical subject’s lateral deviation from the target as a function of the target position (red lines). In the full feedback condition (σ_0) the lateral deviation was close to zero, as subjects could see the target clearly. In contrast, in the no-feedback condition (σ_∞) subjects had to rely on their prior about the target position and should ideally move through the point of maximum prior probability—which is zero in our case, such that the lateral deviation as a function of the target position is described by the identity line. The subject’s behavior in the third panel of Figure 2.2 conforms to this prediction. Furthermore, the model predicts that in the σ_1 -condition subjects should mix prior beliefs with sensory feedback, leading to an intermediate slope for the lateral deviation. We also found this effect in our

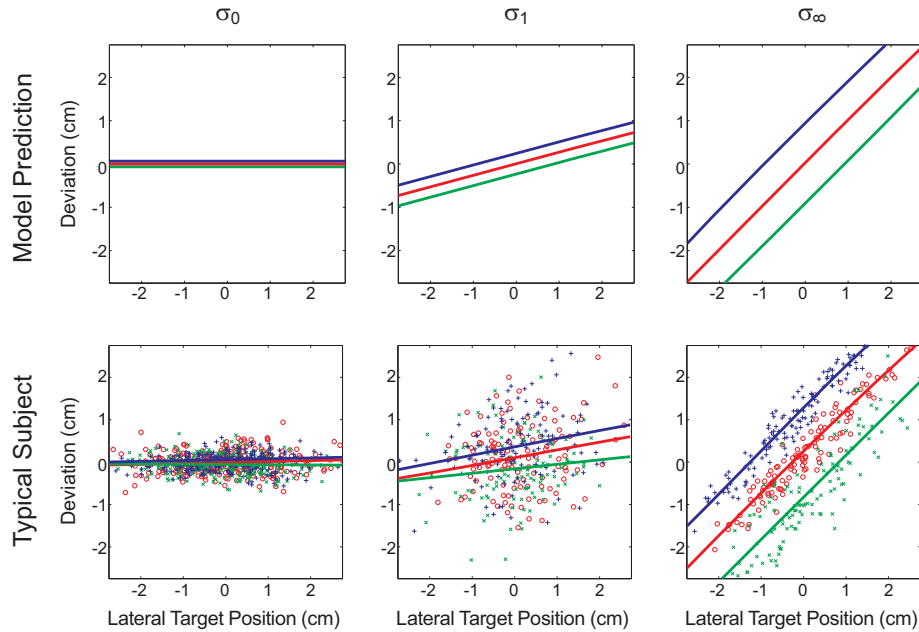


Figure 2.2. Lateral deviation from target as a function of target position in a risk-sensitive model (top row) and in a typical subject (bottom row). The three columns correspond to the three levels of uncertainty of the target feedback (σ_0 , σ_1 and σ_∞). Each panel compares the three different force conditions F_0 (red), F_L (green) and F_R (blue). The model predicts that higher levels of uncertainty are associated with higher slopes and that higher forces are associated with shifts in the intercept that are proportional to the uncertainty.

subjects as displayed in the second panel of Figure 2.2. In summary, when comparing the red lines of the three panels of Figure 2.2, it can be seen that the slope of the lateral deviation increases with the uncertainty, which is exactly what previous studies have reported (K. P. Körding and Daniel M Wolpert, 2004).

To investigate effects of risk-sensitivity we introduced a force landscape that assigned different costs to different beliefs about the target position. The force landscape was given by a viscous force in the forward-backward direction during the second half of the movement between target and goal bar—this is indicated as the red force area in Figure 2.1. We imposed three different force functions (F_0 , F_L and F_R) that were presented consecutively to subjects in three blocks of 750 trials each. The F_0 -function was applied in the first block and corresponded to a zero force condition. The force F_L (“easy left”) was presented in the second block and corresponded to a linear function that increased from left to right. Therefore, pointing to a target position on the left required less effort than pointing to a target position on the right of the center of the target distribution. Finally, the force F_R (“easy right”) was presented in the last block and corresponded to a linear function that decreased from left to right—see Methods for details.

Assigning different costs to different beliefs, predicts an interesting interaction between

uncertainty and cost for a risk-sensitive estimator. In the absence of uncertainty (σ_0 -condition) there is no risk and a risk-sensitive estimator will produce the same predictions as a risk-neutral estimator that is independent of the imposed cost. However, in the presence of uncertainty, there is risk involved and a risk-sensitive estimator will bias its predictions based on cost. Having uncertainty about the target position implies that a risk-sensitive estimator has to consider a range of possible target positions and essentially “hopes” that the target is in one of the possible positions that requires less effort. In the case of linear force functions this “bias” in the belief translates into a parallel shift of the line that describes subjects’ lateral deviation. The magnitude of the shift depends on the uncertainty of the target position, the cost of the presumed target position and subjects’ risk-sensitivity. This prediction can be seen in Figure 2.2.

When reaching for the target, subjects had to combine prior information about the distribution of target positions, visual feedback and the cost of the pointing movement. We examined how they combined these three factors in the following way. For each force block (F_0 , F_L and F_R) we conducted three linear regressions corresponding to the three feedback conditions (σ_0 , σ_1 or σ_∞). In each case we regressed the lateral deviation of subjects’ pointing movement against the true target position and determined slope and intercept of this line. According to the model predictions in Figure 2.2, the slope should only depend on the uncertainty of the feedback independently of the force condition, whereas the intercept should depend on both the cost given by the force and the uncertainty given by the feedback condition.

The slopes and intercepts fitted to every subject are shown in Figure 2.3. In the upper panels of Figure 2.3, one can see that the slopes describing subjects’ lateral deviation increased with higher levels of uncertainty within each force block. This is in line with the prediction and reproduces previous findings. Moreover, in accordance with the prediction from Figure 2.2, this slope increase was not affected by the force condition. To assess the statistical significance of this result we conducted a repeated-measures two-way ANOVA with force and uncertainty as factors. We found that the uncertainty had a significant effect on the slope ($p < 0.01$), whereas the effect of force was not significant ($p > 0.4$).

In the lower panels of Figure 2.3, one can see subjects’ intercepts that describe their mean lateral deviations from a reference target located in the center of the workspace (zero position). In accordance with the prediction from Figure 2.2, our ANOVA revealed that intercepts were affected by both uncertainty ($p < 0.01$) and force condition ($p < 0.01$). In the no-force condition the intercepts are close to zero for all uncertainty levels, as subjects have no incentive to deviate from an unbiased Bayesian estimator. In the force conditions F_L , we found that the intercepts become increasingly negative with growing uncertainty. This means that subjects’ beliefs were biased towards the left, as those beliefs were associated with lower costs. Compared to the no-force condition, subjects deviated on average $8.1 \pm 0.5mm$ more to the left in the no-feedback condition and $2.2 \pm 0.4mm$ more to the left in the σ_1 -condition. Similarly, in the F_R force condition, we found that intercepts increased with growing uncertainty reflecting a low-cost bias towards the right side of the workspace. Compared to the no-force condition,

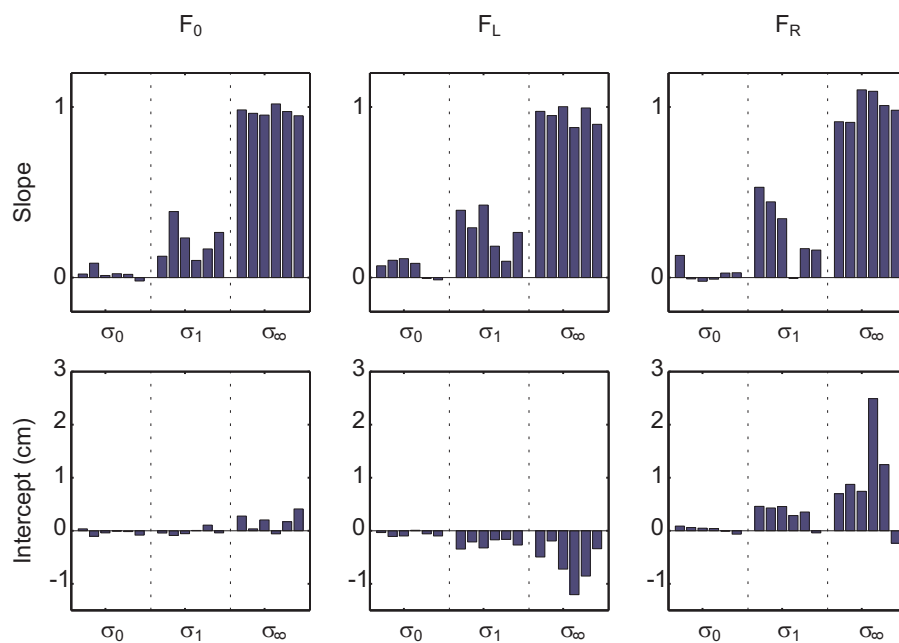


Figure 2.3. Slopes (top row) and intercepts (bottom row) of linear regression for all subjects. Linear regression was performed as in Figure 2.2. The three columns correspond to the three different force conditions F_0 , F_L and F_R . The three different feedback conditions σ_0 , σ_1 and σ_∞ are displayed within each panel. It can be seen that the slope increases with increasing uncertainty. The intercepts are modulated by both uncertainty and force condition.

subjects on average deviated $8.0 \pm 0.5mm$ more to the right in the no-feedback condition and $3.2 \pm 0.4mm$ more to the right in the σ_1 -condition. All subjects but one exhibited this bias pattern—compare Figure 2.3.

Importantly, the model of risk-sensitive estimation not only predicts a fixed bias, but a modulation of bias and uncertainty, such that the bias increases with the amount of uncertainty and vanishes in the limit when uncertainty is absent. In accordance with this prediction, we found that the mean lateral deviations from the center of the target in the σ_0 -condition are negligible in all force conditions. The exact values of the mean lateral deviations were $-0.4 \pm 0.1mm$ in the F_0 -condition, $-0.6 \pm 0.1mm$ in the F_L -condition, and $+0.4 \pm 0.1mm$ in the F_R -condition—all well within the target halfwidth of $2.0mm$. Similarly, the lateral deviations from the center of the starting position at the beginning of the trial was not significantly different between the groups ($p > 0.05$, repeated measures one-way ANOVA). The exact values of the mean lateral deviations were $-1.1 \pm 1.2mm$ in the F_0 -condition, $-1.9 \pm 1.0mm$ in the F_L -condition, and $+0.7 \pm 1.7mm$ in the F_R -condition—all well within the target halfwidth of $2.0mm$. In summary, these results suggests that subjects did not simply avoid high costs, but that their behavior was determined by an interplay of uncertainty and cost as predicted by a risk-sensitive estimation process.

2.2 Discussion

In our study we examined the effects of risk-sensitivity on sensorimotor integration. In line with previous studies, we found that information integration was consistent with Bayesian statistics as long as beliefs are cost-neutral (K. P. Körding and Daniel M Wolpert, 2004). However, once we introduced a cost that was associated to the beliefs, subjects started to bias their beliefs when faced with uncertain feedback. Importantly, subjects did not simply minimize their effort, but they modulated their beliefs based on an interplay between cost and uncertainty. In particular, we found that the higher the uncertainty, the higher the bias. When sensory feedback was unambiguous—i.e. in the (near) absence of uncertainty—this bias vanished. This is in accordance with the predictions of a risk-sensitive estimation process, but violates risk-neutral Bayesian estimation.

Previous studies have found that risk-sensitivity is an important determinant of motor behavior (Braun et al., 2011a). The main finding of these studies was that subjects not only optimize their expectation of success, but also take the performance variability into account. For example, a basket ball player choosing between throwing a three with a 50% success rate and throwing a two with a 75% success rate would prefer the first option if risk-seeking, the second option if risk-averse, and he would be indifferent if risk-neutral. These previous studies have found that risk-sensitive motor behavior can be accounted for by a mean-variance trade-off (Nagengast et al., 2011b) that affects control gains and the speed-accuracy trade-off when performance success becomes more variable (Nagengast et al., 2010; Nagengast et al., 2011a). Importantly, the effects of risk-sensitivity on the estimation process could not be investigated in these previous studies, because they did not contain any latent variables that would have

required estimation.

The differential effects of risk-sensitivity on control and estimation can be readily inspected in the case of risk-sensitive control of linear systems with quadratic costs and Gaussian noise—sometimes abbreviated to risk-sensitive LQG control (Whittle, 1981). The standard LQG control that has often been used in optimal feedback control models of motor behavior (Todorov and Jordan, 2002) can be derived as a special case of the risk-sensitive LQG control in the limit of vanishing risk-sensitivity. Importantly, in risk-neutral LQG controllers the estimation and control processes can be separated such that the solution to the estimation problem is given by the Kalman filter and the solution to the LQ control problem is given by the solution of the Riccati equation in the absence of observation noise. The overall solution to the LQG system is then simply given by the LQ optimal controller where all directly observed variables are replaced by their estimates from the Kalman filter. In the context of our experiment it is important to point out that the risk-neutral estimation process of the Kalman filter does not depend on the cost function of the control process. The beliefs expressed by the Kalman filter are unbiased.

In risk-sensitive LQG systems a separation between control and estimation is still possible (Whittle, 1981), however, with an interesting interplay between estimation and control that is absent in risk-neutral systems. For example, in contrast to risk-neutral estimation, the estimation process of risk-sensitive LQG control can be expressed by a modified Kalman filter that is modulated by the cost function of the control process. This way the beliefs expressed by this modified Kalman filter become biased. The risk-sensitive estimator that we propose in the Methods section corresponds exactly to such a modified Kalman filter—compare Supplementary Material Part I 2.4. Similarly, in the absence of observation noise the solution to the risk-sensitive LQ control problem is given by the solution of a modified Riccati equation. Effects of this modification of the control process have been studied in (Nagengast et al., 2010), for example, where an increase in control gains was found in response to increased process noise that determined the Brownian motion of a virtual ball. However, the observation noise was entirely negligible compared to the process noise in this task, so effects of risk-sensitive estimation did not play any role in this experiment.

An important problem when studying “beliefs” about latent variables is that beliefs are not directly observable, but can only be inferred from observing actions—see for example the notes in (Daniel M Wolpert et al., 1995) for a discussion of a similar problem. In our task the action simply consisted in reporting the belief about the latent variable—as was for example the case in previous studies (K. P. Körding and Daniel M Wolpert, 2004). This can also be seen within the framework of risk-sensitive LQG control, where the control is simply given by the estimate of the modified Kalman filter if we impose a cost on the latent variable—compare Supplementary Material Part I 2.4. In this case control and belief do not need to be further disentangled, because they are essentially the same by design of the experiment. Interestingly, if one wanted to model the experiment by introducing a control cost instead of a cost that is associated to the latent variable, LQG control predicts a constant shift of the control in the presence of force fields that is not modulated by the feedback uncertainty—

compare Supplementary Material Part II 2.4. We can therefore rule out a risk-neutral account of our experiment that is based on the expectation value of control costs.

What makes risk-sensitivity especially interesting in the context of Bayesian inference is that it has also been related to model uncertainty (Hansen and Sargent, 2008). Model uncertainty allows a decision-maker who has a probabilistic model of the environment to deviate from this model if he trusts this model only to a limited extent. In particular, an infinitely pessimistic decision-maker would disregard the probabilistic model entirely and only focus on worst-case outcomes. Since all models are typically prone to error at some precision, taking into account model uncertainty is a crucial aspect of inference.

The distortion of subjective probabilities has also been investigated within the context of prospect theory (D Kahneman and A Tversky, 1979). Subjects are typically found to overweigh low-probability events like plane crashes or natural disasters, whereas they underweigh high-probability events. Interestingly, the opposite pattern has been observed when fitting prospect theory to choice behavior in motor control (Nagengast et al., 2011b; Wu et al., 2009). However, it was also found (Nagengast et al., 2011b) that some choice behavior explicable in terms of prospect theory can also be explained by the mean-variance trade-off as suggested by risk-sensitive control models (Braun et al., 2011a). Our current study provides evidence that risk-sensitivity could also serve as an alternative explanation of the distortion of probabilities and might underlie biasing effects in sensorimotor behavior that is analogous to some cognitive biases.

2.3 Materials and Methods

Subjects. Two female and four male subjects from the Tübingen University student population participated in this experiment after giving informed consent. All experimental procedures were approved by the ethics committee of the medical faculty at the university of Tübingen. Participants were paid the local standard rate of 8 Euros per hour for their participation.

Materials. The experiment was conducted using a vBOT robotic manipulandum (Ian S Howard et al., 2009). Participants controlled the vBOT handle in the horizontal plane. Movement position and velocity were recorded at a rate of $1kHz$. A planar virtual reality projection system was used to overlay images into the plane of movement of the vBOT handle.

Experimental Procedure. Subjects performed reaching movements from a start bar (gray rectangle, width $4cm$, height $1.5cm$) to a goal bar (green rectangle, width $14cm$, height $0.5cm$) $25cm$ away by moving a cursor (red circle, $3mm$ radius) representing their hand position—compare Figure 2.1. The hand position was represented veridically at all times. Subjects could start anywhere from within the start bar and they were told to hit a yellow target that would appear midway during the forward movement to the green bar. When placing the cursor on the start bar, participants heard a beep that informed them to move. At the same

time the target appeared midway at a distance of 12.5cm from the start bar with a lateral displacement drawn from a Gaussian distribution with zero mean and standard deviation $\sigma_p = 1.0\text{cm}$. Movements had to be completed within 0.6s .

In each trial the target position was displayed under one out of three possible feedback conditions ($\sigma_0, \sigma_1, \sigma_\infty$) selected randomly with relative frequencies of (2,1,1) respectively. In the σ_0 -condition, the target was displayed during the whole trial as a small rectangle of 4mm width. The displayed height of the target was 10mm , but only relevant for visualization purposes without consequence for the hitting probability. In the σ_1 -condition, five small circles (radius 2mm) were drawn each trial from a two-dimensional Gaussian distribution (mean 0cm , standard deviation 1.5cm) and shown for 80ms at the beginning of the trial. No feedback was provided in the σ_∞ -condition. In all three conditions subjects had to make a choice in the lateral position h when they were halfway in the movement (12.5cm from the start bar) in order to indicate their belief about the target position. Halfway into the movement they also received auditory feedback, which was a high frequency beep if they hit the target or a low frequency beep if they failed to do so. Another beep of the same frequency informed them when they reached the goal bar.

Between the target and the goal bar subjects entered a “force zone” in which they experienced a viscous force $F = -k(h) \cdot v$ that made movements more strenuous. The viscous force was applied in the forward-backward direction and was proportional to the forward or backward velocity v . The force was also applied in the force zone while subjects returned to the start position to initiate the next trial. The strength $k(h)$ of the force depended only on subjects’ movement position h halfway into the movement (12.5cm from the start bar). To allow for a smooth transition from the no-force zone to the force zone the viscous force was ramped up linearly over the first quarter of the force zone and similarly ramped down during the backward movement. There were three force conditions: F_0 , F_L and F_R . In the F_0 condition there was no force, that is $k(h) \equiv 0$. In the F_L condition the strength $k(h) = ah + b$ was a linear function with $a = 60 \frac{\text{kg}}{\text{cm}\cdot\text{s}}$ and $b = 90 \frac{\text{kg}}{\text{s}}$, such that it increased linearly from left $k_{min} = 0 \frac{\text{kg}}{\text{s}}$ to right $k_{max} = 180 \frac{\text{kg}}{\text{s}}$ over a 3cm range centered around the mean of the target distribution. In the F_R condition the slope was simply inverted to obtain a linear function with $a = -60 \frac{\text{kg}}{\text{cm}\cdot\text{s}}$ and $b = 90 \frac{\text{kg}}{\text{s}}$ that increased linearly from right $k_{min} = 0 \frac{\text{kg}}{\text{s}}$ to left $k_{max} = 180 \frac{\text{kg}}{\text{s}}$ over the same 3cm range.

The experiment consisted of 2250 trials in total and was subdivided in three blocks of 750 trials each corresponding to the three force conditions F_0 , F_L and F_R . In every block of 750 trials only the last 500 were used for analysis, as movement variability in σ_0 -trials had then stabilized—compare Supplementary Figure 2.5.

Risk-neutral estimator. Each trial a target with lateral position h is drawn from a Gaussian distribution with mean zero and standard deviation σ_p . Subjects receive noisy sensory feedback about the target position given by the observation x . We model this noisy feedback by another Gaussian distribution with mean h and standard deviation given by σ_i where $i = \{0, 1, \infty\}$. The Bayesian estimator of the target position given the observation is then

given by

$$p(h|x) = \underbrace{\frac{1}{\sqrt{2\pi}\sigma_i} e^{-\frac{1}{2} \frac{(x-h)^2}{\sigma_i^2}}}_{\text{likelihood}} \underbrace{\frac{1}{\sqrt{2\pi}\sigma_p} e^{-\frac{1}{2} \frac{h^2}{\sigma_p^2}}}_{\text{prior}}$$

In the case of the Gaussian posterior $p(h|x)$ the optimal point estimate h^* that maximizes the a posteriori probability also corresponds to the mean of the Gaussian which is

$$h^* = \frac{\sigma_p^2}{\sigma_p^2 + \sigma_i^2} x.$$

Risk-sensitive estimator. A risk-sensitive estimator does not only depend on the probabilities $p(h|x)$, but also on the costs that are assigned to the hidden variable h (Thrun et al., 2002; Whittle, 1981; Whittle, 1990). In our experiments each target position was assigned a cost $c(h)$ of the form $c(h) = a_j h + b_j$ with $j = \{0, R, L\}$. The cost $c(h)$ models the experimental viscosity function $k(h)$ described in the Experimental Procedures. The parameters a_j and b_j depend on the force condition, where $a_0 = b_0 = 0$ in the F_0 -condition and $a_R = -a_L$ and $b_R = b_L$ in the other force conditions. The cost $c(h)$ is taken into account by a risk-sensitive estimator with risk-parameter α to obtain the biased estimator

$$q_\alpha(h|x) = \frac{p(h|x)e^{\alpha c(h)}}{\int dh' p(h'|x)e^{\alpha c(h')},$$

where $p(h|x)$ is the risk-neutral Bayesian posterior.

Again the optimal point estimate h_α^* that maximizes the a posteriori probability of q_α corresponds to the expectation value which is

$$h_\alpha^* = \frac{\sigma_p^2}{\sigma_i^2 + \sigma_p^2} x - \frac{\sigma_i^2 \sigma_p^2}{\sigma_i^2 + \sigma_p^2} \alpha a_j.$$

For $\alpha \rightarrow 0$ the risk-sensitive point estimator becomes the risk-neutral Bayesian maximum a posteriori estimator, which is given by the first term. The second term incorporates an interaction between marginal cost a_j and the uncertainty given by σ_i and σ_p .

2.4 Supplementary Material

Linear Quadratic Gaussian risk-sensitive control

Part I: Modeling the force as a cost of a latent state

In order to fit with the formalism proposed in (Whittle, 1981), we can translate our experiment into a 3-step system with the following scalar variables

$$\begin{aligned}x_1 &= x_0 + \epsilon \\y_1 &= x_0 + \eta_\infty \\c_1 &= kx_1\end{aligned}$$

$$\begin{aligned}x_2 &= x_1 \\y_2 &= x_1 + \eta \\c_2 &= 0\end{aligned}$$

$$\begin{aligned}x_3 &= x_2 + u_2 \\y_3 &= x_2 + \eta_\infty \\c_3 &= x_3Qx_3.\end{aligned}$$

The integral cost is given by $J = \sum_{t=1}^3 c_t = x_3Qx_3 + kx_1$, where the first term enforces that the difference between the control signal and the target position is minimized, and the second term is a linear state-dependent cost that models the force cost that we imposed in our experiment. The risk-sensitive *stress* function $\gamma(\theta)$ that is to be minimized is given by

$$\gamma(\theta) = -\frac{2}{\theta} \log \mathbb{E} \left[e^{-\frac{\theta}{2}J} \right].$$

The system evolves as follows:

- In the first time step, the target position x_1 is drawn from a Gaussian distribution with mean x_0 and variance σ_n^2 , that is $\epsilon \sim \mathcal{N}(0, \sigma_n^2)$. The mean is assumed to be known precisely, that is $\hat{x}_0 = x_0$ with variance $V_0 = 0$. No observation is made, or formally $\eta_\infty \sim \mathcal{N}(0, \infty)$. No control is applied, that is $u_0 = 0$.
- In the second time step, a noisy observation y_2 of the target position x_1 is made. The observation noise is additive and drawn from a Gaussian distribution with $\eta \sim \mathcal{N}(0, \sigma_m^2)$. The target position does not change during the observation, that is $x_2 = x_1$. No control is applied, that is $u_1 = 0$.
- In the third time step, a control command u_2 can be applied to minimize the quadratic cost $(x_2 + u_2)Q(x_2 + u_2)$, which implies that the control should match the target position. No further observations are made.

Minimizing the stress function $\gamma(\theta)$ can be achieved by computing the *past stress* $P_t(x_t)$ for estimation and the *future stress* $F_t(x_t)$ for control. Whittle (Whittle, 1981) derived the following recursions for past and future stress:

$$P_{t+1}(x_{t+1}) = \text{ext}_{x+t} \left[P_t(x_t) + c_t + \frac{1}{\theta} (n_t + m_t) \right]$$

$$F_t(x_t) = \min_{u_t} \text{ext}_{x_{t+1}} \left[c_t + \frac{1}{\theta} n_t + F_{t+1}(x_{t+1}) \right],$$

where “ext” indicates min or max depending on the sign of θ and the shorthands n_t and m_t are given by

$$n_t = (x_{t+1} - x_t - u_t)^2 \sigma_n^{-2}$$

$$m_t = (y_{t+1} - x_t)^2 \sigma_m^{-2}.$$

Whittle (Whittle, 1981) could show that the optimal control u_t^{opt} that minimizes $\gamma(\theta)$ can be computed by finding the u_t^* that minimizes the future stress $F_t(x_t)$ and finding the \bar{x}_t that extremizes the combined stress $P_t(x_t) + F_t(x_t)$, such that $u_t^{\text{opt}} = u_t^*(\bar{x}_t, t)$. This establishes a risk-sensitive version of certainty-equivalence, where minimizing the past stress leads to a risk-sensitive version of the Kalman filter with recursively updated estimates \hat{x}_t (mean) and V_t (variance) representing a Gaussian belief. For our system equations this results in the following:

Initialization:

$$P(x_0) = 0$$

$$\hat{x}_0 = x_0$$

$$V_0 = 0$$

First time step:

$$P(x_1) = \frac{1}{\theta} (x_1 - \hat{x}_0)^2 \sigma_n^{-2}$$

$$\hat{x}_1 = \hat{x}_0$$

$$V_1 = \sigma_n^2$$

Second time step:

$$P(x_2) = kx_2 + \frac{1}{\theta}(x_2 - \hat{x}_0)^2\sigma_n^{-2} + \frac{1}{\theta}(y_2 - x_2)^2\sigma_m^{-2}$$

$$\hat{x}_2 = \frac{\sigma_n^{-2}\hat{x}_0 + \sigma_m^{-2}y_2 - \frac{\theta}{2}k}{\sigma_n^{-2} + \sigma_m^{-2}}$$

$$V_2 = (\sigma_n^{-2} + \sigma_m^{-2})^{-1}$$

$$F(x_2) = \min_{u_2} \{Q(x_2 + u_2)^2\} = 0$$

$$u_2^* = -x_2$$

$$\bar{x}_2 = \hat{x}_2$$

$$u_2^{\text{opt}} = -\hat{x}_2$$

Consequently, in this system the control signal u_2 directly reveals the risk-sensitive estimate \hat{x}_2 . The estimate \hat{x}_2 gives the same value as the risk-sensitive estimator provided in the Methods section.

Part II: Modeling the force as a control cost

If we assume the same system as in the previous section, but now with costs

$$c_1 = 0$$

$$c_2 = ku_2$$

$$c_3 = x_3Qx_3,$$

where we have exchanged the state-dependent cost $c_1 = kx_1$ with a control-dependent cost $c_2 = ku_2$, then we obtain

Initialization:

$$P(x_0) = 0$$

$$\hat{x}_0 = x_0$$

$$V_0 = 0$$

First time step:

$$P(x_1) = \frac{1}{\theta}(x_1 - \hat{x}_0)^2\sigma_n^{-2}$$

$$\hat{x}_1 = \hat{x}_0$$

$$V_1 = \sigma_n^2$$

Second time step:

$$\begin{aligned}
 P(x_2) &= \frac{1}{\theta}(x_2 - \hat{x}_0)^2 \sigma_n^{-2} + \frac{1}{\theta}(y_2 - x_2)^2 \sigma_m^{-2} \\
 \hat{x}_2 &= \frac{\sigma_n^{-2} \hat{x}_0 + \sigma_m^{-2} y_2}{\sigma_n^{-2} + \sigma_m^{-2}} \\
 V_2 &= (\sigma_n^{-2} + \sigma_m^{-2})^{-1} \\
 F(x_2) &= \min_{u_2} \{Q(x_2 + u_2)^2 + k u_2\} = 0 \\
 u_2^* &= -x_2 - \frac{k}{2Q} \\
 \bar{x}_2 &= \frac{\sigma_n^{-2} \hat{x}_0 + \sigma_m^{-2} y_2 + \frac{1}{2} \theta k}{\sigma_m^{-2} + \sigma_n^{-2}} \\
 u_2^{\text{opt}} &= -\bar{x}_2 - \frac{k}{2Q} = -\hat{x}_2 - \frac{k\theta}{2\sigma_n^{-2}\sigma_m^{-2}} - \frac{k}{2Q}
 \end{aligned}$$

In this system the first two terms correspond to the control of the previous section and the last term represents a shift of $-\frac{k}{2Q}$ in presence of the force field—independent of the risk-sensitivity or the uncertainty associated with the latent variable. In particular, a risk-neutral controller ($\theta = 0$) that simply minimizes the expected cost would only show the last deviation, that is a constant shift for all uncertainty conditions. This is in contrast to our experiment where we found that subjects' modulated their response depending on their uncertainty.

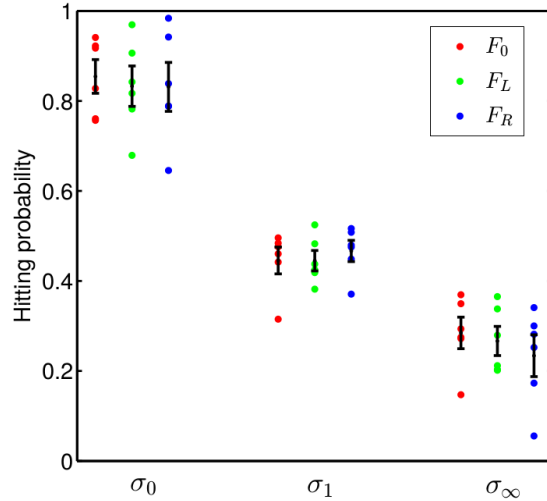


Figure 2.4. Hitting probabilities. Success probability of hitting the target for the three different feedback conditions (σ_0 , σ_1 and σ_∞) and the three different force conditions F_0 (red), F_L (green) and F_R (blue). The hitting probability decreases with increasing feedback uncertainty.

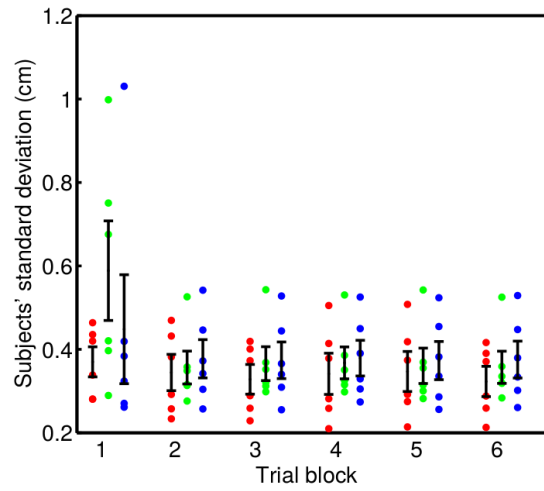


Figure 2.5. Movement variability. Standard deviation of hitting movements in trials of the σ_0 -condition over blocks of 125 trials. Since variability was increased for some subjects in the first block, we only analyzed the last 500 trials of each force condition in the experiment.

Acknowledgments

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

Chapter 3

Framing Effects in Decision-Making under Ambiguity

This chapter is a reproduction of the already published work in ([Grau-Moya et al., 2016b](#)).

Abstract

A number of recent studies have investigated differences in human choice behavior depending on task framing, especially comparing economic decision-making to choice behavior in equivalent sensorimotor tasks. Here we test whether decision-making under ambiguity exhibits effects of task framing in motor vs. non-motor context. In a first experiment, we designed an experience-based urn task with varying degrees of ambiguity and an equivalent motor task where subjects chose between hitting partially occluded targets. In a second experiment, we controlled for the different stimulus design in the two tasks by introducing an urn task with bar stimuli matching those in the motor task. We found ambiguity attitudes to be mainly influenced by stimulus design. In particular, we found that the same subjects tended to be ambiguity-preferring when choosing between ambiguous bar stimuli, but ambiguity-avoiding when choosing between ambiguous urn sample stimuli. In contrast, subjects' choice pattern was not affected by changing from a target hitting task to a non-motor context when keeping the stimulus design unchanged. In both tasks subjects' choice behavior was continuously modulated by the degree of ambiguity. We show that this modulation of behavior can be explained by an information-theoretic model of ambiguity that generalizes Bayes-optimal decision-making by combining Bayesian inference with robust decision-making under model uncertainty. Our results demonstrate the benefits of information-theoretic models of decision-making under varying degrees of ambiguity for a given context, but also demonstrate the sensitivity of ambiguity attitudes across contexts that theoretical models struggle to explain.

Introduction

Should you continue reading this paper? The uncertainty involved in this decision is difficult to quantify. This is in contrast to uncertainties arising for example in dice or roulette games, where the decision-maker has a pretty good idea of the probabilities that are involved, even though individual outcomes cannot be predicted. In the economic literature there is a long-standing debate about known vs. unknown uncertainty (Knight, 1921), sometimes also called risk vs. ambiguity. The question is, whether these two kinds of uncertainty are the same or whether they are processed in a different way by human decision-makers. This question has been famously addressed by Ellsberg in what is now an eponymous experiment (Ellsberg, 1961). In a simplified version, it requires subjects to choose between a *risky urn* with a known composition of differently colored balls, for example 50 blue balls and 50 red balls, and an *ambiguous urn* with an unknown color composition, for example 100 balls with unknown proportion of blue and red. When setting a prize on drawing a blue ball, most subjects (typically around 70% (Pulford and Coleman, 2008)) prefer drawing from the risky urn, implying the belief that there are more blue balls in the risky urn than in the ambiguous one. The paradox arises when leaving the urns untouched and swapping the prize money. When setting a prize on drawing red, most subjects still prefer drawing from the risky urn, implying the belief that there are more red balls in the risky urn than in the ambiguous one. Crucially, there is no single probability that can represent the two beliefs that there are simultaneously more blue balls and more red balls in the risky urn than in the ambiguous urn. Ever since the experiments of Ellsberg there has been growing evidence, both behaviorally (Camerer and Weber, 1992; Keren and Gervitsen, 1999) and neurally (Chumbley et al., 2012; Hsu et al., 2005; Huettel et al., 2006; Krain et al., 2006; I. Levy et al., 2010; Smith et al., 2002), that there are indeed two different kinds of uncertainty considered by humans engaged in economic decision-making. However, it is unclear how ambiguity is modulated by the task context and by framing.

Previous studies have investigated, for example, how decision-making in sensorimotor tasks compares to economic pen-and-paper decision-making. A number of these studies have reported that the human sensorimotor system operates in line with expected utility theory, that is Bayes-optimal decision-making with known probabilities (Braun et al., 2009a; Diedrichsen et al., 2010; K. Körding, 2007; K. P. Körding and Daniel M Wolpert, 2004; Todorov, 2004; Trommershäuser et al., 2003b; Trommershäuser et al., 2008; Daniel M. Wolpert and Landy, 2012). Other studies have shown discrepancies of sensorimotor decision-making with Bayes-optimal decision-making. In particular, Wu et al. (Wu et al., 2009) have previously compared economic decision-making with an equivalent motor task where participants had to choose between different targets they had to hit. In particular, they investigated a well-known decision-making paradox under risk—the so-called Allais paradox—and its occurrence in the two types of tasks. They found that subjects had different attitudes towards risk in the two tasks but did not investigate the origin of this behavioral difference. Additionally, in their motor task the targets were always fully visible and, therefore, were not subject to ambiguity.

In this study we ask the same question as Wu et al. (Wu et al., 2009) did for risk in the Allais paradox now for ambiguity in the Ellsberg paradox. We investigate a generalized version of Ellsberg’s paradox in decision-making under ambiguity and test how ambiguity is modulated by task context and framing. Similar to Wu et al. (Wu et al., 2009), we compare motor and non-motor context. As a motor context we use a target hitting task, as a non-motor context we use an urn task. Additionally, we investigate the effects of visual framing by manipulating the stimulus presentation, in particular the way how uncertainty is visually displayed. Finally, we compare Bayes-optimal expected utility predictions to predictions of an information-theoretic free energy model of decision-making under varying degrees of ambiguity.

3.1 Results

3.1.1 An Information-Theoretic Model of Decision-Making under Ambiguity

In Ellsberg’s urn experiment subjects have to choose between two options, a risky urn and a fully ambiguous urn. We generalize this paradigm by also including partially ambiguous urns, which can be experimentally achieved for example by revealing samples from the ambiguous urn with unknown ratio. We assume that subjects’ choice between the risky option x_{risk} and the ambiguous option x_{amb} can be described by a probability distribution $p(x)$ with $x \in \{x_{\text{amb}}, x_{\text{risk}}\}$, and that subjects have no prior preference between the options, that is $p_0(x) = 1/2$. Each option x is characterized by a latent variable h corresponding to the ratio of blue and red balls. Each h implies a utility $U(h)$ indicating the expected payoff under h . For the risky option h is known, for the ambiguous option it is unknown. The decision-maker holds a Bayesian belief $q(h|x, D)$ about h for option x after observing data D corresponding for example to the observed samples in the urn experiment. Accordingly, we have the belief $q(h|x_{\text{amb}}, D)$ for the ambiguous option and the belief $q(h|x_{\text{risk}}, D) = \delta(h - h^*)$ for the risky option with a ratio h^* of red and blue balls.

The crucial point of Ellsberg’s original experiment was to show that standard models of economic decision-making that only care about maximizing expected utility cannot explain subjects’ choice behavior under ambiguity. In our experiment an expected utility maximizer would assign the value V_0 to option x according to

$$V_0(x) = \mathbb{E}_{q(h|x,D)}[U(h)]. \quad (3.1)$$

A perfect expected utility maximizer chooses the option $x^* = \operatorname{argmax}_x V_0(x)$ that maximizes the overall expected utility. A more general imperfect expected utility maximizer can be modeled for example by a soft-max decision rule, such that the decision-maker chooses according to hypothesis \mathbf{H}_1

$$\mathbf{H}_1 : \quad p_1(x) = \frac{e^{\alpha V_0(x)}}{\sum_{x'} e^{\alpha V_0(x')}} \quad (3.2)$$

with the soft-max parameter α .

Our alternative hypothesis \mathbf{H}_2 is that the decision-maker optimizes a free energy function that trades off utilities against information-theoretic constraints that can be derived from axiomatic principles (Braun et al., 2011b; Ortega and Braun, 2011; Ortega and Braun, 2013). Such information-theoretic constraints can reflect for example a lack of available information which makes them interesting for modeling ambiguity. Intuitively, such a decision-maker is sensitive to ambiguity by biasing their belief q towards best-case or worst-case utilities depending on whether the decision-maker is ambiguity-seeking or ambiguity-averse. Such ambiguity-sensitive decision-makers would assign the value $V_\beta(x)$ to option x , where

$$\begin{aligned} V_\beta(x) &= \underset{\tilde{q}(h|x,D)}{\text{ext}} \left\{ \mathbb{E}_{\tilde{q}(h|x,D)} [U(h)] - \frac{1}{\beta} D_{\text{KL}}(\tilde{q}(h|x,D) || q(h|x,D)) \right\} \\ &= \frac{1}{\beta} \log \mathbb{E}_{q(h|x,D)} \left[e^{\beta U(h)} \right] \end{aligned} \quad (3.3)$$

This valuation allows for pessimistic deviations from the Bayesian posterior q towards worst-case (ext = min) outcomes if the decision-maker is ambiguity-averse ($\beta < 0$); or for optimistic deviations towards best-case (ext = max) outcomes if the decision-maker is ambiguity-seeking ($\beta > 0$). The deviation from the Bayesian posterior q is measured by the “information distance” $D_{\text{KL}}(\tilde{q}||q)$ and scaled by $1/\beta$. The larger the magnitude of β , the higher the ambiguity regarding q . In Fig 3.1A it can be seen that $V_\beta(x) \leq V_0(x)$ for $\beta \leq 0$. In the economic literature the free energy valuation of Equation (3.3) is known as multiplier preference models (Hansen and Sargent, 2008) that are part of the more general family of variational preference models (Maccheroni et al., 2006). According to (Braun et al., 2011b; Ortega and Braun, 2011; Ortega and Braun, 2013), the decision-maker also optimizes a free energy to determine its action by following the strategy

$$\begin{aligned} \mathbf{H}_2 : \quad p_2(x) &= \underset{\tilde{p}(x)}{\text{argmax}} \left\{ \mathbb{E}_{\tilde{p}} [V_\beta(x)] - \frac{1}{\alpha} D_{\text{KL}}(\tilde{p}(x) || p_0(x)) \right\} \\ &= \frac{p_0(x) e^{\alpha V_\beta(x)}}{\sum_{x'} p_0(x') e^{\alpha V_\beta(x')}} \end{aligned} \quad (3.4)$$

which is equivalent to a soft-max choice rule when assuming an indifferent prior choice probability of $p_0(x) = \frac{1}{2}$. Such a free energy optimizing decision-maker can be interpreted as a bounded rational decision-maker that can only afford to deviate from the prior choice strategy $p_0(x)$ by a limited number of information bits quantified by the relative entropy $D_{\text{KL}}(p||p_0)$ (Ortega and Braun, 2013). Equation (3.4) describes the choice of option x with value $V_\beta(x)$ under both sensitivity to ambiguity and limited information-processing resources—see Fig 3.1B. Note that the two hypotheses are nested, as \mathbf{H}_2 includes \mathbf{H}_1 in the limit of $\beta \rightarrow 0$ (no sensitivity to ambiguity) and also includes the perfect Bayes-optimal decision-maker for $\alpha \rightarrow \infty$ and $\beta \rightarrow 0$.

To distinguish between the two hypotheses in our experiment we investigate subjects’ choice probabilities in *probe trials* in which a decision-maker that only cares about expected

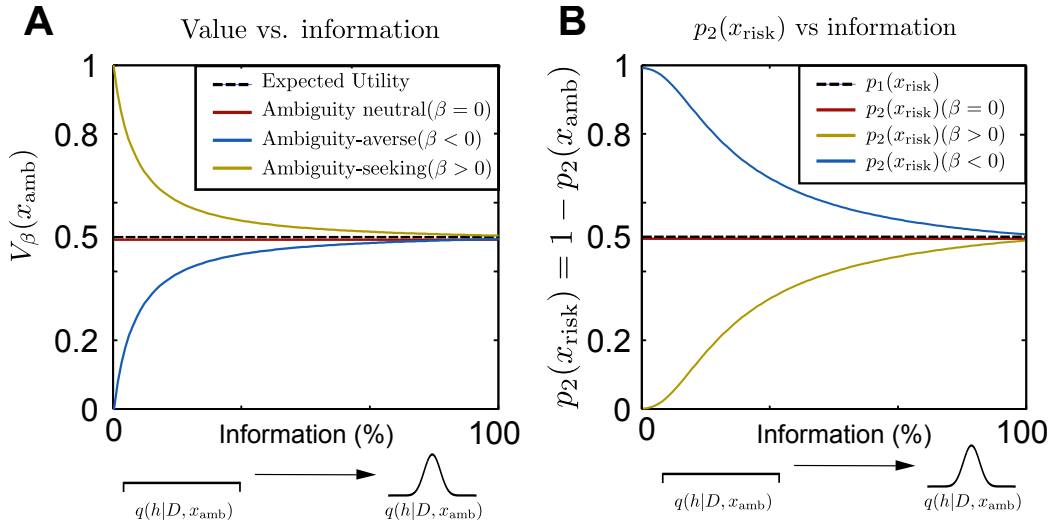


Figure 3.1. Illustration of model predictions. Predictions for probe trials where the risky urn (with equal composition of blue and red balls) has the same expected utility as the ambiguous urn—the number of observed red balls is equal to the number of observed blue balls. In panel **A** we show the value of Equation (3.3) assigned to an ambiguous option depending on the ambiguity attitude β and on the available information. In the case of the urn, information is quantified with the number of observations. The more information becomes available the more concentrated is the Bayesian posterior $q(h|D, x)$, so a high number of observed balls reflect a peaked posterior. We show in yellow line the value of the ambiguous option $V_\beta(x_{\text{amb}})$ according to Equation (3.3) for positive β (optimistic) being higher or equal than the expected utility value $\mathbb{E}_{q|x_{\text{risk}}} = 0.5$ (indicated by the dashed line). In blue we show that for negative β (pessimistic) the value V_β is always lower or equal than the expected utility value. The value V_β converges to the expected-utility value if the decision-maker is ambiguity-neutral (red line for $\beta \rightarrow 0$) or in the absence of ambiguity when the posterior becomes a delta function $q(h|D, x) = \delta(h - h^*)$. In panel **B** we show the predicted choice probability in probe trials according to Equation (3.4) for different β . Translating the value into a choice probability requires an additional parameter α that regulates the level of stochasticity like in a soft-max choice rule. For example, we show in yellow for a particular $\alpha > 0$ how the probability of choosing the risky option is modulated by the information available. The dashed line indicates the perfectly rational expected utility maximizer that is indifferent between the risky and the ambiguous option in the probe trials.

success would be indifferent between the risky and the ambiguous option. The expected utility hypothesis \mathbf{H}_1 predicts that subjects should be indifferent between the risky and ambiguous option in probe trials, that is $p_1(x) = 1/2$. In contrast, the free energy hypothesis \mathbf{H}_2 predicts that subjects should modulate their choice behavior in probe trials depending on the degree of ambiguity according to Equations (3.4) and (3.3). In particular, ambiguity-averse individuals ($\beta < 0$) should prefer the risky option in the face of ambiguity, but do so less and less the more information about the ambiguous option becomes available (that is the more concentrated their belief $q(h|D, x)$ becomes). Similarly, ambiguity-seeking individuals ($\beta > 0$) should prefer the ambiguous option, but less and less so with increasing information. These predictions are illustrated in Fig 3.1. Note that while the model explains choice behavior depending on a given ambiguity attitude β , it does not explain how β changes across task contexts. Details of the model can be found in the Materials and Methods section.

3.1.2 Experiments

We designed an experiment to test for differences in choice behavior in motor versus non-motor contexts. Furthermore, we designed a second experiment to control for framing effects that could be induced by the different stimulus designs used in the two tasks. In both experiments subjects had to choose between a risky and an ambiguous option in every trial. The risky option provided full information about the probabilities of the possible outcomes. The ambiguous option was always characterized by a lack of information with respect to the probabilities. We could manipulate the degree of ambiguity by varying the amount of information revealed about the ambiguous option. After the decision was made subjects received a payoff depending on the chosen option.

In Experiment 1 we compare two tasks, an urn task and a sensorimotor task under ambiguity—see Fig 3.2 top and middle row. In the case of the urn task the stimuli are sampled balls from both urns and the uncertainty about the outcomes after making the choice is computer generated. In case of the motor task, the stimuli are bars that subjects had to hit and therefore the uncertainty about the outcome is internally generated by subjects due to their skill and motor variability. Any difference in behavior between the two tasks might be attributable to either motor vs. non-motor context or to the stimulus design. The goal of Experiment 2 is to distinguish between the two possibilities.

3.1.3 Experiment 1: Urn Task vs Motor Task

3.1.3.1 Urn task

In the urn task—see Fig 3.2 top row—the risky option was always fully visible and displayed by a sample of 100 balls drawn from an urn with 50 : 50 composition of red and blue whereas the ambiguous option had a possibly different composition with varying degree of ambiguity depending on the number of samples that were shown, ranging from zero (full ambiguity) to one hundred samples (no ambiguity). Ellsberg’s original task corresponds to the fully

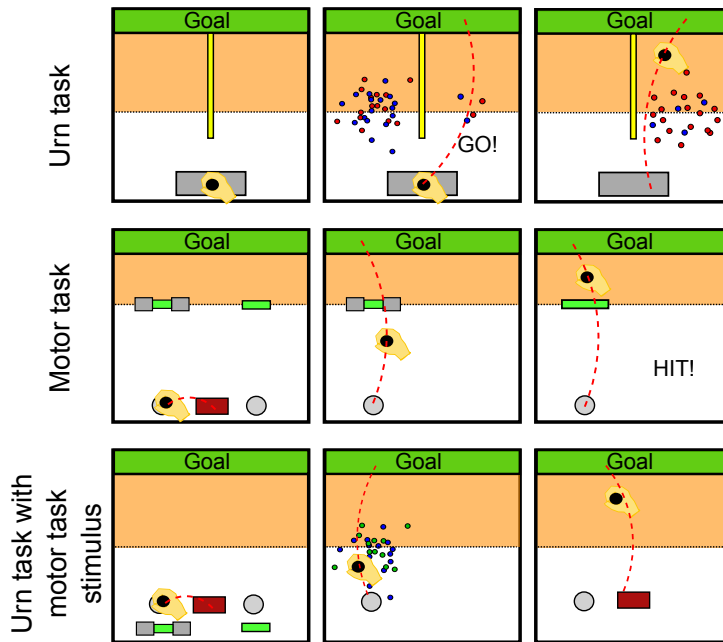


Figure 3.2. Experimental design. **Top row.** In the urn experiment, the trial was initiated by moving on a gray start bar. Two point clouds appeared showing samples from two urns with different underlying ratios of blue and red balls. The risky urn was always displayed by 100 samples drawn from a 50 : 50 ratio. The ambiguous urn had a variable number of samples drawn from an unknown ratio. Subjects made their decision about the urn they believed to have a higher red ratio by crossing into the area highlighted in orange. In case of choosing the ambiguous urn, the composition of the urn was revealed. The payoff was given by a viscous force field, which was switched on with the probability determined by the chosen urn’s ratio for blue. **Middle row.** In the motor task, subjects had to decide to hit one of two targets by moving into the corresponding decision circle. In case they chose a (partially) occluded target, the target became fully visible after crossing into the highlighted area. When failing to hit the target, subjects had to move against a viscous force field. In both experiments subjects had to move towards the goal bar and back. The orange color is only for illustration and was not displayed during the experiment. **Bottom row.** Experiment 2. Subjects are presented with the same stimulus as in the motor task but perform an urn task where the random outcome is computer generated, in contrast to the motor task where the outcome is determined by the subjects behavior. After choosing between an ambiguous and a risky option a cloud of points appeared revealing the composition of the hidden urn that determined the payoff in the same way as in the urn task.

ambiguous limit case in which the ambiguous urn shows zero samples.

Subjects decided between the risky and the ambiguous urn displayed in the two halves of the workspace respectively by moving a manipulandum to the respective half—compare Fig 3.2 top row. To complete the trial they had to move to a goal bar and back to the start position. Instead of a monetary payoff as used in Ellsberg’s original experiment, we used viscous force fields that subjects tried to avoid. The force payoff was stochastic and constituted a risk probability. The probability to experience a force in any individual trial was determined by the probability of drawing a blue ball from the urn chosen by the decision-maker. We recorded subjects’ choice in each trial and determined their choice probabilities as choice frequencies over many trials with the same stimulus. Importantly, we designed symmetric *probe trials* in which half of the shown samples from the ambiguous urn were red and the other half blue, such that the most likely hypothesis to explain this observation is a 50 : 50 composition of red and blue balls. Crucially, subjects should be indifferent between the ambiguous and the risky urn in these trials, if they base their decision solely on their expected success, as posited by expected utility theory given that in our experiment all possible ratios for the ambiguous urn are equiprobable. Additionally, we ascertained subjects’ preference for no-force outcomes in trials without ambiguity (i.e. the ambiguous option was fully revealed), where subjects preferred in more than 93% of cases the urn with the higher ratio of red. In fact, we found the payoff given as a force not to be critical, as subjects in a control experiment that received point scores as payoffs showed the same behavior—compare Figure 3.5 in the Supplementary Information. This is in line with previous studies (Inukai and Takahashi, 2009; Smith et al., 2002) that have found ambiguity attitude to be robust in gains vs. losses scenarios.

In accordance with Ellsberg’s results, we found in our urn experiment that the majority of subjects were averse to the fully ambiguous urn in the probe trials—see Fig 3.3A. For 13 out of 16 subjects the choice probability for the risky urn was significantly elevated from 50 : 50 in the fully ambiguous condition ($p < 0.05$, binomial test)—compare Figure 3.6 in the Supplementary Information for single subject choice data. This deviation from expected utility theory in the zero information limit (full ambiguity) was also significant at the population level ($p < 0.05$, Wilcoxon signed-rank test). In the case of full information (zero ambiguity), the ambiguous urn showed as many samples as the risky urn. Naturally, subjects were indifferent between these two indistinguishable options ($p > 0.6$, Wilcoxon signed-rank test). In between the two information limits, the ambiguous urn was partially revealed by showing a smaller number of samples. We found that subjects’ preference for the risky urn decreased with an increasing amount of information about the ambiguous urn ($p < 0.05$ for 11 out of 16 subjects, Cochran-Armitage trend test with linear weights). Moreover, we found that the time to take the decision increased with an increasing amount of information about the ambiguous urn—compare Figure 3.7 in the Supplementary Information. Since in the analyzed probe trials we ensured that half of the observed samples were red and the other half blue, a decision-maker that only cares about the expected success would be indifferent between the two options regardless of the amount of information. Such a decision-maker is represented by the dashed

flat line in Fig 3.3A. We found that all but one subject significantly differ from this choice pattern and thereby exhibit ambiguity aversion (13 subjects) or ambiguity-seeking behavior (2 subjects) depending on the degree of available information.

3.1.3.2 Motor task

In the sensorimotor task—see Fig 3.2 middle row—we translated Ellsberg’s urn task into an equivalent a motor task where subjects had to hit a risky target or an ambiguous target. The risky target was always fully visible, whereas the ambiguous target was occluded in varying degrees, such that subjects could not precisely assess the hitting probability associated with the hidden target size. We manipulated the degree of ambiguity by varying the size of the occluder from no occlusion to full occlusion.

In the motor task subjects chose in every trial between a risky target and an ambiguous target. Once selected, they had to try and hit the target. If they failed to do so, they experienced a viscous force on their way to the goal bar and back. To test the impact of varying ambiguity on choice behavior, we again introduced symmetric *probe trials* in which a decision-maker that only cares about expected success would be indifferent between the risky and the ambiguous option. In the case of the ambiguous target, the size of the occluder and the hidden target size were adjusted in a way such that the expected hitting probability for the subjects in probe trials was also 50%, given that in our experiment all hidden sizes compatible with the occlusion were equiprobable—see Materials and Methods for details. To assess subjects’ hitting probabilities and to adjust the displayed target sizes accordingly, we measured subjects’ endpoint variability and ensured that their performance was stable over at least 500 trials—see Materials and Methods for details. Finally, we ascertained subjects’ preference for no-force outcomes in trials without ambiguity, where subjects preferred the larger target in more than 87% of cases. Again we found the fact that the payoff was given as a force not to be critical, as subjects in a control experiment that received point scores as payoffs showed the same behavior—compare Figure 3.5 in the Supplementary Information.

In contrast to the expected utility prediction in probe trials that is represented by the dashed lines in Fig 3.3B, we found that most subjects’ choice probability differed significantly from this prediction, and that consequently their behavior cannot be simply explained by expected utility maximization. However, unlike in the urn probe trials, most subjects had a preference for the ambiguous option in the motor probe trials. When choosing between the risky and the fully ambiguous target—corresponding to Ellsberg’s choice scenario—, 13 out of 16 subjects’ choice probability for the risky target was significantly reduced from 50 : 50 ($p < 0.01$, binomial test)—compare Figure 3.6 in the Supplementary Information for single subject choice data. This deviation from expected utility theory in the zero information limit was also significant at the population level ($p < 0.01$, Wilcoxon signed-rank test)—compare Fig 3.3B. In the case of full information (zero ambiguity), both targets were fully visible and indistinguishable ($p > 0.6$, Wilcoxon signed-rank test). In between the two extremes of zero and full information, the ambiguous target was only partially occluded. We found that

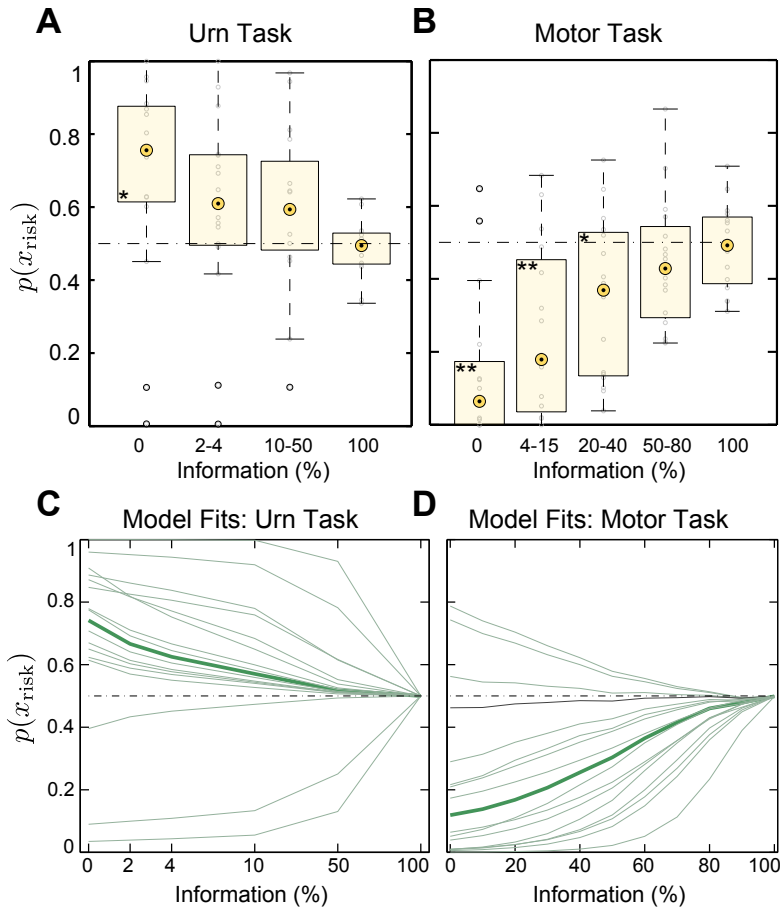


Figure 3.3. Experiment 1: Experimental data and model fits. Aggregate choice probabilities over all subjects in probe trials of **A** the urn task , **B** the motor task. The boxes are centered around the median across subjects and the edges of the box are the 25th and 75th percentiles. Panels **C** and **D** show the corresponding model fits. The thin green lines represent individual subjects' choice probabilities according to Equation (3.4), the thick green line indicates the group mean. The dashed lines show the indifference choice probabilities predicted by expected utility. Probabilities above the dashed line imply that subjects prefer the risky choice (ambiguity aversion), probability values below the dashed line imply that subjects prefer the ambiguous choice (ambiguity preference). Asterisks denote significant deviation from the expected utility prediction: one asterisk signifies $p < 0.05$, two asterisks signify $p < 0.01$. In the urn task information (%) corresponds to the ratio of the number of revealed balls to the total number of balls, in the motor task and in the urn task with motor stimulus to the ratio of visible size to total size of the ambiguous target.

for 14 out of 16 subjects, preference for the ambiguous target decreased with an increasing amount of information ($p < 0.05$, Cochran-Armitage trend test with linear weights). Unlike in the urn task, we found the decision time in the motor task not to vary with the degree of ambiguity—compare Figure 3.7 in the Supplementary Information.

3.1.3.3 Model fits

In the information-theoretic free energy model of decision-making there are two free parameters per subject to fit, that are the soft-max parameter α and the ambiguity parameter β . In contrast, the expected utility model has only one free parameter per subject given by the softmax-parameter α . The two decision-making models are nested in the sense that the expected utility model is a special case of the free energy model in the limit $\beta \rightarrow 0$. To compare the two hypotheses we maximized the log-likelihood of the experimental data over all trials by varying the free parameters of the two models. We performed a likelihood ratio test to investigate which model fits the data better. Importantly, the likelihood ratio test with nested models trades off the extra complexity of the more general model against its better fitting performance. We found that we can reject the expected utility model with a p-value of $p < 0.01$. The model fits are shown in green in Figures 3.3C,D and in Figure 3.6 in the Supplementary Information for individual choice data. In Fig 3.3 it can be seen that unlike the expected utility model, the free energy model can explain how subjects' choice probabilities change depending on the amount of available information.

While there are a number of alternative ambiguity models, the difficulty in our task is that these models have to be dynamically consistent—that is they have to be updated with new data in a consistent way—and they have to allow for both ambiguity-seeking and ambiguity-averse behavior. For the urn task we adapt one of the most popular ambiguity models from Gilboa and Schmeidler (Gilboa and Schmeidler, 1989), because in this case the prior can be easily parameterized as a beta distribution. The Gilboa-Schmeidler model assumes that decision-makers have multiple beliefs arising from multiple priors. We assume that within that set of priors decision-makers can update their beliefs according to Bayesian inference procedures and select the worst-case possible belief for every option. This will lead to ambiguity averse behavior. To allow for ambiguity-seeking behavior we will also allow for best-case selection of beliefs. We model directly the best- or worst-case belief selection with a single Beta prior for the ratio of the urn. In this case the Gilboa-Schmeidler model has three parameters, the soft-max parameter and two more parameters of the Beta prior. In contrast the information theoretic model has two parameters, the ambiguity parameter β and the rationality parameter α . We compare these two models using the Bayesian Information Criterion (BIC) and find that the model comparison clearly favors the information-theoretic model (BIC = 8139) over the dynamic Gilboa-Schmeidler model (BIC = 8165), with $\Delta\text{BIC} = 26$.

3.1.3.4 Comparison: motor task and urn task

Importantly, both experiments were performed by the same subjects. In total, 11 subjects that were ambiguity averse in the urn task under full ambiguity, preferred the fully ambiguous option in the motor task. Moreover, 2 subjects that were ambiguity averse in the urn task under full ambiguity, were indifferent to full ambiguity in the motor task. Three subjects did not change their preferences across tasks, two of them consistently preferred the fully ambiguous option, one of them remained indifferent. This difference in behavior of subjects between the two tasks might be attributable to either motor vs. non-motor context or to the stimulus design. This is the subject of the second experiment.

3.1.4 Experiment 2: Stimulus versus Motor Framing

In the first experiment we found a clear difference in choice behavior between the motor task and the urn task—compare Fig 3.3. This difference in subjects' behaviour between the two tasks might be attributable to either motor vs. non-motor context or to the stimulus design. In Experiment 2, we distinguish between the two possibilities. In this experiment, a group of subjects performed the urn task but at the moment of choice they were presented with the motor task stimulus instead of the urn task stimulus. If the preference reversal was mainly induced by the stimulus, we would expect most subjects to prefer the ambiguous option in the probe trials of Experiment 2, as the stimulus is identical to the motor task. However, if the preference reversal was mainly a function of the underlying source of uncertainty (Beers et al., 2004) (external source for the urn task or internal source for the motor task), we would expect them to be mostly ambiguity averse as in the urn task. We found that most subjects in Experiment 2 still preferred the ambiguous option as in the motor task—compare Fig 3.4A. This deviation from expected utility theory was significant both at the population level ($p < 0.05$, Wilcoxon signed-rank test on the full ambiguity condition) and at the level of individual choice: for 14 out of 16 subjects in Experiment 2, preference for the ambiguous target decreased with an increasing amount of information as in the motor task ($p < 0.05$, Cochran-Armitage trend test with linear weights). This suggests that subjects' ambiguity preference critically depends on the stimulus display, whereas the context of motor and non-motor task and the framing of gains and losses do not seem to be critical.

We conducted a control experiment to discern if the stimulus affects directly ambiguity attitude or whether it induces a perceptual distortion in such a way that after all subjects' behavior can be explained according to expected utility with perceptual bias. Importantly, both the control experiment and Experiment 2 were performed by the same subjects. The control experiment was identical to Experiment 2 except that the force payoff was now associated with the opposite color (inverse utility condition). Effectively, this implied that now smaller bar stimuli were preferable to larger bar stimuli. We can then compare subjects' choices when presented with the same target bar stimulus under the two utility conditions. If they prefer the same option—either ambiguous or risky—under both conditions, their choice cannot be explained by a single belief probability, as they would effectively believe the same

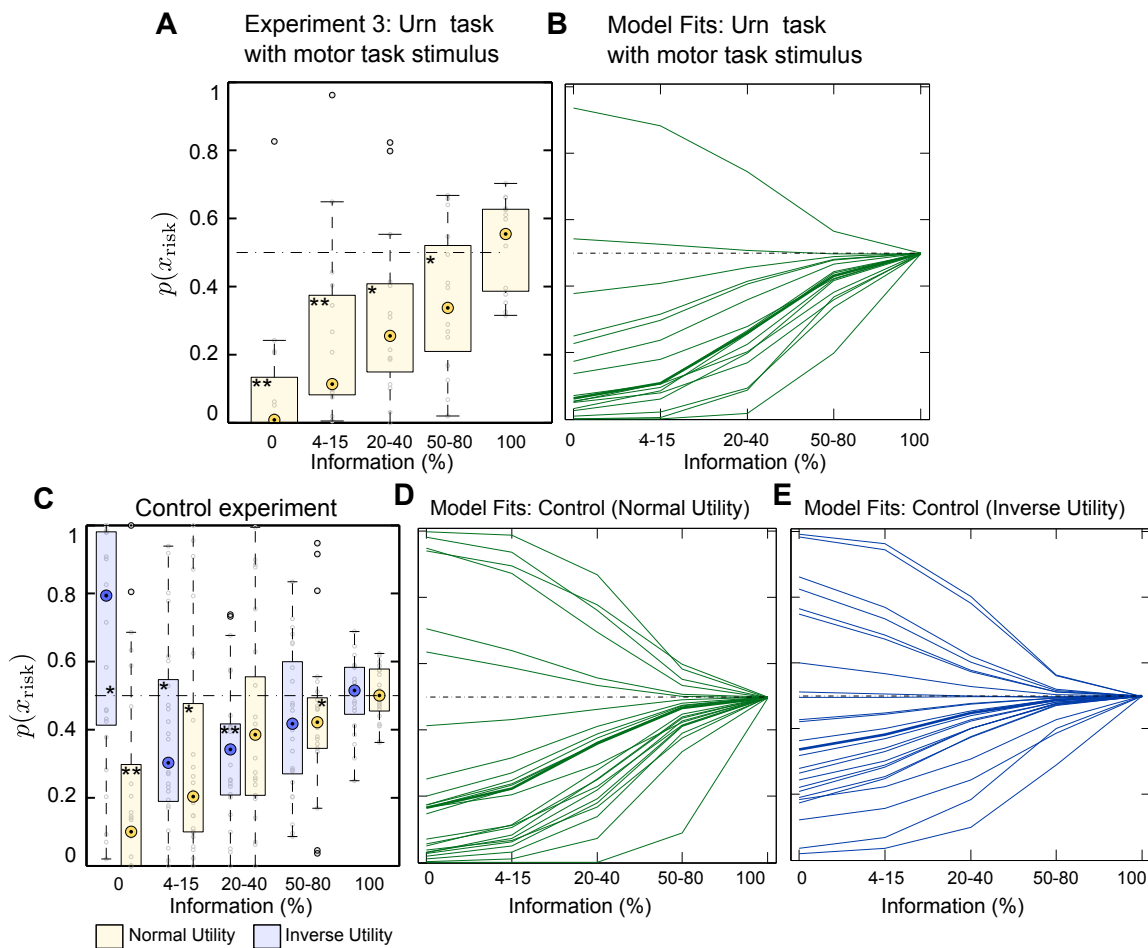


Figure 3.4. Experiment 2. Panel A shows Experiment 3 choice data where subjects are ambiguity-seeking as in the motor task and not ambiguity-averse as in the urn task—compare Fig 3.3. Panel C shows experimental choice data of the control experiment where subjects are ambiguity-seeking in most trials independent of the utility function—normal utility condition as in Experiment 3 or inverse utility condition. Probability values above the dashed line predicted by expected utility imply that subjects prefer the risky choice (ambiguity aversion), probability values below the dashed line imply that subjects prefer the ambiguous choice (ambiguity preference). The normal utility condition is colored in light orange, the inverse utility condition is colored in light blue. The boxes are centered around the median across subjects and the edges of the box are the 25th and 75th percentiles. Asterisks denote significant deviation from the expected utility prediction: one asterisk signifies $p < 0.05$, two asterisks signify $p < 0.01$. Panels B, D and E show the corresponding model fits. The thin lines represent individual subjects’ choice probabilities according to Equation (3.4), the thick lines indicate the group mean. Note in panel E, how the our information-theoretic model (and the expected utility model) is unable to produce simultaneously ambiguity aversion in the zero information limit and ambiguity seeking behavior in the other information cases.

stimulus to be larger and smaller at the same time. Crucially, this could be explained in terms of ambiguity attitude. If, however, they believed the ambiguous target to be smaller in one utility condition, but larger in the other, then their behavior might also be explicable as a perceptual bias or a biased belief that discards experienced statistics.

Fig 3.4C shows subjects' choice behavior under the two utility conditions. For the normal utility condition (light orange boxes), we see the gradual increase in ambiguity preference as in the previous subject groups. However, for the inverse utility condition (blue boxes) the results are mixed. For partially occluded stimuli subjects still mostly prefer the ambiguous option, but in the case of fully occluded stimuli they mostly prefer the risky option—compare Fig 3.4C. The corresponding significance values (Wilcoxon signed-rank test) are indicated by asterisks in the figure. This mixed behavior is also visible in single subject choice data—compare Figure 3.8 in the Supplementary Information. For both, the Experiment 2 with normal utility condition and the inverse utility condition we found the decision time not to vary with the degree of ambiguity—compare Figure 3.7 in the Supplementary Information. In Figs 3.4B,D,E we show the model fits according to equation (3.4). In particular, we observe the limitations of the information-theoretic model and the expected utility model when modeling the inverse utility condition experiment (Fig 3.4E). None of them is able to show both ambiguity averse behavior in the zero-information limit and ambiguity seeking behavior in the remaining ambiguity levels.

A possible reason for the deviation observed in the full ambiguity condition might be the nonlinear relationship between bar size and hitting probability, which only plays a minor role in the partial ambiguity condition. Thus, these mixed results suggest that both perceptual distortion and stimulus-dependent ambiguity attitude play a role in sensorimotor choices. Crucially, it is impossible to exclusively explain the observed preference reversal within expected utility theory with biased beliefs or by perceptual distortion, as it is impossible to probabilistically represent the belief or the perception that the same stimulus is at the same time smaller and bigger than a gauge stimulus, as observed in the partial ambiguity conditions.

3.2 Discussion

In our study we found that human subjects continuously modulate their choice behavior in an experience-based urn task and in a motor task depending on the level of ambiguity in line with the prediction of an information-theoretic free energy choice model and contrary to the prediction of expected utility. We found that the ambiguity preference changed in the two tasks for the same subjects, where subjects were mostly ambiguity-averse in the urn task and ambiguity-seeking in the motor task. Additionally, we found that subjects' ambiguity sensitivity is not affected by the framing of motor and non-motor context. However, in a second experiment we show that this reversal was mainly a consequence of a framing effect induced by the different stimuli in the two tasks.

Framing effects induced by presenting decision-problems in terms of gains and losses were first studied by Tversky and Kahneman ([Amos Tversky and Daniel Kahneman, 1981](#)), showing

how framing could greatly affect choice behavior. In our study we found that the framing effect induced by stimulus display significantly affected behavior. In fact, the visualization of uncertainty has recently become an active research topic (Brodie et al., 2012; Johnson, 2004; Marx, 2013; Pang et al., 1996). One of the reasons for this surge in interest is the realization that the way uncertainty is communicated can affect policy making, for example in the context of the climate change debate. Similarly, our results suggest that the way that ambiguity is presented to users can make striking differences in the way they respond to this uncertainty, both in economic decision-making tasks and in motor tasks. A previous study by Wu et al (Wu et al., 2009) have also reported striking differences between economic and sensorimotor decision-making under risk. Our results suggest that these differences could be explained by the way uncertainty is displayed and not by the fact of how uncertainty is generated—externally in case of economic task or internally in case of a sensorimotor task. In principle, the representation of uncertainty can induce both perceptual biases or elicit particular ambiguity attitudes. In our tasks, we found that perceptual biases alone cannot explain subjects’ choice behavior and that ambiguity attitude is affected in stimuli with partial ambiguity.

Previous studies in behavioral economics have shown that risk-attitudes can be distinguished experimentally from ambiguity attitudes (Camerer and Weber, 1992). Risk attitudes are usually modeled by the curvature of the utility function (Kenneth J. Arrow, 1965; John W. Pratt, 1964). This model of risk is also included in Equation (3.3). The ambiguity attitude in the free energy model is expressed by an additional temperature parameter that quantifies deviations from a Bayesian model (Ortega and Braun, 2013). The same variational principle can also be applied to acting of bounded rational decision-makers. In this case, the temperature parameter β can be interpreted in terms of the degree of control a decision-maker has as a result of the available computational resources. Accordingly, one could interpret Equation (3.3) equivalently as anticipating the choice of a bounded rational opponent with boundedness parameter β . Therefore, our results encourage a more general investigation of free energy variational principles for perception and action. One such avenue might be the study of decision-makers’ perceived degree of control, for example in the context of illusions of control (Gino et al., 2011; Langer, 1975). In the case where utilities are restricted to informational surprise or absorbed into prior distributions, such free energy variational principles have for example been recently investigated by Friston and colleagues (Friston, 2009; Friston, 2010).

In the economic literature there have been an extensive effort in developing models that formalize decision-making under ambiguity. From the first models where decisions are evaluated by looking exclusively at its worst possible outcome (Wald, 1945), to models that take into account both the worst and the best possible outcome (Kenneth J Arrow and Hurwicz, 1972). There are also more mathematically elaborate models such as Choquet Expected Utility (CEU) model (Schmeidler, 1989) where beliefs are not considered subjective probabilities but by capacities that can possibly be non-additive. Extensions of CEU include the Cumulative Prospect Theory (Amos Tversky and Daniel Kahneman, 1992) that uses two capacities,

one for gains and another one for losses. There are other popular models such as the Maxmin expected utility model from Gilboa and Schmeidler that use multiple priors to define the beliefs of decision-makers with built-in ambiguity aversion (Gilboa and Schmeidler, 1989) and also a variation of it that drops the axiom of ambiguity aversion (Ghirardato et al., 2004). The smooth ambiguity aversion model (Klibanoff et al., 2005) can be viewed as an extension of the maxmin model. It regards the maxmin criterion as too extreme and opts for modeling second order beliefs and introducing a convex function to model ambiguity aversion—in the same way that the curvature of the utility function models risk aversion.

The information-theoretic model relates to the above-mentioned models in several ways. First, Equation (3.3) that assigns value to an option under ambiguity presented here is known in the economic literature as a multiplier preference model (Hansen and Sargent, 2008), that is a type of variational preference model for decision-making under ambiguity (Etner et al., 2012; Maccheroni et al., 2006). In our formalism, the temperature parameter can assume positive and negative values corresponding to ambiguity-seeking or ambiguity-averse behavior without changing the general form of the solution equations. Second, just like the multiplier preference model the information-theoretic model has dynamic consistency, because it can incorporate new information in line with Bayesian updating (Hanany and Klibanoff, n.d.). Accordingly, it provides a neat way to include ambiguity into the Bayesian formalism, unlike many other models that abandon the concept of Bayesian probability. Third, the majority of the decision-making models under ambiguity include an *ad hoc* soft-max function to determine the probabilities of decisions. In contrast to these previous models, we use a single free energy principle for both action (Equation (3.4)) and perception (Equation (3.3)) which can also be reconciled with Bayesian updating (that also obeys a free energy principle) and dynamic choice under new incoming data (Gilboa and Marinacci, 2011). Therefore, the information-theoretic model provides a powerful generalization to Bayes optimal decision-making allowing for ambiguity and limited resources.

Variational ambiguity models build on earlier work on robust control where decision-makers consider the possibility that their current model q may not be the appropriate model for the observed phenomenon and therefore bias their predictions towards worst-case outcomes to ensure robustness (Hansen and Sargent, 2008). The concept of robustness is also closely related to the concept of risk-sensitivity as the relative entropy contains the information of all the higher-order statistical moments. Previously, risk-sensitivity was shown to play an important role in motor tasks, showing that subjects care about higher order moments of the cost function. This is often modeled as a mean-variance trade-off, that can be used to express risk attitudes towards observable random variables (Braun et al., 2011a; Nagengast et al., 2010; Nagengast et al., 2011b). Previously, it was also shown that risk-sensitivity affects sensorimotor integration when different beliefs are associated with different sensorimotor costs (Grau-Moya et al., 2012) and it also affects the amount of cooperation in two-player games when different beliefs represent the strategy of the other player (Grau-Moya et al., 2013). In this study we show that the same framework that is used to model risk-sensitivity can also be applied to model ambiguity.

Bayes optimal decision-making has been applied as a very general optimality principle to explain behavior from the scale of single neurons (Wei J. Ma et al., 2006; Wei Ji Ma et al., 2008) to whole-body motor control (Stevenson et al., 2009). When probability models are accurate, optimal decision-making is indeed optimal and accomplished by computing expected utilities. However, if probability models are inaccurate or even plain wrong, then maximizing expected utility can be far from optimal. In such scenarios, one might be interested in robust control and decision-making strategies with guaranteed performance bounds within defined neighborhoods of a proposed model (Hansen and Sargent, 2008). Robustness is also a core feature of biological organisms coping with model uncertainty (Kitano, 2004), which has so far been neglected in many optimality models. Our results suggests a way of how to combine model uncertainty, optimality and inference in the study of adaptive behavior.

3.3 Materials and Methods

3.3.1 Ethics Statement

The study was approved by the ethics committee of the Max Planck Society (reference number: 0269/2010BO2). All participants gave written informed consent.

3.3.2 Subjects

68 subjects (30 male, 38 female) from the Tübingen University student population participated in this experiment after giving informed consent. We excluded one subject in the motor task with force payoffs and two subjects in the motor task with point payoffs, because the standard deviation did not stabilize over the course of the experiment. The remaining 65 subjects were assigned as follows to the three experiments: 16 subjects participated in Experiment 1, 16 subjects participated in Experiment 2, and in Experiment 2 there were 16 subjects in the main experiment (8 of which overlapped with 8 subjects from Experiment 1), and 25 subjects in the control. Participants were paid the local standard rate of 8 Euros per hour for their participation.

3.3.3 Materials

The experiments were conducted using a vBOT robotic manipulandum (Ian S. Howard et al., 2009). Participants controlled the vBOT handle in the horizontal plane. Movement position and velocity were recorded at a rate of $1kHz$. A planar virtual reality projection system was used to overlay images into the plane of movement of the vBOT handle. Subjects hand position was displayed by a cursor that could move across the planar screen. Subjects were using their preferred hand throughout the entire experiment.

3.3.4 Information-Theoretic Model Details

In our experiment decision-makers have ambiguity about a latent variable h , which is the unknown ratio of blue and red balls in case of the urn, or the size of the hidden target in case of the motor task. The expected utility for a known h is determined by $U(h) = \sum_o p(o|h)r(o)$, where $r(o) = -1$ is the reward for the outcome $o = \textit{blue}$ in the urn task or $o = \textit{fail}$ in the motor task, and $r(o) = 0$ for the outcomes $o = \textit{red}$ or $o = \textit{hit}$. $p(o|h)$ indicates the probability of drawing color o from an urn with known ratio h or the probability of hitting a fully visible target of known size h depending, of course, on subjects' skill level. Note that we took into account changes in subjects' performance during the whole experiment—see *Experimental design: motor task* in the Materials and Methods section

The decision-maker's model q is given by a Bayesian posterior $q(h|D, x)$ over the latent variable h when observing data D of option x . In the urn task, the data corresponds to the number of observed red and blue balls and the distribution $q(h|D, x_{\text{amb}})$ can be represented by a Beta distribution over the ratio of the ambiguous urn. In the motor task, the data corresponds to observing the occluded target, where $q(h|D, x_{\text{amb}})$ is a uniform distribution over the possible target sizes covered by the occluder as we sampled the target sizes from this uniform distribution.

The critical trials for model comparison are the probe trials, in which the expected utility of the ambiguous option is exactly the same as the expected utility of the risky option. Importantly, a pure expected utility decision-maker with $\beta = 0$ values the ambiguous option x_{amb} according to its expected utility $V_0(x_{\text{amb}}) = \int dh q_{\text{amb}}(h|D, x_{\text{amb}})U(h)$, which in the illustration in Fig 3.1 simply corresponds to the mean of the distribution $q(h|D, x_{\text{amb}})$. In probe trials the mean is given by $\mathbb{E}_{q(h|D, x)}[U(h)] = 1/2$ by design. Crucially, in probe trials the expected utility value is independent of the number of observed data points. In contrast, the valuation given by Equation (3.3) is sensitive to the number of data points that determine the spread of the distribution $q(h|D, x)$. The more data becomes available the more concentrated the posterior becomes around the true value h^* (that is the true ratio of the ambiguous urn or the true target size of the ambiguous target). In the limit of exact knowledge only the true value h^* has non-zero probability mass, that is $q(h|x, D) \rightarrow \delta(h - h^*)$. In the limit of infinite data, all ambiguity vanishes and the value of the ambiguous option according to Equation (3.3) converges to $V_\beta(x_{\text{amb}}) \rightarrow U(h^*)$ independent of the value of β . The limit value $U(h^*)$ exactly corresponds to the value $V_0(x_{\text{risk}})$ of the risky option in probe trials—compare Fig 3.1.

The solution to Equation (3.4) that describes the choice probability of subjects choosing the risky option is given by

$$p_2(x_{\text{risk}}) = \frac{p_0(x_{\text{risk}})e^{\alpha V_\beta(x_{\text{risk}})}}{p_0(x_{\text{risk}})e^{\alpha V_\beta(x_{\text{risk}})} + p_0(x_{\text{amb}})e^{\alpha V_\beta(x_{\text{amb}})}} \quad (3.5)$$

where $p_0(x_{\text{risk}}) = p_0(x_{\text{amb}}) = 1/2$. Naturally, the probability of choosing the ambiguous option is modeled by $p(x_{\text{amb}}) = 1 - p(x_{\text{risk}})$. Note that the case of uniform prior the choice probabilities follow the common soft-max rule but for non-uniform prior it is a weighted

version of this rule.

The value $V_\beta(x)$ is the solution to Equation (3.3) and is given by

$$V_\beta(x) = \frac{1}{\beta} \log Z_\beta(x) = \frac{1}{\beta} \log \int q(h|D, x) e^{\beta U(h)} dh.$$

As the utility function $U(h)$ and the Bayesian posterior $q(h|D, x)$ are given by our modeling assumptions, there are only two free parameters per subject to fit in the information-theoretic free energy model of decision-making, that are the soft-max parameter α and the ambiguity parameter β .

3.3.5 Experimental Design: Experiment 1 (Urn Task)

Experiment. Subjects performed 600 trials of reaching movements from a start bar (gray rectangle with size $4 \times 1.5\text{cm}$) to a goal bar (green rectangle with size $20 \times 0.5\text{cm}$) that was 24cm away by moving a cursor (red circle, radius 0.3cm) representing their hand position—compare Fig 3.2. After holding still for 0.2s at the start bar, subjects heard a beep indicating trial start and stimulus appearance. By moving to the left or right side of the workspace when entering the force zone at 12cm in the forward direction (orange zone in Fig 3.2), they made a choice between the “risky” urn and the “ambiguous” urn. The display location of the risky and the ambiguous urn was randomly selected between left and right. The choice had to be made within a maximum time window of 1s after stimulus appearance, otherwise a new trial was generated.

Subjects were informed that both urns contained 100 balls. Subjects were also told that the risky urn always had 50 blue balls and 50 red balls and that the ambiguous urn had an unknown proportion of red and blue balls. Before they had to make their decision they were shown a sample of 100 balls drawn with replacement from the risky urn and a sample of varying size from the ambiguous urn. The number of samples shown from the ambiguous urn was determined randomly from the set $\{0, 2, 4, 10, 50, 100\}$. Thus, showing 0 balls corresponds to a completely ambiguous urn and showing 100 balls corresponds to a non-ambiguous urn. We devised two methods to indicate the missing information to ensure robustness of our results. The first eight subjects were explicitly told that the ambiguous urn had 100 balls with only a small subset shown as a sample. The second eight subjects were shown gray balls in ambiguous trials to visualize the missing information directly. Samples from the ambiguous urn were generated as follows. In 50% of trials the composition of the ambiguous urn was determined randomly from the set $\{(0, 100), (10, 90), \dots(50, 50), \dots(90, 10), (100, 0)\}$ of red balls and blue balls respectively. The specified amount of samples was then drawn from the ambiguous urn. In the other 50% of trials, we designed *probe trials* where a perfectly symmetric stimulus was presented where half of the samples was red and the other half was blue. These *probe trials* are important for the model comparison.

To show subjects the samples stemming from either urn, circles of red and blue colors (radius 2mm) were drawn from a two-dimensional Gaussian distribution with mean $\mu_{left} = -5.0\text{cm}$ $\mu_{right} = 5.0\text{cm}$ and standard deviation $\sigma_{left} = \sigma_{right} = 1.0\text{cm}$. The circles were

displayed in the horizontal plane as illustrated in Fig 3.2. In case subjects opted for the ambiguous urn, the content of the urn was revealed after their choice in order to provide them with feedback. In probe trials, the feedback was given by a sample of 100 balls drawn from a 50 : 50 urn. In other trials, the 100 balls sample was drawn from the ambiguous urn with the specified composition.

Subjects were told to imagine that a ball would be randomly drawn from the urn that they chose, and if the ball was blue they would experience a viscous force F in the forward-backward direction when trying to reach the goal bar. The force was set to $F = -kv_y$ with $k = 1.25 \frac{Kg}{s}$ and v_y the velocity of the robot handle in the forward-backward direction. In contrast, if the ball was red subjects would experience no force. The constant k of the viscous force was ramped up from $k = 0$ to $k = 1.25$ in the first third of the force area ($12cm - 16cm$) in the forward movement and similarly was ramped down in the backward movement to have a smooth transition between the non-force area and the force area.

Sampling procedure. In the urn task, the manipulation of the degree of ambiguity was controlled by the number of samples shown from the ambiguous urn. For finite size urns, the problem of inferring the true ratio of red and blue balls of the ambiguous urn depends in general on whether assuming a sampling scheme with replacement or without replacement. While subjects could in principle use either inference strategy, importantly this does not affect our conclusions. In the critical trials with symmetric evidence (probe trials), the expected utility of the ambiguous option is exactly the same under both sampling schemes and equal to the expected utility of the risky option, because in this case the mean of the beta distribution (inference with replacement) is equal to the mean of a beta-binomial distribution (inference without replacement). Our results are consequently independent of the sampling scheme that subjects were using for inference.

3.3.6 Experimental Design: Experiment 1 (Motor Task)

Experiment. Subjects had to move a cursor to a red start rectangle that was placed in the bottom middle of the workspace—compare Fig 3.2. When subjects entered the red rectangle, two decision circles (radius $0.6cm$) appeared to the left ($-4cm$) and to the right ($4cm$) of the center of the start rectangle together with their respective targets, and the red bar disappeared. One decision circle was associated with the risky target, while the other one was associated with the ambiguous target. The targets were displayed $18cm$ in the forward direction from the decision circles. The location of the risky target and the ambiguous target was randomized between left and right with 50 : 50 probability. Subjects could compare the two targets and move towards the decision circle associated with the target that they intended to hit. When holding still in the decision circle, the other decision circle and target disappeared and they heard a beep that urged them to move towards the target they selected. In order to increase the difficulty of the target hitting task, we imposed a lateral gain $g = 3$ between hand and cursor movement, thereby artificially increasing the variance of subjects' reaching

endpoints. When hitting the target, they heard a high frequency beep. When missing the target, they heard a low frequency beep. In the latter case, they also experienced a viscous force F impeding their movement in the forward-backward direction between the target and a goal bar (between 18cm and 27cm from the decision circle) they had to reach to complete the trial. The viscous force $F = -kv_y$ was proportional to subjects' movement velocity. To provide a smooth transition between the non-force area and the force area, k was ramped up in the forward direction from $k = 0$ to $k^* = 0.6\frac{Kg}{s}$ within the first quarter between the target and the goal bar, and similarly, ramped down in the backward direction. When subjects reached the goal bar, they heard another beep with the same frequency as before to inform them that the trial was completed. At this point they had to move back to the red rectangle to initiate the next trial. Each trial had to be completed within 0.6s .

In case of a fully visible target with half-width s , the probability of hitting the target P_{hit} can be computed as

$$P_{hit}(s) = 2 (F(s; \sigma_0^2) - F(0; \sigma_0^2)), \quad (3.6)$$

assuming that subjects' reaching endpoints can be described by a zero-mean Gaussian distribution with variance σ_0^2 such that $F(x; \sigma_0^2) = \int_{-\infty}^x \mathcal{N}(x; 0, \sigma_0^2) dx$. In case of an ambiguous target with visible size $2s$ and gray occluders of size d on each side, the average hitting probability is

$$P_{hit}(s, d) = \frac{2}{d} \int_s^{s+d} (F(x; \sigma_0^2) - F(0; \sigma_0^2)) dx, \quad (3.7)$$

assuming that all possible target sizes are equally probable.

Training and tracking performance. At the beginning of the experiment subjects were exposed to a training session where they had to hit a single fully visible target (width 2cm) displayed randomly at the left or right target position. After 200 trials, the training session ended allowing us to estimate subjects' hitting accuracy. In particular, we could compute the target half-width s^* , such that subjects' hitting probability was $P_{hit}(s^*) = 0.5$. To determine s^* we computed the median of subjects' unsigned endpoints. In order to keep track with potential changes in performance after the training session, we continuously adapted s^* and all other target sizes based on the penultimate 200 trials ensuring that subjects keep a constant performance over the entire course of the experiment. In total, subjects performed at least 750 choice trials under the condition that the relative standard deviation of s^* lies within a band of 10% over the last 500 trials.

Trial generation. Ambiguous trials were generated in the following way. Analogous to our urn experiment, in 50% of the choice trials the hitting probability of the ambiguous target was set to $P_{hit} = 0.5$. In the other 50% of trials the ambiguous target was set to have a hitting probability drawn randomly from $P_{hit} \in \{0.3, 0.4, 0.5, 0.6, 0.7\}$. Larger hitting probabilities were not considered because of the disproportionate target sizes required. Once P_{hit} was determined, the maximum ambiguous size d_f was computed according to Equation (3.7) for $s = 0$. Then an ambiguity index a was drawn randomly from the set $\{0, 0.1 \dots 0.5 \dots 0.9, 1\}$

and the actual size of the occluders was computed as $d = ad_f$. This way it could be ensured that all degrees of ambiguity were equally probable. Finally, given the expected hitting probability P_{hit} and the occluder size d , the displayed target half-width s was chosen to satisfy Equation (3.7).

3.3.7 Experimental Design: Experiment 2

In the second experiment subjects were shown the same stimulus as in the motor task, but then experienced the an externally imposed uncertainty as in the urn task. Instead of red and blue, the urn stimulus consisted of green and blue balls to match the color of the target and to establish an association between target size and ratio. The ratio of blue and green balls in the urn was determined by the hitting probability of the true target size under a variance of $1.44cm^2$. Subjects initiated each trial by moving their cursor to a red start rectangle as in the sensorimotor experiment. Then two decision circles (radius $0.6cm$) appeared to the left ($-4cm$) and to the right ($4cm$) of the center of the start rectangle and the two respective target stimuli—one risky, one ambiguous—were displayed $2cm$ below the decision circles. Once the target was selected, a cloud of points was shown $4cm$ above the decision circle to represent the urn. Once the goal bar was crossed the red start bar reappeared so that subjects could trigger the next trial.

Main experiment. In each trial of the Experiment 2, subjects chose between a risky target and an ambiguous target with the same statistics as in the motor task, including the occurrence of probe trials with equal expected utility for both options. Once subjects made their choice based on the target stimulus by moving the cursor into one of the decision circles, a cloud of points appeared as in the urn task representing a sample of blue and green balls (instead of red), as subjects were told that the size of the green target bar indicated the ratio of green balls in the urn. As in the urn task, this ratio also determined the probability of the force payoff.

We recorded eight subjects in this control experiment after they performed in the motor task, and another eight control subjects after they performed in the urn task to account for order effects. Unlike the first group, the second group of eight subjects were not yet acquainted with the bar stimulus once they started the control experiment. We therefore adapted the control experiment for them in such a way that they could learn the relationship between ambiguous target stimuli and true target size. Once they selected the target, the true target size was revealed at the same time as showing the true composition of the urn. In contrast, the first eight subjects already knew the bar stimulus from the preceding motor task. Once they selected the target in the control task, they were shown a point cloud that consisted of $100 \times a$ gray balls determined by the ambiguity index a associated with the ambiguous target and a sample of $100 \times (1 - a)$ green and blue balls drawn from the corresponding urn with composition equal to P_{hit} . The true composition of the urn was revealed when entering the force zone as in the urn task. This way they could learn a direct mapping from ambiguous bar

stimulus to ambiguous urn stimulus. The remainder of the trial proceeded for both control groups as in the urn task. There was no qualitative difference between both groups, as both predominantly preferred the ambiguous option. Accordingly, we found that the preference of subjects first performing in the motor task remained stable across tasks, but the preference of subjects first performing in the urn task changed—compare Figure 3.9 in the Supplementary Information.

Control experiment. In this control experiment we tested a group of subjects performing both in the normal and in an inverted utility condition using the same design as in Experiment 2. In the normal utility condition, a force payoff was associated with drawing a blue ball from the associated urn as in the previous experiment (Experiment 2). In the inverse utility condition, a force payoff was associated with drawing a green ball (instead of red) from the urn. Using the sensorimotor stimulus, effectively, subjects had to decide in the first condition which of the two target bars—risky or ambiguous—they believed to be larger, whereas in the second condition they had to decide which one they believed to be smaller.

In the inverse utility control experiment, sixteen subjects performed the inverse utility condition before performing in the normal utility condition, and nine subjects performed the inverse utility condition after performing in the normal condition. Subjects were told that in both conditions the size of the green bar indicated the proportion of green balls in the urn and that green balls would either be associated with no force (normal condition) or with a force (inverse condition) according to the probability of drawing a green ball from the urn. Since these subjects did not previously perform in the motor task, they underwent the same procedure as the second group of eight subjects in Experiment 2.

Order effects. In all experiments, the order in which subjects performed the experiments was permuted. From the sixteen subjects performing the urn and motor task, the first eight subjects started with the urn task, while the second eight subjects started with the motor task. Before performing Experiment 2, the first eight subjects of Experiment 2 performed the motor task and the second eight subjects performed the urn task. In the control experiment, the first sixteen subjects performed the inverse utility condition before the normal condition, and nine subjects performed the normal condition before the inverse condition. To test for order effects we devised both a logistic generalized linear mixed model that depended on an order variable and another logistic generalized linear mixed model that did not depend on it. In both models the other fixed effects were given by the ambiguity condition and the expected hitting probability. The random effect in both models was given by the subject index. We found that in none of the above cases the order played a significant role ($p > 0.05$ in all cases, χ^2 difference test).

3.4 Supporting Information

Data

The data can be found in doi:10.1371/journal.pone.0153179.s002.

Acknowledgments

This study was supported by the DFG, Emmy Noether grant BR4164/1-1 and by Grants from the U.S. National Science Foundation, Office of Naval Research and Department of Transportation.

Supplementary Figures

Experiment: Gain vs losses

In order to test if there exists a framing effect depending if the payoffs in the experiments are changed from losses to gains we conducted a second experiment. In this second experiment (reward vs. force payoff), sixteen subjects performed the urn and motor task experiment as described above. The only difference was the payoff mode. Subjects did not experience any viscous forces, but instead received point rewards. In urn task trials, a point was awarded whenever a red ball was drawn from the urn selected by the subject. In motor task trials, a point was awarded whenever the subject managed to hit the target. In all other cases no points were awarded. The total point score was displayed on the screen at all times. Even though we found stronger significance in the urn task and weaker significance in the motor task for specific ambiguity levels, overall we found that our results were not significantly affected when comparing the subject population receiving force payoff to the subject population receiving point payoffs ($p > 0.15$, Wilcoxon ranksum test for each ambiguity condition in the urn task and $p > 0.5$, Wilcoxon ranksum test for each ambiguity condition in the motor task). The aggregate choice probabilities and model fits are shown in Fig. S3.5. This suggests, in line with a previous study (Inukai and Takahashi, 2009), that ambiguity attitude is not sensitive to positive or negative payoff.

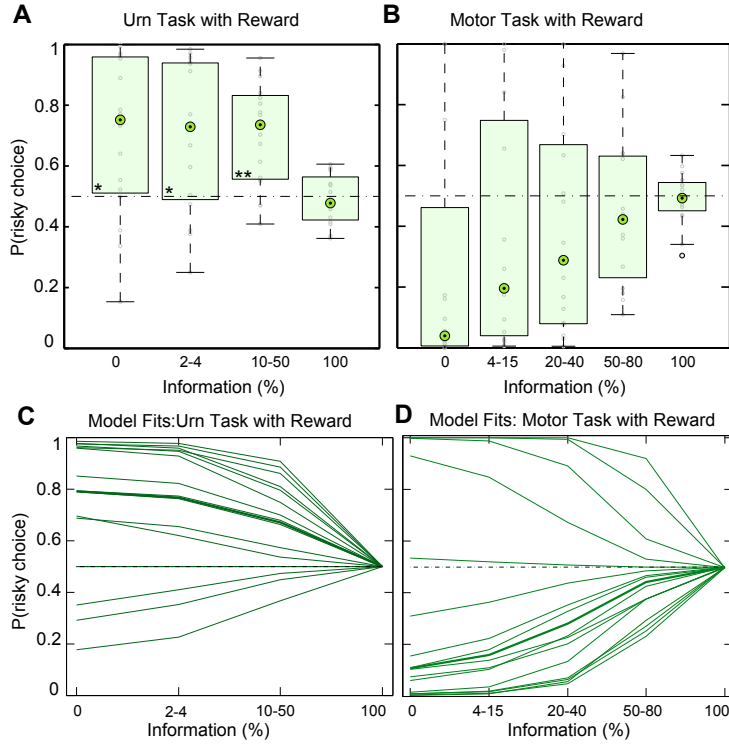


Figure 3.5. Experiment: Reward versus force payoff. Aggregate choice probabilities over all subjects in symmetric probe trials of the urn task **A** and the motor task **B**, when subjects received point rewards instead of experiencing viscous forces. In the urn task, subjects received one point whenever a red ball was drawn from the urn the subjects selected. No points were awarded if a blue ball was drawn. In the motor task, subjects received one point for hitting the target, otherwise no point was awarded. In both tasks the total point score was shown at all times on the screen. The boxes are centered around the median across subjects and the edges of the box are the 25th and 75th percentiles. Panels **C** and **D** show the corresponding model fits. The thin green lines represent individual subjects' choice probabilities according to Equation 3.4 in the main text, the thick green line indicates the group mean. The dashed lines show the indifference choice probabilities predicted by expected utility. Probabilities above the dashed line imply that subjects prefer the risky choice (ambiguity aversion), probability values below the dashed line imply that subjects prefer the ambiguous choice (ambiguity preference). Asterisks denote significant deviation from the expected utility prediction: one asterisk signifies $p < 0.05$, two asterisks signify $p < 0.01$. In the urn task information (%) corresponds to the ratio of the number of revealed balls to the total number of balls, in the motor task to the ratio of visible size to total size of the ambiguous target.

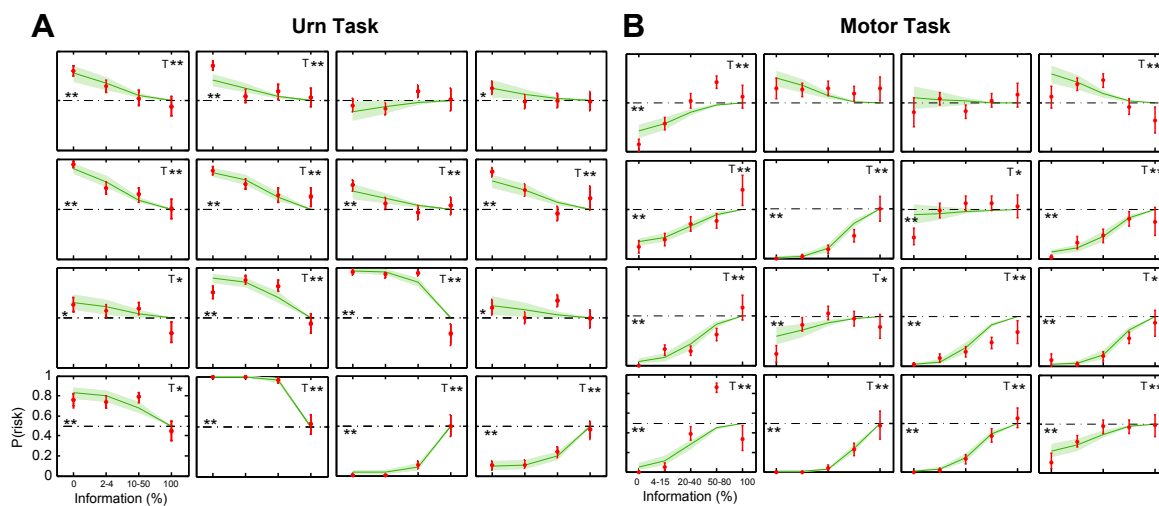


Figure 3.6. Experiment 1: Individual choice probabilities in probe trials. in the urn task (A) and in the motor task (B). The red data points show subjects' probability of choosing the risky option in dependence of the amount of information revealed from the ambiguous option. The error bars indicate 80% confidence intervals. In probe trials, an expected utility decision-maker should always be indifferent between the two options, independent of the information (dashed lines). The shaded green line shows maximum likelihood model fits for subjects' choice probability according to Equation 3.4 in the main manuscript. Asterisks on the first data point in each panel denote a significant difference from the dashed expected utility line. Asterisks in the top right corner of each panel indicate significance of trend. One asterisk signifies $p < 0.05$, two asterisks signify $p < 0.01$. In the urn task information (%) corresponds to the ratio of the number of revealed balls to the total number of balls, in the motor task to the ratio of visible size to total size of the ambiguous target. In total there were 16 subjects performing the two tasks. Each subject can be identified by their panel position.

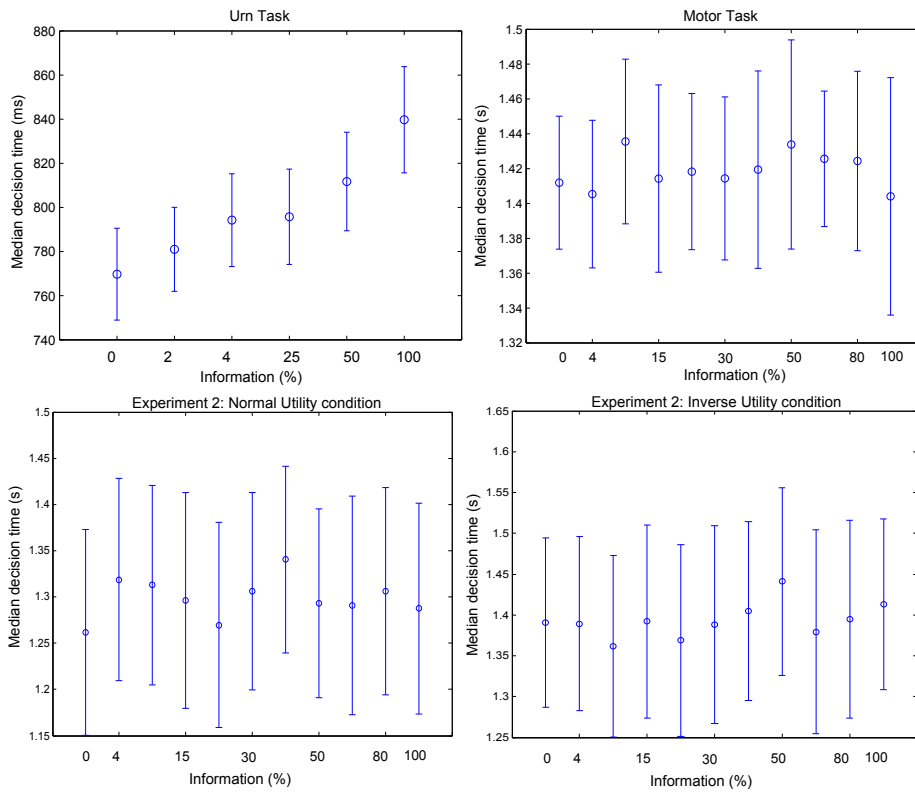


Figure 3.7. Average decision time across subjects in dependence of the amount of information revealed: in the urn task (top-left panel), the motor task (top-right panel) and the control tasks of Experiment 2 (bottom-left and bottom-right panel). Error bars indicate standard errors. In the urn task, the decision time was defined as the time from entering the grey start bar to crossing into the orange zone displayed in Fig. 3.1 of the main text. In the motor task and Experiment 2, the decision time was defined as the time from entering the red square to entering one of the decision circles. Note that the decision time in the urn task was recorded for all 16 subjects, but in the motor task only for the last 8 subjects, and in Experiment 2 for all 25 subjects.

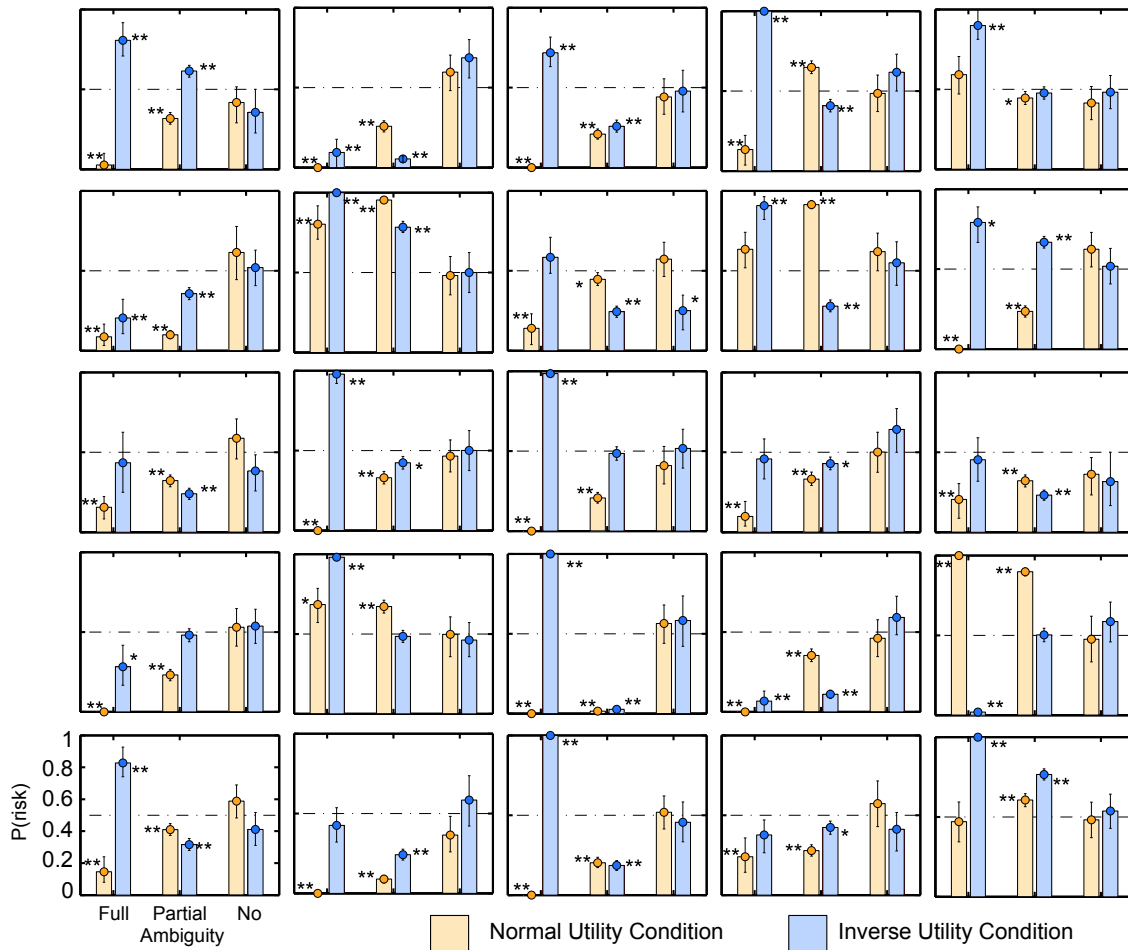


Figure 3.8. Choice probabilities of control experiment. Individual choice probabilities in probe trials of the same subjects performing in the inverse utility (A) and the normal utility condition (B) under full, partial and no ambiguity. The data points show subjects' probability of choosing the risky option in dependence of the amount of information revealed from the ambiguous option. In probe trials, an expected utility decision-maker should always be indifferent between the two options (dashed lines). In 25 subjects, 11 subjects changed from general ambiguity preference in the normal utility condition to a mixed behavior in the inverse utility condition as reflected in the population average shown in the main manuscript. These subjects maintain ambiguity preference for partially ambiguous target bars, but become ambiguity averse in the full ambiguity condition. Six subjects maintained their ambiguity preference across utility conditions in line with the hypothesis that the stimulus induces a stable ambiguity attitude across all ambiguity conditions. Four subjects switched their ambiguity preference across utility conditions in line with a biased belief or perceptual distortion hypothesis. Asterisks on the data points denote a significant deviation from the dashed expected utility line. One asterisk signifies $p < 0.05$, two asterisks signify $p < 0.01$.

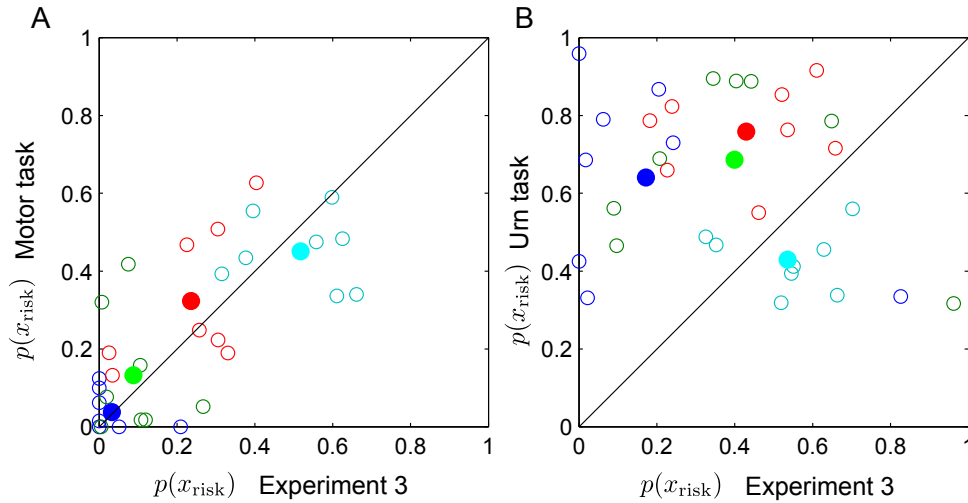


Figure 3.9. Comparison between choice probabilities. Comparison between choice probabilities in Experiment 3 against choice probabilities in the motor task ((**A**), 8 subjects, first group) and in the urn task ((**B**), 8 subjects, second group)—see Methods. Each open circle corresponds to a subject’s choice probability in one of the ambiguity conditions. The different colors indicate the ambiguity condition ranging from cyan, red, green and blue to denote the range from zero ambiguity to full ambiguity. Data points close to the diagonal line imply that the ambiguity preference of subjects remains stable across tasks (as in (**A**)), data points far from the diagonal line indicate that ambiguity attitudes of subjects changed (as in (**B**)), thus meaning that Experiment 3 and the motor task induced similar ambiguity attitudes. Filled circles denote the average across individual data points for each ambiguity condition.

Chapter 4

The Effect of Model Uncertainty on Cooperation in Sensorimotor Interactions

This chapter is a reproduction of the already published work in ([Grau-Moya et al., 2013](#)).

Abstract

Decision-makers have been shown to rely on probabilistic models for perception and action. However, these models can be incorrect or partially wrong, in which case the decision-maker has to cope with model uncertainty. Model uncertainty has recently also been shown to be an important determinant of sensorimotor behavior in humans that can lead to risk-sensitive deviations from Bayes optimal behavior towards worst-case or best-case outcomes. Here we investigate the effect of model uncertainty on cooperation in sensorimotor interactions similar to the stag hunt game, where players develop models about the other player and decide between a payoff-dominant cooperative solution and a risk-dominant non-cooperative solution. In simulations we show that players who allow for optimistic deviations from their opponent model are much more likely to converge to cooperative outcomes. We also implemented this agent model in a virtual reality environment and let human subjects play against a virtual player. In this game subjects' payoffs were experienced as forces opposing their movements. During the experiment we manipulated the risk-sensitivity of the computer player and observed human responses. We found not only that humans adaptively changed their level of cooperation depending on the risk-sensitivity of the computer player, but also that their initial play exhibited characteristic risk-sensitive biases. Our results suggest that model uncertainty is an important determinant of cooperation in two-player sensorimotor interactions.

Introduction

When interacting with its environment, the human sensorimotor system has been shown to employ predictive models for control and estimation (Blakemore et al., 1998; Flanagan and Wing, 1997; Kawato, 1999; Mehta and Schaal, 2002; Shadmehr and Mussa-Ivaldi, 1994; Daniel M Wolpert et al., 1995). These models are thought to be probabilistic in nature and considerable evidence suggests that learning of such models is consistent with the process of Bayesian inference (Doya et al., 2007; Ernst and Banks, 2002; Knill and Pouget, 2004; K. P. Körding and Daniel M Wolpert, 2004). Such probabilistic models are not only important for perception, but they can also be used for decision-making and motor control (Todorov, 2004; Todorov and Jordan, 2002; Trommershäuser et al., 2003a; Trommershäuser et al., 2003b; Trommershäuser et al., 2008; Wu et al., 2009). Importantly, decision-makers that maximize expected gain (or minimize expected costs) require probabilistic models of their environment so that they can determine an expectation value. However, such optimal decision-makers have no performance guarantees if their model happens to be partially incorrect or plain wrong (Hansen and Sargent, 2008). This raises the issue of decision-making strategies that do not rely on accurate probabilities. An extreme example strategy that completely dispenses with probabilities altogether are maximin-strategies where the decision-maker picks an action that is optimal under the assumption of a worst-case scenario (or minimax-strategies in the case of costs). Such a decision-maker, for example, would take out insurance not for the calamity with highest expected costs, but the most disastrous (possibly low-probability) calamity, because of not knowing the probability. Similarly, an extremely optimistic decision-maker would assume a best-case scenario following a maxmax-strategy (or a minmin-strategy in the case of costs), for example, by buying lottery tickets with the highest prize, independent of the presumed winning probabilities. Risk-sensitive decision-makers strike a compromise between the two extremes: they have a probabilistic model that they distrust to some extent, but they do not completely dismiss it—though the extreme cases of robust or optimistic and expected gain decision-making can also be considered as risk-sensitive limit cases (Whittle, 1981).

More formally, we can think of a decision-maker that considers model uncertainty, in the following way (Hansen and Sargent, 2008; Maccheroni et al., 2006). Initially, the decision-maker has a probabilistic model p_0 , but knowing that this model may not be entirely accurate, the decision-maker allows deviations from it, which leads to a new effective probabilistic model p . The transformation between p_0 and p has to be constrained when the decision-maker is very confident about the model. Conversely, when the decision-maker is very insecure about the correctness of the model, there should be leeway for larger deviations. The effective value of a choice set with outcomes x under the effective probability p can then be stated as

$$V = \operatorname{ext}_{p(x)} \left[\int dx p(x) U(x) - \frac{1}{\beta} \int dx p(x) \log \frac{p(x)}{p_0(x)} \right] \quad (4.1)$$

where the utility $U(x)$ quantifies the desirability of x . The first term is the expected utility under p , and the second term—formed by the cost factor $\frac{1}{\beta}$ times the Kullback-Leibler divergence—captures the cost of the transformation from p_0 to p . When $\frac{1}{\beta} > 0$ we have to

replace the extremum operator with a max operator (concave maximization), when $\frac{1}{\beta} < 0$ we have to replace it with a min operator (convex minimization). Sensitivity to model uncertainty is modulated by β . When $\beta \rightarrow 0$ we recover a decision-maker without model uncertainty. For $\beta \rightarrow -\infty$ we get a maximin decision-maker who picks the choice set with maximum V , where each V considers the worst-case scenario of the choice set. In fact, the quantity V is a free energy difference, and equation (4.1) can be motivated by statistical physics—see section 4.1.1.

Recently, it was found that model uncertainty also affects decision-making in sensorimotor integration tasks where subjects have to form beliefs about latent variables, for example the position of a hidden target (Grau-Moya et al., 2012). However, latent variables do not only play an important role in single-player environments, but also in multi-player sensorimotor interactions (Braun et al., 2009b; Braun et al., 2011c), where the policy of the other player can be considered as a particular latent variable. Sensorimotor interactions in humans range from hand shaking and avoiding collisions with another passerby to tandem riding, tango dancing and arm wrestling. An important latent variable in such two-player interactions is for example the strategy of the other player. As in the case of single-player environments, the presence of a latent variable suggests the formation of a belief model that can be exploited for prediction and planning (Doya et al., 2007; Ernst and Banks, 2002; Knill and Pouget, 2004; K. P. Körding and Daniel M Wolpert, 2004). And as in the case of the single-player environment, decision-makers might exhibit model uncertainty (Grau-Moya et al., 2012). Especially, when meeting a player for the first time, only a little information about this player’s strategies is available. The initial trust or distrust with respect to this player can be thought of as an optimistic or pessimistic bias. However, as more information about the unknown player becomes available such deviations should vanish and be replaced by accurate statistical estimates.

Sensorimotor interactions can be of cooperative nature, as in the case of dancing, or of competitive nature as in the case of arm wrestling. To investigate the effect of model uncertainty on cooperation, we study sensorimotor interactions similar to the stag hunt game. In the stag hunt game each player decides whether to hunt a highly valued stag or a lower-valued hare. However, the stag is caught only if both players have decided to hunt stag. In contrast, a hare can be caught by each player independently. The stag hunt game is a coordination game with two pure Nash equilibria, given by the payoff-dominant stag solution, where both players hunt stag and achieve the highest possible payoff, and the risk-dominant hare solution, where both players hunt hare and obtain a lower payoff. The latter solution is called risk-dominant, because a player hunting hare knows exactly the payoff he will receive, which is higher than he would get if he hunted stag by himself. The stag hunt game is therefore often used to study the emergence of cooperation.

In our study we investigate a decision-making model that forms Bayesian beliefs about the other player’s strategy based on empirically observed choice frequencies. In simulations we study how model uncertainty with respect to these beliefs affects cooperation in a stag-hunt-like setting. To test human behavior in stag-hunt-like sensorimotor interactions, we employ a previously developed paradigm that allows translating 2x2 matrix games into sensorimotor

interactions (Braun et al., 2009b; Braun et al., 2011c). In the experiment one of the players is simulated by a virtual computer player that is based on our risk-sensitive decision-making model. This way, we can directly manipulate the risk-sensitivity of the artificial player and observe the response of the human player.

4.1 A Risk-sensitive Model of Interaction

Classic models in game theory are usually equilibrium models that predict the occurrence of Nash equilibria, that is joint settings of strategies where no individual player has any incentive to deviate unilaterally from their strategy (Osborne and Rubinstein, 1999). In evolutionary game theory this problem is addressed by developing dynamic learning models that converge to the equilibria (Jörgen W Weibull, 1997). One of the simplest classes of such learning models is *fictitious play* (Berger, 2007; Koopmans et al., 1951; Krishna and Sjostrom, 1998). In fictitious play it is assumed that the other player plays with a stationary strategy, which is estimated by the hitherto observed empirical choice frequencies. In our model we also adopt the assumption of modeling the other player with a stationary strategy, but form a Bayesian belief about this strategy. In the case of the stag hunt game this strategy is a distribution over a binary random variable that indicates the two possible actions, namely whether to hunt stag or hare. This distribution can be expressed as a beta distribution. After observing s choices of stag and h choices of hare from the opponent, the decision-maker's belief about the strategy x of the other player is then given by

$$P(x|s, h) = \frac{x^s(1-x)^h}{\int_0^1 dx x^s(1-x)^h}, \quad (4.2)$$

where the opponent's stationary strategy is represented by the probability x of choosing stag. For a known strategy x^* of the opponent where $p(x) = \delta(x - x^*)$, the decision-maker faces the following expected payoff

$$EU_1(a_1|x^*) = x^*U(a_1, a_2 = S) + (1 - x^*)U(a_1, a_2 = H), \quad (4.3)$$

with $U(a_1, a_2)$ denoting the player's payoff if he chooses action a_1 and the opponent chooses a_2 . Under strategy x^* the opponent chooses $a_2 = S$ with probability x^* and $a_2 = H$ with probability $1 - x^*$. In fictitious play the decision-maker simply gives a best response to this expected payoff, where x^* is given by the empirical frequencies and corresponds to the mean of the beta distribution. In contrast, we construct a decision-maker that takes the uncertainty over the x -estimate into account and exhibits risk-sensitivity with respect to this belief over x . This can be achieved by inserting (4.2) as p_0 and (4.3) as $U(x)$ into equation (4.1), which results in

$$V(a_1) = \frac{1}{\beta} \log \int_0^1 dx P(x|s, h) e^{\beta(xU(a_1, a_2=S) + (1-x)U(a_1, a_2=H))}$$

The value $V(a_1)$ assigned to each action depends on the parameter β , that in our case represents the risk-sensitivity. For action selection, we assume a soft-max decision rule

$$P(a_1) = \frac{e^{\alpha V(a_1)}}{\sum_{a'_1} e^{\alpha V(a'_1)}}, \quad (4.4)$$

where α is a rationality parameter that regulates how deterministic the response is. Soft-max decision rules are prevalent in Quantal Response Equilibrium models to formalize the bounded rationality of decision-makers in games (McKelvey and Palfrey, 1995). This includes the theoretically best response in the limit $\alpha \rightarrow \infty$ that corresponds to a perfect rational agent that is able to distinguish between tiny differences in the values V . At the other end of the spectrum is a decision-maker with $\alpha \rightarrow 0$, which leads to $P(a_1) \rightarrow 0.5$ corresponding to an irrational agent that only produces random actions. In the remainder of the paper we will refer to $P(a_1)$ also as λ_1 if chosen by player 1 and λ_2 if chosen by player 2.

The expression for the value V also models the learning process of the parameter x . In the limit when x is determined completely, then the distribution $p(x|H, S)$ is going to approach a delta-function in x . In that case the integral collapses and the free energy becomes equal to the expected payoff. Fictitious play is therefore obtained in the limit of $p(x|H, S) \rightarrow \delta(x - x^*)$ and $\alpha \rightarrow \infty$. Before this limit is reached the distribution $p(x|H, S)$ captures the uncertainty over the opponent and the temperature parameter beta determines the risk-sensitivity with respect to this distribution. In the infinitely risk-seeking limit $\beta \rightarrow \infty$ the decision-maker is so optimistic about the stag outcome that he will ignore any information to the contrary, and such a player will always cooperate independent of the history of the game. This is because

$$\begin{aligned} \lim_{\beta \rightarrow \infty} V(a_1) &= \max_{a_1} (xU(a_1, a_2 = S) + (1 - x)U(a_1, a_2 = H)) \\ &= U(a_1, a_2 = S) \end{aligned}$$

Similarly, an infinitely risk-averse decision-maker ($\beta \rightarrow -\infty$) is so pessimistic that he will only expect the worst case scenario. This decision-maker will never cooperate independent of any experienced play. For any finite settings of α and β both cooperative and non-cooperative solutions can occur.

4.1.1 Model Uncertainty and Statistical Physics

The central idea of having model uncertainty is that we do not fully trust our probabilistic model $p_0(x)$ of a latent variable x that we are trying to model. We therefore bias our estimates of x taking into account our utility function $U(x)$. If we are extremely pessimistic and cautious, for example, we will completely dismiss our probability model and simply assume a worst-case scenario. We then pick the action with the best worst-case scenario. If we fully trust our probability model, we will pick the action with the highest *expected* utility. But if we are a risk-averse decision-maker with a finite amount of model uncertainty, we compromise between the two extremes and bias our probability model towards the worst-case to some extent.

This decision-making scenario can be translated in terms of state changes in physical systems, where we start with a probability distribution $p_0(x)$ and end up with a new distribution $p(x)$, because we have added an energy potential $\Delta\phi(x)$ to the system. In this analogy energy plays the role of a negative utility. In physics, a statistical system in equilibrium can be described by a Boltzmann distribution $p_0(x) = \frac{1}{Z_0} e^{-\beta\phi_0(x)}$ with inverse temperature $\beta = \frac{1}{kT}$, energy potential $\phi_0(x)$ and partition sum Z_0 . The distribution p_0 is called an equilibrium distribution, because it minimizes the free energy

$$F_0[q] = \sum_x q(x)\phi_0(x) + \frac{1}{\beta} \sum_x q(x) \log q(x)$$

such that $p_0 = \operatorname{argmin}_q F[q]$ with $F[p_0] = -\frac{1}{\beta} \log Z_0$. If an energy potential $\Delta\phi(x)$ is now added to the system, the new equilibrium distribution that will arise is given by $p(x) = \frac{1}{Z_1} e^{-\beta(\phi_0 + \Delta\phi(x))} = \frac{1}{Z_1} q(x) e^{-\beta\Delta\phi(x)}$. This equilibrium distribution minimizes a free energy $F_1[q]$

$$F_1[q] = \sum_x q(x) \left(\phi_0(x) + \Delta\phi(x) \right) + \frac{1}{\beta} \sum_x q(x) \log q(x).$$

The distribution $p = \operatorname{argmin}_q F_1[q]$ can be interpreted as the biased model. If the inverse temperature β is low, p is going to be very similar to p_0 , if the inverse temperature β is high then p is going to be biased towards low-energy outcomes of the added potential $\Delta\phi$. In the KL-control setting (Kappen, 2005a; Kappen et al., 2012; Todorov, 2009), p_0 is the equilibrium distribution resulting from the uncontrolled dynamics, whereas p corresponds to the controlled dynamics.

Both free energies can be combined into a free energy difference as a single variational principle such that

$$\Delta F[q] = F_1[q] - F_0[p_0]$$

and $p = \operatorname{argmin}_q \Delta F[q]$ such that $\Delta F[p] = -\frac{1}{\beta} \log \sum_x q(x) e^{-\beta\Delta\phi(x)}$. When replacing $\Delta\phi(x) = -U(x)$, we recognize in $-\Delta F[q]$ the same variational principle as suggested in equation (4.1) to describe model uncertainty. This variational principle has recently been suggested as a principle for decision-making with information-processing costs (Braun et al., 2011b; Ortega and Braun, 2011; Ortega and Braun, 2013). Moreover, in non-equilibrium thermodynamics the same expression for the free energy difference $\Delta F[p]$ can be obtained from the Jarzynski equation for infinitely fast switching between the two states. Crucially, the Jarzynski equation holds for any switching process between the two states, and generalizes classical results for infinitely slow and fast switching (C. Jarzynski, 1997). When the utilities are negative log-likelihoods of outcomes under a generative model, this becomes the free energy principle that has recently been proposed to model action and perception in living systems trying to minimize surprise (Friston, 2010).

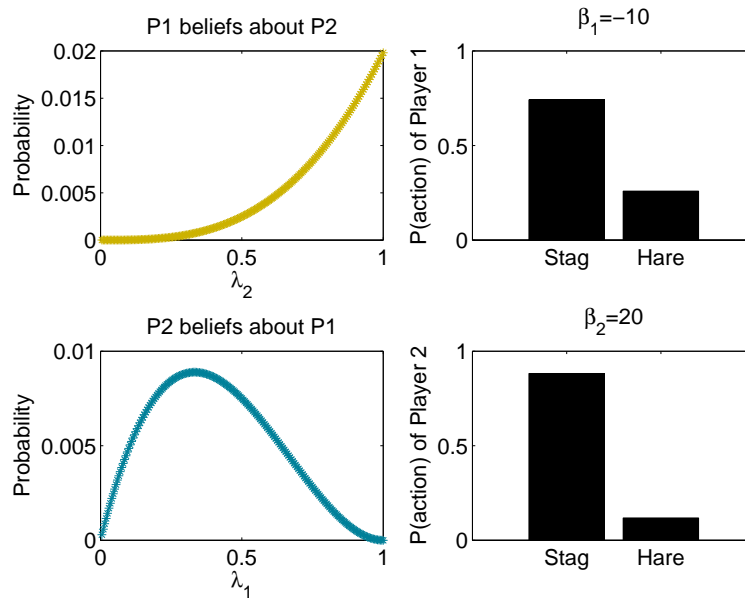


Figure 4.1. Belief and action probabilities. (Left) Belief probability of player 1 (top) and player 2 (bottom) after observing three actions of the other player. Player 1 observed three cooperative actions and player 2 observed one cooperative and two non-cooperative actions. Accordingly, player 1 has more probability mass on the right half, whereas player 2 has more probability mass on the left half. (Right) Action probability of player 1 (top) and player 2 (bottom) resulting from the beliefs and the player’s risk-sensitivity. Player 1 has a higher probability to cooperate even though he is risk-averse, due to the strong evidence of cooperation. Player 2 also places high probability on cooperation because he is strongly risk-seeking, even though the evidence points more towards a non-cooperative opponent.

4.2 Simulation Results

To illustrate the behavior that arises when two decision-makers interact following Equation (4.4), we simulated two model players with rationality parameter $\alpha_1 = \alpha_2 = 10$ and risk-sensitivity parameters $\beta_1 = -10$ and $\beta_2 = 20$ for player 1 and 2 respectively. In Figure 4.1 we depict beliefs and action probabilities of the two players after the pessimistic player 1 played stag once and hare twice, and the very optimistic player 2 played stag three times in a row. Accordingly, player 1’s belief about player 2 is biased towards cooperative strategies (top left panel), whereas player 2’s belief about player 1 is biased towards non-cooperative strategies (bottom left panel). Despite being risk-averse player 1 has a higher probability for cooperation, given the strong evidence of cooperative behavior of player 2. In contrast, player 2 has evidence of non-cooperativeness of player 1, but because he is optimistic, he most probably chooses to cooperate anyway. In Figure 4.2 it can be seen how both players converge to a cooperative equilibrium after 25 interactions. In the left panel the mean and standard deviations of the beta distribution beliefs of the two players are shown over the course of the 25 trials. It can be seen that both beliefs converge towards cooperative strategies, implying

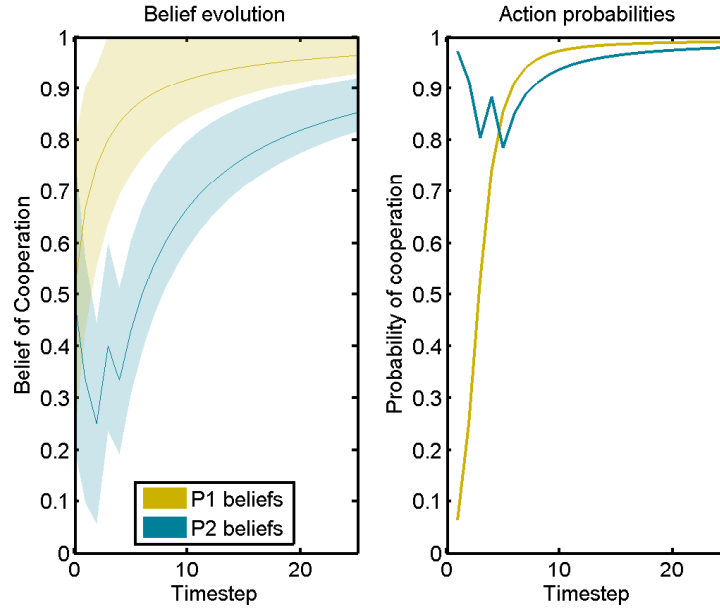


Figure 4.2. Evolution of belief and action probabilities over 25 trials. (Left) Mean and standard deviation of the beta distribution reflecting each player’s beliefs. (Right) Action probabilities of the players according to Equation (4.4). The third trial corresponds to the beliefs and actions displayed in Figure 4.1.

that both players believe in the cooperativeness of the other player. In the right panel the action probabilities of choosing stag for both players are shown. Both action probabilities converge to cooperative strategies.

In the bottom row of Figure 4.3 we show the probability of a cooperative equilibrium after 25 interactions depending on all possible combinations of risk-sensitivities of the two players ranging from risk-averse ($\beta = -20$) to risk-seeking ($\beta = +20$). In this simulation the rationality of player 1 was always set to $\alpha_1 = 10$, whereas the rationality of player 2 was set to $\alpha_2 = 2$ (right panels) or $\alpha_2 = 10$ (left panels). The prior probability of cooperation before any interaction has taken place is shown in the upper panels. For uninformative priors the probability of cooperation in the first trial is greater than one half for all risk-seeking decision-makers and lower than one half for all risk-averse decision-makers independent of the opponent’s risk-sensitivity. Naturally, in later interactions the opponent’s risk-sensitivity comes to bear. If both players have positive risk-sensitivities there is a higher probability they will end up cooperating, and similarly if both players have negative risk-sensitivities there is an increased probability they will end up with a non-cooperative equilibrium. If one of the players is risk-seeking and the other one risk-averse, then the player whose risk-sensitivity has higher absolute value will more probably drive the behavior of the interaction towards cooperation if risk-seeking or non-cooperation if risk-averse. If player 2 has a low rationality $\alpha_2 = 2$ the overall pattern is similar, but more noisy.

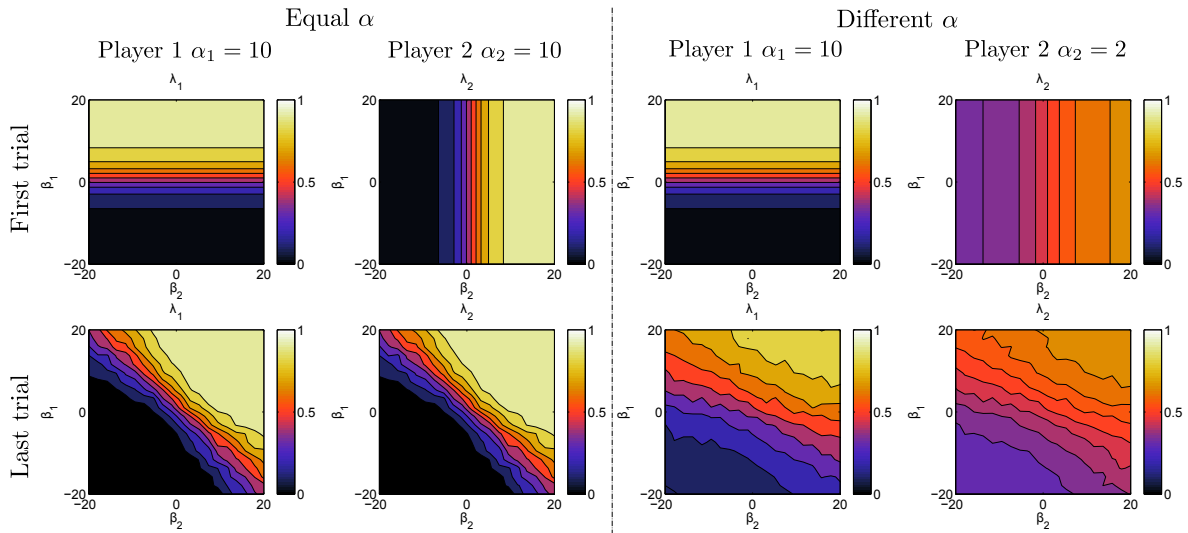


Figure 4.3. Probability of choosing the cooperative action for both players with different risk-sensitivities and rationality parameters. (Top) Prior probability of cooperating. In the first trial the probability of cooperation only depends on the risk-sensitivity of the player and does not depend on the risk-sensitivity of the opponent. (Bottom) Probability of cooperating after 25 trials. In later trials the probability of cooperation depends on the risk-sensitivity of both the player and the opponent. (First two columns) Probability of cooperation when both players have equal rationality α . (Last two columns) Probability of cooperation when players have different rationality α . The probability of cooperating was computed according to Equation (4.4).

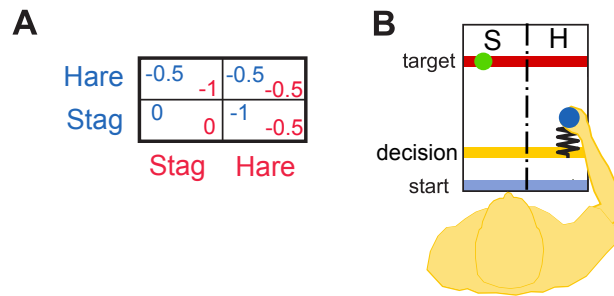


Figure 4.4. The sensorimotor stag hunt game. A. Payoff matrix of the game. B. Experimental Methods. Subjects had to move a cursor from the start bar to the target bar. The left half of the workspace corresponded to selecting “Stag”, and the right half of the workspace corresponded to selecting “Hare”. Once they crossed the decision line a circle on the target line indicated the choice of the virtual player and subjects experienced a force opposing their forward movement that depended on their selection and the virtual opponent’s action selection which followed equation (4.4).

4.3 Experimental Methods

To investigate the effect of risk-sensitivity in sensorimotor interactions in human subjects, we employed a previously developed virtual reality paradigm to translate 2x2 matrix games into sensorimotor games (Braun et al., 2009b; Braun et al., 2011c). One of the players was always simulated by a virtual agent modeled by equation (4.4). This way, we could directly manipulate the risk-sensitivity of the virtual player and record subjects’ responses to these changes.

4.3.1 Experimental Design

As illustrated in Figure 4.4B, participants held the handle of a robotic interface with which they could control the position of a cursor on a display. On each trial, participants had to move the cursor from a start bar to a target bar and back. Importantly, they could do so choosing any lateral position within the width of the target bar. Therefore, participants could achieve the task with their final hand position anywhere between the left and right target bounds. During the forth-and-back movement to the target, subjects had to cross a yellow decision line at 3cm into the movement. Once the line was crossed, both the subject’s and the virtual player’s decision were made. The left half of the subject’s lateral workspace represented the cooperative stag solution, while the right half represented the non-cooperative hare solution.

An implicit payoff was placed on the movements beyond the decision line by using the robot to generate a resistive force opposing the forward motion of the handle. The forces were generated by simulating springs that acted between the handle and the yellow decision bar. The stiffness of the spring during the movement depended on the lateral position of the handle at the time of crossing the decision line and the computer player’s choice. The spring constant was determined by the payoff indicated in Fig. 4.4A and multiplied by a constant

factor of $1.9N/cm$. For successful trial completion, the target bar had to be reached within $1200ms$. The distance of the target bar from the start bar was sampled randomly each trial from a uniform distribution between $15cm$ and $25cm$. Subjects performed two sessions where they faced virtual players with two different rationality parameters. In the first session the rationality of the virtual player was $\alpha_2 = 10$ and subjects performed 40 sets of 25 trials, where the virtual player could assume one of five different β_2 -values from the set $[\pm 20, \pm 10, 0]$. At the beginning of each set the β_2 -parameter of the virtual player was determined and remained constant throughout the set. Each β_2 -parameter was chosen eight times, but in randomized order. In the second session the rationality of the virtual player was set to $\alpha_2 = 2$ and subjects performed again 40 sets of 25 trials each with different β -parameters. At the start of every session they had between 100 and 125 training trials where they could see the degree of risk-sensitivity of the virtual player displayed on a bar.

4.3.2 Experimental Apparatus

The experiments were conducted using two planar robotic interfaces (vBOTs) (Ian S Howard et al., 2009). Participants held a handle of the vBOT, which constrained hand movements to the horizontal plane. A virtual reality system was used to overlay visual feedback onto the plane of movement and players were prevented from seeing their own hand. The vBOT allowed us to record the position of the handle and to generate forces on the hand with a 1 kHz update rate.

4.3.3 Participants

Six naïve participants from the student pool of the Eberhard-Karls-Universität Tübingen took part in the study. All experimental procedures were approved by the ethics committee of the medical faculty at the university of Tübingen.

The precise instructions given to subjects are described below. Subjects were told that they were playing a game against a virtual player and that they could choose between two actions in every trial: either to cooperate or not to cooperate. They were instructed to make their choice by moving the handle across the decision line either in the right or left half of the workspace and that the left half corresponded to cooperation, whereas the right half corresponded to non-cooperation. They were also informed that there would be a force opposing their movement between the decision line and the target line. They were told that in case of non-cooperation they would always experience the same medium force, but that in case of cooperation the force would depend on the choice of the virtual player, who could choose to cooperate or not to cooperate. In case both players cooperate their would be no force, but if the virtual player chooses not to cooperate there would be a very high force. Subjects were also told that the virtual player can learn and adapt to the subject's play.

At the beginning of each block of training trials, subjects could see a bar displaying the degree of the virtual player's risk-sensitivity and they were told that the bar indicates the virtual player's attitude towards cooperation. They were also told that there was a different

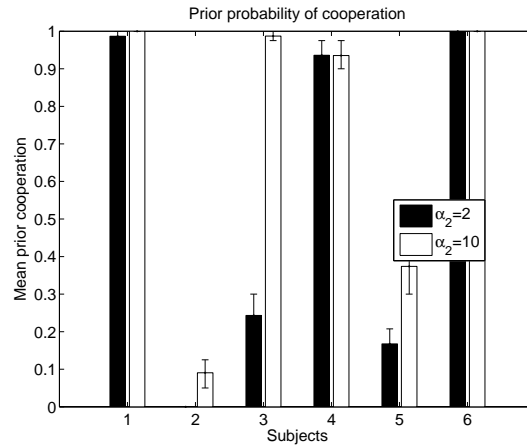


Figure 4.5. Prior cooperation probabilities in human subjects playing virtual opponents with high ($\alpha_2 = 10$, white color) and low ($\alpha_2 = 2$, black color) rationality. In the first trial, when facing a new opponent, subjects knew the rationality of the opponent, but not their risk-sensitivity.

player with a different attitude every 25 trials. After the training trials, they were told that the bar would be no longer displayed and that they can learn the player’s attitude towards cooperation only from actual play, and that there would be a different player every 25 trials. Between blocks of 25 trials there was a short break to mark this transition clearly.

4.4 Results

In Figure 4.5 we display subjects’ prior cooperation probabilities in the first trial of every set of 25, when they face a novel virtual player. In white color this is shown for virtual players with rationality $\alpha_2 = 10$ and in black color this is shown for virtual players with rationality $\alpha_2 = 2$. In the $\alpha_2 = 10$ condition, we found that four out of six subjects chose to cooperate most of the time in the first trial. In the $\alpha_2 = 2$ condition, only three out of six subjects chose to cooperate. This implies that about half of our subjects were risk-seeking and optimistic about cooperation, whereas the others were risk-averse and pessimistic.

After the first trial, subjects received feedback about the choice of the virtual player and could make a first inference about the virtual player’s willingness to cooperate. Accordingly, subjects’ probability of cooperation in subsequent trials in a set of 25 needs to be investigated separately for the different risk-sensitivities of the virtual players. For the extreme risk-sensitivities of $\beta_2 = 20$ and $\beta_2 = -20$ this is depicted in Figure 4.6. When playing a risk-averse opponent ($\beta_2 = -20$), subjects mostly converged to non-cooperative behavior (right panels), whereas when playing a risk-seeking opponent ($\beta_2 = 20$), subjects mostly converged to cooperative behavior (left panels). This pattern is clearly demonstrated when facing virtual players with high rationality $\alpha_2 = 10$ (bottom panels), but much more diffuse in the case of virtual players with low rationality $\alpha_2 = 2$ (top panels).

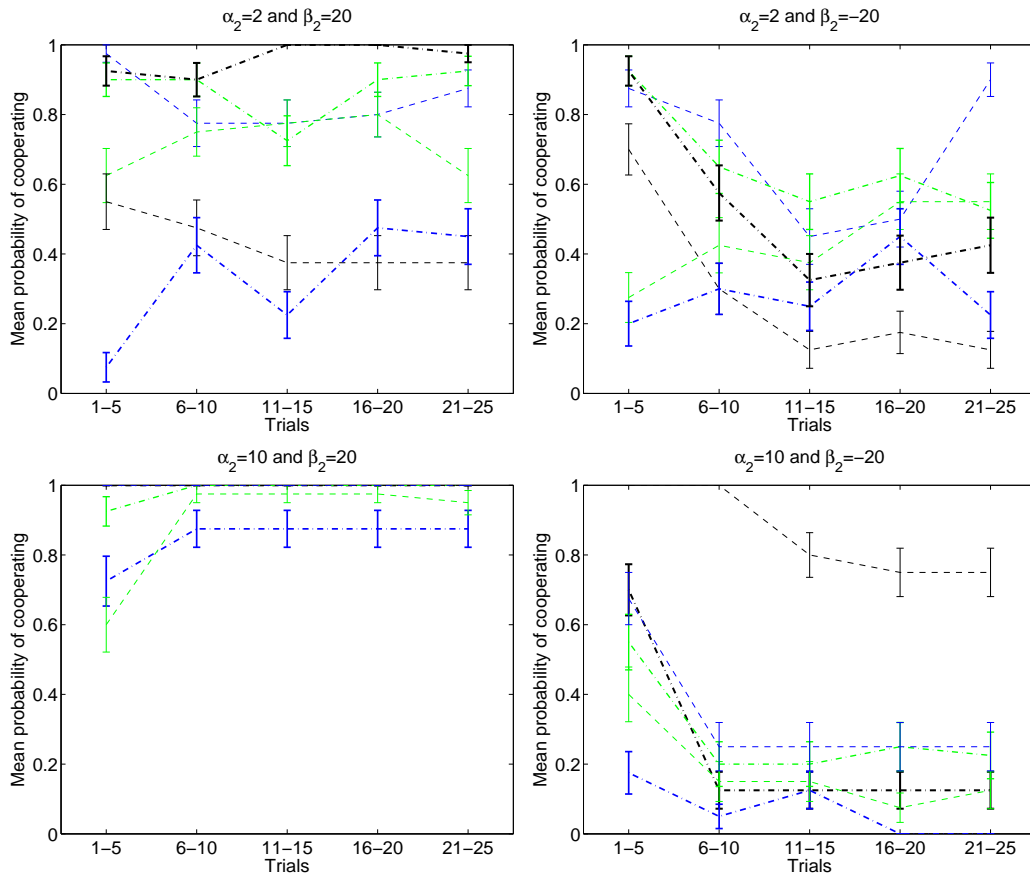


Figure 4.6. Evolution of cooperation over the course of 25 trials of human subjects facing virtual opponents with low ($\alpha_2 = 2$, top) or high ($\alpha_2 = 10$, bottom) rationality and positive ($\beta_2 = 20$, left) and negative ($\beta_2 = -20$, right) risk-sensitivity. Different lines indicate different subjects.

To directly assess the effect of risk-sensitivity on the cooperative behavior of human subjects over all trials, we computed the mean probability of cooperation averaged over all trials where the opponent had the same risk-sensitivity β_2 and rationality α_2 . In Figure 4.7 this is shown for all six subjects playing an opponent with rationality $\alpha_2 = 10$ (left panel) and rationality $\alpha_2 = 2$ (right panel) respectively. For both rationalities, the risk-sensitivity β_2 of the opponent has a significant effect on the probability of cooperation (non-parametric Jonckheere-Terpstra trend test $p < 0.05$ for $\alpha_2 = 2$ and $p < 0.001$ for $\alpha_2 = 10$). However, in the case of high rationality α_2 of the virtual player this effect is stronger and clearer than in the case of inconsistent play resulting from an opponent with low α_2 . The general trend is that subjects' tendency to cooperate increases for higher β_2 and decreases for lower β_2 . Importantly, most subjects deviated on average from a 50 : 50 cooperation probability when playing a risk-neutral opponent of high rationality ($\alpha_2 = 10$), which is another signature of subjects' risk-sensitivity.

To compare the predictive power of our model with the traditional fictitious play model,

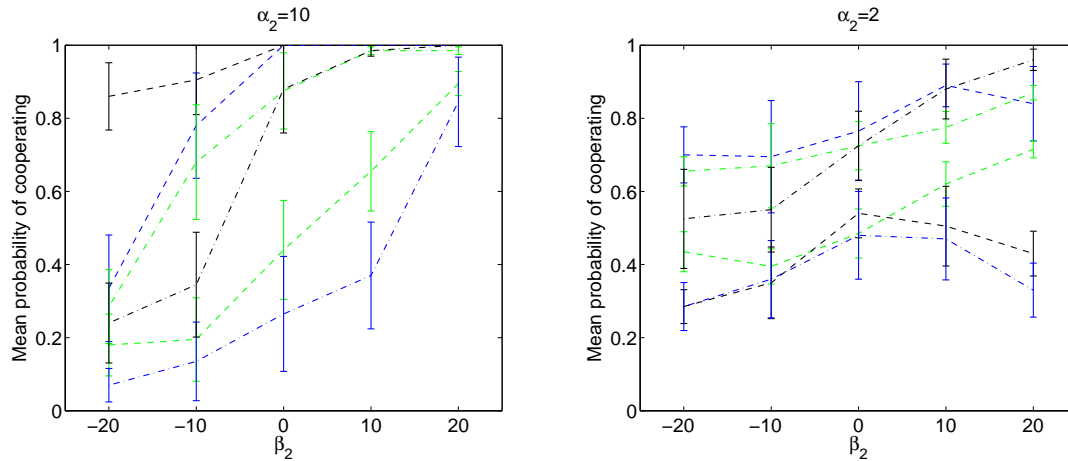


Figure 4.7. (Top row) Average probability of cooperation depending on risk-sensitivity of the opponent with either high ($\alpha_2 = 10$, left) or low ($\alpha_2 = 2$, right) rationality for the six different subjects. Different lines indicate different subjects.

we investigated the ratio of cooperation after subjects had experienced an (approximately) 50 : 50 sequence of actions of the virtual player—i.e. the opponent had (roughly) cooperated half the time and refused cooperation the other half of the time. Importantly, we did this at two different stages of the game such that the 50 : 50-ratio was the result of either a small number of trials (after 2 trials) or a large number of trials (after 10 trials). In the 2-trial case only trials with one Stag- and one Hare-choice were included, however, for the 10-trial case there were not enough instances with an exact 50 : 50 ratio. Therefore, we also included trials between 40% and 60% of cooperation, but still this analysis was only possible in the case of a virtual player with low rationality ($\alpha_2 = 2$). The crucial observation is that in the case of 2 trials the estimate of the other player’s cooperation is highly uncertain, whereas in the case of 10 trials this estimate is much more consolidated. In both cases, fictitious play makes the same prediction, which is the best response to the ratio—compare dashed line in Figure 4.8. In contrast, a risk-sensitive model predicts that the best response should depend on the uncertainty of the estimate of the ratio. For our model predictions we fitted to each subject an α_1 - and a β_1 -parameter by maximizing the log-likelihood of subjects choices given the predicted choice probabilities of Equation (4.4). In particular this predicts that a risk-seeking player will deviate towards cooperation in early trials, whereas a risk-averse player will deviate towards non-cooperation in early trials—compare left plot in Figure 4.8. In late trials, when a large part of the uncertainty has been removed, both players converge to fictitious play.

In the right plot of Figure 4.8 it can be seen that most subjects’ behavior was inconsistent with fictitious play. Subjects 1, 4 and 6 were risk-seeking and deviated significantly towards cooperation in the third trial (one-sided t-test $p < 0.01$). Subject 5 was risk-averse and refused cooperation in early trials ($p < 0.01$). Subjects 2 and 3 were risk-neutral and consistent with

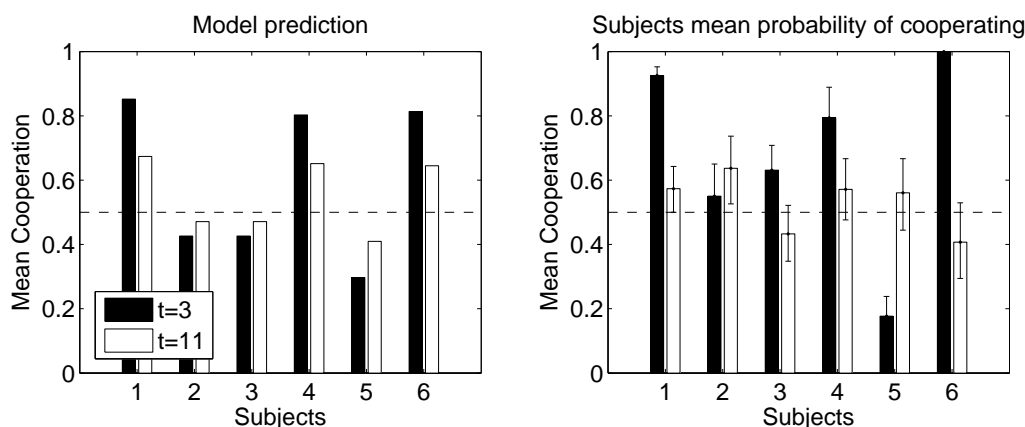


Figure 4.8. Comparison of risk-sensitive predictions to fictitious play and human subjects' behavior. (Left) Predictions of probability of cooperation when observing a sequence with 50% cooperation after 2 trials (black) or 10 trials (white). The dashed line is the prediction of fictitious play. (Right) Subjects' cooperation probabilities when observing a sequence with roughly 50% cooperation after 2 trials (black) or 10 trials (white).

fictitious play and therefore the deviation from 0.5 choice probability was not significant ($p > 0.1$). Importantly, after 10 trials all subjects were consistent with fictitious play and were best-responding to the observed sequence of the opponent's play, hence the deviation from 0.5 choice probability was not significant for all of them ($p > 0.1$).

4.5 Discussion

Most current theoretical frameworks of motor control rely on probabilistic models that are used for prediction, estimation and control. However, when such models are partially incorrect or wrong, there are usually no performance guarantees (Hansen and Sargent, 2008). Model uncertainty is therefore an important factor in real world control problems, because in practice one can never be absolutely sure about one's model. In this paper we investigated risk-sensitive deviations arising from having model uncertainty in sensorimotor interactions. We found that human subjects adapted their cooperation depending on the risk-sensitivity of a virtual computer player. Furthermore, we found that subjects did not only best-respond to the frequency of observed play, but that they were sensitive to the certainty of this estimate. In particular, they allowed for risk-sensitive deviations in initial interaction trials when uncertainty was high. This behavior is consistent with a risk-sensitive decision-maker with model uncertainty.

Recently, it was found that risk-sensitivity is an important determinant in human sensorimotor behavior (Braun et al., 2011a). Risk-sensitive decision-makers do not base their choices exclusively on the expectation value of a particular cost function, but they also consider higher order moments of this cost function. This can be seen when approximating the risk-sensitive

cost function with a Taylor series

$$\begin{aligned} J &= \frac{1}{\beta} \log \int dx p(x) e^{\beta U(x)} \\ &\approx \mathbb{E}[U] + \frac{\beta}{2} \text{VAR}[U], \end{aligned}$$

assuming that $\beta \text{VAR}[U]$ is small (Whittle, 1981). Sensitivity to the second-order moment of the cost function was found, for example, in motor tasks with speed-accuracy trade-off (Nagengast et al., 2011a). Such risk-sensitive decision-makers can be thought of as trading off the mean cost versus the variability of the cost. A mean-variance trade-off in effort was found, for example, in a motor task where subjects had to decide between hitting differently sized targets that were associated with different levels of effort (Nagengast et al., 2011b). Sensitivity to the variance of the control cost was also found in continuous motor tasks, where subjects had to control a cursor undergoing a random walk (Nagengast et al., 2010). The sensitivity to the variance can also be exploited by assistive technologies that consider the human as a (useful) noise source (Medina et al., 2012; Saida et al., 2012).

When x is a latent variable that needs to be inferred, risk-sensitivity also allows decision-makers to take model uncertainty into account. This can be seen when rewriting the risk-sensitive cost function as in equation (4.1) yielding

$$\begin{aligned} J &= \frac{1}{\beta} \log \int dx p(x) e^{\beta U(x)} \\ &= \text{ext}_{p(x)} \left[\int dx p(x) U(x) - \frac{1}{\beta} \int dx p(x) \log \frac{p(x)}{p_0(x)} \right], \end{aligned}$$

where J can be re-expressed as a variational principle that trades off the maximization of a utility term and the deviation from p_0 to p (Hansen and Sargent, 2008). Such model uncertainty was recently found to play a role in a sensorimotor integration task, where subjects had to infer the position of a hidden target (the latent variable) (Grau-Moya et al., 2012). When given feedback information about the target position with varying degree of reliability, subjects' estimates of the target position was consistent with a Bayesian estimator that optimally combines prior knowledge of the distribution of target positions with the actual feedback information. Subjects' behavior was therefore also consistent with previous reports on information integration in sensorimotor tasks (K. P. Körding and Daniel M Wolpert, 2004). However, when subjects' beliefs were associated with control costs, study (Grau-Moya et al., 2012) found that subjects exhibited characteristic deviations from the Bayes optimal response that could be described by a risk-sensitive decision-making model that depended on the level of model uncertainty, the reliability of the feedback and the control cost. These risk-sensitive deviations were particularly prominent in trials with high uncertainty and vanished in the absence of uncertainty as more and more information about the latent variable becomes available.

In the context of model uncertainty, risk-sensitivity can be distinguished from risk-attitudes modeled by the curvature of the utility function, both theoretically and experimentally (Chakravarty

and Roy, 2009; Gilboa and Marinacci, 2011). Utility functions generally express the subjective desirability of an outcome and not necessarily its nominal value. For example, the subjective value of money does typically not increase linearly with the nominal amount. Accordingly, receiving a monetary increase of \$1,000 has more utility for a beggar than for a millionaire. The utility function is said to be marginally decreasing. Intriguingly, this property can also be used to model risk-attitudes. For example, people with a marginally decreasing utility function of money will prefer \$50 for sure over a gamble between a 50 : 50 lottery, where one outcome is \$0 and the other is \$100, because $U(\$50) > \frac{1}{2}U(\$100)$, assuming that $U(\$0) = 0$. Importantly, these risk-attitudes are independent of the level of information about the probabilities. In fact, the probabilities are assumed to be perfectly known. Thus, risk-attitudes are conceptually very different from model uncertainty that vanishes in the limit of perfect information about the probabilities. Model uncertainty captures the lack of information about a lottery.

The effect of risk-attitudes on cooperation in the stag hunt game is investigated in behavioral economics tasks (Büyükboyacı, 2014; Neumann and Vogt, 2009; Al-Ubaydli et al., 2011) in which the risk-attitude of subjects is determined by subjects' choice behavior when deciding between risky and safe lotteries. In these studies it was found that subjects' risk-attitude does not predict their cooperation in the stage hunt game, although players consider information about the other player's risk-attitude. In particular, subjects are less likely to cooperate if they know that their opponent is risk-averse. However, the fact that subjects' risk-attitude is a poor predictor of their cooperation in the game suggests that not risk-attitude, but model uncertainty might be a stronger factor affecting cooperation in the game.

In the traditional stag hunt game payoffs are usually framed as gains, whereas in our experiment the payoffs are framed as losses in shape of forces subjects have to exert. In the economics literature it is well known that the framing of losses versus gains can have a strong influence on human choice behavior (Daniel Kahneman and Amos Tversky, 1979). It is therefore not surprising that different payoff levels have also been found to influence choice behavior in the stag hunt game (Feltovich et al., 2012), in particular, it was found that having losses increases players' probability of choosing the more risky stag. Crucially, our results showing sensitivity to model uncertainty do not depend on the exact shape of the utility function. Expected utility players that have experienced 50 : 50 play of their opponent after N amount of trials will choose between $a_1 = S$ and $a_1 = H$ according to equation (4.3) where $x^* = 0.5$. The decision-maker's preference depends of course on the utilities $U(a_1, a_2)$, but crucially these utilities and the resulting expected utility does not change with varying the amount of trials N as long as the empirical frequency is 50 : 50. The fact that we have used a loss scenario does therefore not invalidate our results on model uncertainty, although the exact choice probabilities might look different in a gain scenario.

Fictitious play is one of the earliest models that were developed to explain learning in games (Fudenberg and Levine, 1998; Koopmans et al., 1951). Crucially, it assumes stationary strategies for both players. It can be shown to converge for a wide class of problems, including all two-player interactions (Berger, 2005). However, it can also be shown that fictitious play

can lead to non-converging limit-cycles for very simple games (Shapley, 1964). In our study we found that subjects were not simply best-responding to the observed frequency of the opponent's play, as presumed by fictitious play. Rather, subjects were sensitive to the amount of information they had gathered about the other player when deciding whether to cooperate or not—compare Figure 4.8. Our risk-sensitive model of cooperation can account for this dependency. However, it still makes the simplistic assumption of stationary strategy beliefs. This limitation may be overcome in the future by considering more complex belief models.

An important objection to risk-sensitive models is often that they could be replaced by a standard risk-neutral Bayesian model under a different (post-hoc) prior (Hansen and Sargent, 2008). This is also true in our case: subjects could develop biased prior beliefs about the population of virtual players. Importantly, the population of virtual players was statistically balanced and there is therefore no statistical reason why subjects should develop biased priors. However, if the prior is thought to reflect not only the (prior) statistics of the environment, but also traits of the decision-maker, then a risk-neutral Bayesian model with a biased prior could, in principle, also explain our data. This is sometimes also discussed in the context of so-called complete class theorems, in which the existence of priors is investigated when modeling Bayesian decision-makers with different loss functions (Brown, 1981; Friston et al., 2012).

The results of our study also speak to cognitive theories of (dyadic) social interactions and joint actions. Several recent studies have investigated how humans mutually adjust and synchronize their behavior during on-line joint actions, revealing the role of several mechanisms that range from automatic entrainment to action prediction (Braun et al., 2011c; Pezzulo and Dindo, 2011; Sebanz et al., 2006; Vesper et al., 2013). An open research question is if and how sensorimotor interactions are influenced by the co-actors' goals and attitudes. Given that socially- and culturally-relevant information (e.g., facial expression, racial or social group membership) is automatically processed in the brain (Cosmides et al., 2003) and can automatically modulate imitation (Losin et al., 2012) and empathy (Avenanti et al., 2010), most studies have focused on the impact of socially-relevant variables in joint actions, with the hypothesis that it could favor pro-social or anti-social behavior. It has been shown that interpersonal perception and (positive and negative) attitude towards the co-actor modulate cooperation and joint actions (Iani et al., 2011; Sacheli et al., 2012). In turn, sensorimotor interactions can modulate a co-actors' attitude; for example, it has been reported that dyads engaged in synchronous interactions improve their altruistic behavior (Valdesolo et al., 2010).

The aforementioned studies focus on social attitudes and leave unanswered the issue of how personal traits and non-social attitudes influence sensorimotor interactions. Here we studied the influence of model uncertainty on the evolution of sensorimotor interactions. We designed a sensorimotor task that is equivalent to the stag hunt game. Our results show that model uncertainty modulates sensorimotor interactions and their success. In particular, optimistic (risk-seeking) adaptive agents are much more likely to converge to cooperative outcomes. Furthermore, humans adaptively change their level of cooperation depending on the risk-sensitivity of their co-actor (in our study, a computer player). Effects of model uncertainty

are particularly strong in early interactions with a novel player. In summary, our results indicate that interacting agents can build sophisticated models of their co-actors ([Yoshida et al., 2008](#)) and use them to modulate their level of cooperation taking model uncertainty into account.

Acknowledgments

This study was supported by the DFG, Emmy Noether grant BR4164/1-1 and EU's FP7 under grant agreement no FP7-ICT-270108 (Goal-Leaders).

Chapter 5

Planning with Information-Processing Constraints and Model Uncertainty in Markov Decision Processes

This chapter is a reproduction of the already published work in ([Grau-Moya et al., 2016a](#)).

Abstract

Information-theoretic principles for learning and acting have been proposed to solve particular classes of Markov Decision Problems. Mathematically, such approaches are governed by a variational free energy principle and allow solving MDP planning problems with information-processing constraints expressed in terms of a Kullback-Leibler divergence with respect to a reference distribution. Here we consider a generalization of such MDP planners by taking model uncertainty into account. As model uncertainty can also be formalized as an information-processing constraint, we can derive a unified solution from a single generalized variational principle. We provide a generalized value iteration scheme together with a convergence proof. As limit cases, this generalized scheme includes standard value iteration with a known model, Bayes Adaptive MDP planning, and robust planning. We demonstrate the benefits of this approach in a grid world simulation.

Introduction

The problem of planning in Markov Decision Processes was famously addressed by Bellman who developed the eponymous principle in 1957 (Bellman, 1957). Since then numerous variants of this principle have flourished in the literature. Here we are particularly interested in a generalization of the Bellman principle that takes information-theoretic constraints into account. In the recent past there has been a special interest in the Kullback-Leibler divergence as a constraint to limit deviations of the action policy from a prior. This can be interesting in a number of ways. Todorov (Todorov, 2006; Todorov, 2009), for example, has transformed the general MDP problem into a restricted problem class without explicit action variables, where control directly changes the dynamics of the environment and control costs are measured by the Kullback-Leibler divergence between controlled and uncontrolled dynamics. This simplification allows mapping the Bellman recursion to a linear algebra problem. This approach can also be generalized to continuous state spaces leading to path integral control (Braun et al., 2011b; Broek et al., 2010). The same equations can also be interpreted in terms of *bounded rational* decision-making where the decision-maker has limited computational resources that allow only limited deviations from a prior decision strategy (measured by the Kullback-Leibler divergence in bits) (Ortega and Braun, 2013). Such a decision-maker can also be instantiated by a sampling process that has restrictions in the number of samples it can afford (Ortega and Braun, 2014). Disregarding the possibility of a sampling-based interpretation, the Kullback-Leibler divergence introduces a control information cost that is interesting in its own right when formalizing the perception action cycle (Tishby and Polani, 2011).

While the above frameworks have led to interesting computational advances, so far they have neglected the possibility of model misspecification in the MDP setting. Model misspecification or model uncertainty does not refer to the uncertainty arising due to the stochastic nature of the environment (usually called risk-uncertainty in the economic literature), but refers to the uncertainty with respect to the latent variables that specify the MDP. In Bayes-Adaptive MDPs (Duff, 2002), for example, the uncertainty over the latent parameters of the MDP is explicitly represented, such that new information can be incorporated with Bayesian inference. However, Bayes-Adaptive MDPs are not robust with respect to model misspecification and have no performance guarantees when planning with wrong models (Mannor et al., 2007). Accordingly, there has been substantial interest in developing robust MDP planners (Iyengar, 2005; Nilim and El Ghaoui, 2005; Wiesemann et al., 2013). One way to take model uncertainty into account is to bias an agent’s belief model from a reference Bayesian model towards worst-case scenarios; thus avoiding disastrous outcomes by not visiting states where the transition probabilities are not known. Conversely, the belief model can also be biased towards best-case scenarios as a measure to drive exploration—also referred in the literature as *optimism in face of uncertainty* (Szita and Lőrincz, 2008; Szita and Szepesvári, 2010).

When comparing the literature on information-theoretic control and model uncertainty, it is interesting to see that some notions of model uncertainty follow exactly the same mathematical principles as the principles of relative entropy control (Todorov, 2009). In this paper

we therefore formulate a unified and combined optimization problem for MDP planning that takes *both*, model uncertainty and bounded rationality into account. This new optimization problem can be solved by a generalized value iteration algorithm. We provide a theoretical analysis of its convergence properties and simulations in a grid world.

5.1 Background and Notation

In the MDP setting the agent at time t interacts with the environment by taking action $a_t \in \mathcal{A}$ while in state $s_t \in \mathcal{S}$. Then the environment updates the state of the agent to $s_{t+1} \in \mathcal{S}$ according to the transition probabilities $T(s_{t+1}|a_t, s_t)$. After each transition the agent receives a reward $R_{s_t, a_t}^{s_{t+1}} \in \mathcal{R}$ that is bounded. For our purposes we will consider \mathcal{A} and \mathcal{S} to be finite. The aim of the agent is to choose its policy $\pi(a|s)$ in order to maximize the total discounted expected reward or value function for any $s \in \mathcal{S}$

$$V^*(s) = \max_{\pi} \lim_{T \rightarrow \infty} \mathbb{E} \left[\sum_{t=0}^{T-1} \gamma^t R_{s_t, a_t}^{s_{t+1}} \right]$$

with discount factor $0 \leq \gamma < 1$. The expectation is over all possible trajectories $\xi = s_0, a_0, s_1 \dots$ of state and action pairs distributed according to $p(\xi) = \prod_{t=0}^{T-1} \pi(a_t|s_t) T(s_{t+1}|a_t, s_t)$. It can be shown that the optimal value function satisfies the following recursion

$$V^*(s) = \max_{\pi} \sum_{a, s'} \pi(a|s) T(s'|a, s) \left[R_{s, a}^{s'} + \gamma V^*(s') \right]. \quad (5.1)$$

At this point there are two important implicit assumptions. The first is that the policy π can be chosen arbitrarily without any constraints which, for example, might not be true for a bounded rational agent with limited information-processing capabilities. The second is that the agent needs to know the transition-model $T(s'|a, s)$, but this model is in practice unknown or even misspecified with respect to the environment's true transition-probabilities, specially at initial stages of learning. In the following, we explain how to incorporate both bounded rationality and model uncertainty into agents.

5.1.1 Information-Theoretic Constraints for Acting

Consider a one-step decision-making problem where the agent is in state s and has to choose a single action a from the set \mathcal{A} to maximize the reward $R_{s, a}^{s'}$, where s' is the next the state. A perfectly rational agent selects the optimal action $a^*(s) = \operatorname{argmax}_a \sum_{s'} T(s'|a, s) R_{s, a}^{s'}$. However, a bounded rational agent has only limited resources to find the maximum of the function $\sum_{s'} T(s'|a, s) R_{s, a}^{s'}$. One way to model such an agent is to assume that the agent has a prior choice strategy $\rho(a|s)$ in state s *before* a deliberation process sets in that refines the choice strategy to a posterior distribution $\pi(a|s)$ that reflects the strategy *after* deliberation. Intuitively, because the deliberation resources are limited, the agent can only afford to deviate from the prior strategy by a certain amount of information bits. This can be quantified by

the relative entropy $D_{\text{KL}}(\pi||\rho) = \sum_a \pi(a|s) \log \frac{\pi(a|s)}{\rho(a|s)}$ that measures the average information cost of the policy $\pi(a|s)$ using the source distribution $\rho(a|s)$. For a bounded rational agent this relative entropy is bounded by some upper limit K . Thus, a bounded rational agent has to solve a constrained optimization problem that can be written as

$$\max_{\pi} \sum_a \pi(a|s) \sum_{s'} T(s'|a, s) R_{s,a}^{s'} \quad \text{s.t. } D_{\text{KL}}(\pi||\rho) \leq K$$

This problem can be rewritten as an unconstrained optimization problem

$$F^*(s) = \max_{\pi} \sum_a \pi(a|s) \sum_{s'} T(s'|a, s) R_{s,a}^{s'} - \frac{1}{\alpha} D_{\text{KL}}(\pi||\rho) \quad (5.2)$$

$$= \frac{1}{\alpha} \log \sum_a \rho(a|s) e^{\alpha \sum_{s'} T(s'|a, s) R_{s,a}^{s'}}. \quad (5.3)$$

where F^* is a free energy that quantifies the value of the policy π by trading off the average reward against the information cost. The optimal strategy can be expressed analytically in closed-form as

$$\pi^*(a|s) = \frac{\rho(a|s) e^{\alpha \sum_{s'} T(s'|a, s) R_{s,a}^{s'}}}{Z_{\alpha}(s)}$$

with partition sum $Z_{\alpha}(s) = \sum_a \rho(a|s) \exp\left(\alpha \sum_{s'} T(s'|a, s) R_{s,a}^{s'}\right)$. Therefore, the maximum operator in (5.2) can be eliminated and the free energy can be rewritten as in (5.3). The Lagrange multiplier α quantifies the boundedness of the agent. By setting $\alpha \rightarrow \infty$ we recover a perfectly rational agent with optimal policy $\pi^*(a|s) = \delta(a - a^*(s))$. For $\alpha = 0$ the agent has no computational resources and the agent's optimal policy is to act according to the prior $\pi^*(a|s) = \rho(a|s)$. Intermediate values of α lead to a spectrum of bounded rational agents.

5.1.2 Information-Theoretic Constraints for Model Uncertainty

In the following we assume that the agent has a model of the environment $T_{\theta}(s'|a, s)$ that depends on some latent variables $\theta \in \Theta$. In the MDP setting, the agent holds a belief $\mu(\theta|a, s)$ regarding the environmental dynamics where θ is a unit vector of transition probabilities into all possible states s' . While interacting with the environment the agent can incorporate new data by forming the Bayesian posterior $\mu(\theta|a, s, D)$, where D is the observed data. When the agent has observed an infinite amount of data (and assuming $\theta^*(a, s) \in \Theta$) the belief will converge to the delta distribution $\mu(\theta|s, a, D) = \delta(\theta - \theta^*(a, s))$ and the agent will act optimally according to the true transition probabilities, exactly as in ordinary optimal choice strategies with known models. When acting under a limited amount of data the agent cannot determine the value of an action a with the true transition model according to $\sum_{s'} T(s'|a, s) R_{s,a}^{s'}$, but it can only determine an expected value according to its beliefs $\int_{\theta} \mu(\theta|a, s) \sum_{s'} T_{\theta}(s'|a, s) R_{s,a}^{s'}$.

The Bayesian model μ can be subject to model misspecification (e.g. by having a wrong likelihood or a bad prior) and thus the agent might want to allow deviations from its model towards best-case (optimistic agent) or worst-case (pessimistic agent) scenarios up to a certain

extent, in order to act more robustly or to enhance its performance in a friendly environment (Hansen and Sargent, 2008). Such deviations can be measured by the relative entropy $D_{\text{KL}}(\psi|\mu)$ between the Bayesian posterior μ and a new biased model ψ . Effectively, this allows for mathematically formalizing model uncertainty, by not only considering the specified model but all models within a neighborhood of the specified model that deviate no more than a restricted number of bits. Then, the effective expected value of an action a while having limited trust in the Bayesian posterior μ can be determined for the case of optimistic deviations as

$$F^*(a, s) = \max_{\psi} \int_{\theta} \psi(\theta|a, s) \sum_{s'} T_{\theta}(s'|a, s) R_{s,a}^{s'} - \frac{1}{\beta} D_{\text{KL}}(\psi|\mu) \quad (5.4)$$

for $\beta > 0$, and for the case of pessimistic deviations as

$$F^*(a, s) = \min_{\psi} \int_{\theta} \psi(\theta|a, s) \sum_{s'} T_{\theta}(s'|a, s) R_{s,a}^{s'} - \frac{1}{\beta} D_{\text{KL}}(\psi|\mu) \quad (5.5)$$

for $\beta < 0$. Conveniently, both equations can be expressed as a single equation

$$F^*(a, s) = \frac{1}{\beta} \log Z_{\beta}(a, s)$$

with $\beta \in \mathbb{R}$ and $Z_{\beta}(s, a) = \int_{\theta} \mu(\theta|a, s) \exp\left(\beta \sum_{s'} T_{\theta}(s'|a, s) R_{s,a}^{s'}\right)$ when inserting the optimal biased belief

$$\psi^*(\theta|a, s) = \frac{1}{Z_{\beta}(a, s)} \mu(\theta|a, s) \exp\left(\beta \sum_{s'} T_{\theta}(s'|a, s) R_{s,a}^{s'}\right)$$

into either equation (5.4) or (5.5). By adopting this formulation we can model any degree of trust in the belief μ allowing deviation towards worst-case or best-case with $-\infty \leq \beta \leq \infty$. For the case of $\beta \rightarrow -\infty$ we recover an infinitely pessimistic agent that considers only worst-case scenarios, for $\beta \rightarrow \infty$ an agent that is infinitely optimistic and for $\beta \rightarrow 0$ the Bayesian agent that fully trusts its model.

5.2 Model Uncertainty and Bounded Rationality in MDPs

In this section, we consider a bounded rational agent with model uncertainty in the infinite horizon setting of an MDP. In this case the agent must take into account all future rewards and information costs, thereby optimizing the following free energy objective

$$F^*(s) = \max_{\pi} \text{ext}_{\psi} \lim_{T \rightarrow \infty} \mathbb{E} \sum_{t=0}^{T-1} \gamma^t \left(R_{s_t, a_t}^{s_{t+1}} - \frac{1}{\beta} \log \frac{\psi(\theta_t|a_t, s_t)}{\mu(\theta_t|a_t, s_t)} - \frac{1}{\alpha} \log \frac{\pi(a_t|s_t)}{\rho(a_t|s_t)} \right) \quad (5.6)$$

where the extremum operator ext can be either \max for $\beta > 0$ or \min for $\beta < 0$, $0 < \gamma < 1$ is the discount factor and the expectation \mathbb{E} is over all trajectories $\xi = s_0, a_0, \theta_0, s_1, a_1, \dots, a_{T-1}, \theta_{T-1}, s_T$

with distribution $p(\xi) = \prod_{t=0}^{T-1} \pi(a_t|s_t) \psi(\theta_t|a_t, s_t) T_{\theta_t}(s_{t+1}|a_t, s_t)$. Importantly, this free energy objective satisfies a recursive relation and thereby generalizes Bellman's optimality principle to the case of model uncertainty and bounded rationality. In particular, equation (5.6) fulfills the recursion

$$F^*(s) = \max_{\pi} \operatorname{ext}_{\psi} \mathbb{E}_{\pi(a|s)} \left[-\frac{1}{\alpha} \log \frac{\pi(a|s)}{\rho(a|s)} + \mathbb{E}_{\psi(\theta|a,s)} \left[-\frac{1}{\beta} \log \frac{\psi(\theta|a,s)}{\mu(\theta|a,s)} + \mathbb{E}_{T_{\theta}(s'|a,s)} \left[R_{s,a}^{s'} + \gamma F^*(s') \right] \right] \right]. \quad (5.7)$$

Applying variational calculus and following the same rationale as in the previous sections (Ortega and Braun, 2013), the extremum operators can be eliminated and equation (5.7) can be re-expressed as

$$F^*(s) = \frac{1}{\alpha} \log \mathbb{E}_{\rho(a|s)} \left[\mathbb{E}_{\mu(\theta|a,s)} \left[\exp \left(\beta \mathbb{E}_{T_{\theta}(s'|a,s)} \left[R_{s,a}^{s'} + \gamma F^*(s') \right] \right) \right]^{\frac{\alpha}{\beta}} \right] \quad (5.8)$$

because

$$F^*(s) = \max_{\pi} \mathbb{E}_{\pi(a|s)} \left[\frac{1}{\beta} \log Z_{\beta}(a, s) - \frac{1}{\alpha} \log \frac{\pi(a|s)}{\rho(a|s)} \right] \quad (5.9)$$

$$= \frac{1}{\alpha} \log \mathbb{E}_{\rho(a|s)} \left[\exp \left(\frac{\alpha}{\beta} \log Z_{\beta}(a, s) \right) \right], \quad (5.10)$$

where

$$\begin{aligned} Z_{\beta}(a, s) &= \operatorname{ext}_{\psi} \mathbb{E}_{\psi(\theta|a,s)} \left[\mathbb{E}_{T_{\theta}(s'|a,s)} \left[R_{s,a}^{s'} + \gamma F^*(s') \right] - \frac{1}{\beta} \log \frac{\psi(\theta|a,s)}{\mu(\theta|a,s)} \right] \\ &= \mathbb{E}_{\mu(\theta|a,s)} \exp \left(\beta \mathbb{E}_{T_{\theta}(s'|a,s)} \left[R_{s,a}^{s'} + \gamma F^*(s') \right] \right) \end{aligned} \quad (5.11)$$

with the optimizing arguments

$$\begin{aligned} \psi^*(\theta|a, s) &= \frac{1}{Z_{\beta}(a, s)} \mu(\theta|a, s) \exp \left(\beta \mathbb{E}_{T_{\theta}(s'|a,s)} \left[R_{s,a}^{s'} + \gamma F^*(s') \right] \right) \\ \pi^*(a|s) &= \frac{1}{Z_{\alpha}(s)} \rho(a|s) \exp \left(\frac{\alpha}{\beta} \log Z_{\beta}(a, s) \right) \end{aligned} \quad (5.12)$$

and partition sum

$$Z_{\alpha}(s) = \mathbb{E}_{\rho(a|s)} \left[\exp \left(\frac{\alpha}{\beta} \log Z_{\beta}(a, s) \right) \right].$$

With this free energy we can model a range of different agents for different α and β . For example, by setting $\alpha \rightarrow \infty$ and $\beta \rightarrow 0$ we can recover a Bayesian MDP planner and by setting $\alpha \rightarrow \infty$ and $\beta \rightarrow -\infty$ we recover a robust planner. Additionally, for $\alpha \rightarrow \infty$ and when $\mu(\theta|a, s) = \delta(\theta - \theta^*(a, s))$ we recover an agent with standard value function with known state transition model from equation (5.1).

5.2.1 Free Energy Iteration Algorithm

Solving the self-consistency equation (5.8) can be achieved by a generalized version of value iteration. Accordingly, the optimal solution can be obtained by initializing the free energy at some arbitrary value F and applying a value iteration scheme $B^{i+1}F = BB^iF$ where we define the operator

$$BF(s) = \max_{\pi} \text{ext}_{\psi} \mathbb{E}_{\pi(a|s)} \left[-\frac{1}{\alpha} \log \frac{\pi(a|s)}{\rho(a|s)} + \mathbb{E}_{\psi(\theta|a,s)} \left[-\frac{1}{\beta} \log \frac{\psi(\theta|a,s)}{\mu(\theta|a,s)} + \mathbb{E}_{T_{\theta}(s'|a,s)} \left[R_{s,a}^{s'} + \gamma F(s') \right] \right] \right] \quad (5.13)$$

with $B^1F = BF$, which can be simplified to

$$BF(s) = \frac{1}{\alpha} \log \mathbb{E}_{\rho(a|s)} \left[\mathbb{E}_{\mu(\theta|a,s)} \left[\exp \left(\beta \mathbb{E}_{T_{\theta}(s'|a,s)} \left[R_{s,a}^{s'} + \gamma F(s') \right] \right) \right]^{\frac{\alpha}{\beta}} \right]$$

In Algorithm (1) we show the pseudo-code of this generalized value iteration scheme. Given state-dependent prior policies $\rho(a|s)$ and the Bayesian posterior beliefs $\mu(\theta|a, s)$ and the values of α and β , the algorithm outputs the equilibrium distributions for the action probabilities $\pi(a|s)$, the biased beliefs $\psi(\theta|a, s)$ and estimates of the free energy value function $F^*(s)$. The iteration is run until a convergence criterion is met. The convergence proof is shown in the next section.

Algorithm 1: Iterative algorithm solving the self-consistency equation (5.8)

Input: $\rho(a|s), \mu(\theta|a, s), \alpha, \beta$
Initialize: $F \leftarrow 0, F_{\text{old}} \leftarrow 0$
while not converged do
 forall $s \in \mathcal{S}$ **do**
 $F(s) \leftarrow \frac{1}{\alpha} \log \mathbb{E}_{\rho(a|s)} \left[\mathbb{E}_{\mu(\theta|a,s)} \left[\exp \left(\beta \mathbb{E}_{T_{\theta}(s'|a,s)} \left[R_{s,a}^{s'} + \gamma F_{\text{old}}(s') \right] \right) \right]^{\frac{\alpha}{\beta}} \right]$
 end
 $F_{\text{old}} \leftarrow F$
end
 $\pi(a|s) \leftarrow \frac{1}{Z_{\alpha}(a,s)} \rho(a|s) \exp \left(\frac{\alpha}{\beta} \log Z_{\beta}(a, s) \right)$
 $\psi(\theta|a, s) \leftarrow \frac{1}{Z_{\beta}(a,s)} \mu(\theta|a, s) \exp \left(\beta \mathbb{E}_{T_{\theta}(s'|a,s)} \left[R_{s,a}^{s'} + \gamma F(s') \right] \right)$
return $\pi(a|s), \psi(\theta|a, s), F(s)$

5.3 Convergence

Here, we show that the value iteration scheme described through Algorithm 1 converges to a unique fixed point satisfying Equation (5.8). To this end, we first prove the existence of a unique fixed point (Theorem 5.3.1) following (Bertsekas and Tsitsiklis, 1996; Rubin et al., 2012), and subsequently prove the convergence of the value iteration scheme presupposing that a unique fixed point exists (Theorem 5.3.2) following (Strehl et al., 2009).

Theorem 5.3.1. Assuming a bounded reward function $R_{s,a}^{s'}$, the optimal free-energy vector $F^*(s)$ is a unique fixed point of Bellman's equation $F^* = BF^*$, where the mapping $B : \mathbb{R}^{|S|} \rightarrow \mathbb{R}^{|S|}$ is defined as in equation (5.13)

Proof. Theorem 5.3.1 is proven through Proposition 5.3.1 and 5.3.2 in the following. \square

Proposition 5.3.1. The mapping $T_{\pi,\psi} : \mathbb{R}^{|S|} \rightarrow \mathbb{R}^{|S|}$

$$T_{\pi,\psi}F(s) = \mathbb{E}_{\pi(a|s)} \left[-\frac{1}{\alpha} \log \frac{\pi(a|s)}{\rho(a|s)} + \mathbb{E}_{\psi(\theta|a,s)} \left[-\frac{1}{\beta} \log \frac{\psi(\theta|a,s)}{\mu(\theta|a,s)} + \mathbb{E}_{T_\theta(s'|a,s)} \left[R_{s,a}^{s'} + \gamma F(s') \right] \right] \right]. \quad (5.14)$$

converges to a unique solution for every policy-belief-pair (π, ψ) independent of the initial free-energy vector $F(s)$.

Proof. By introducing the matrix $P_{\pi,\psi}(s, s')$ and the vector $g_{\pi,\psi}(s)$ as

$$P_{\pi,\psi}(s, s') := \mathbb{E}_{\pi(a|s)} \left[\mathbb{E}_{\psi(\theta|a,s)} [T_\theta(s'|a, s)] \right],$$

$$g_{\pi,\psi}(s) := \mathbb{E}_{\pi(a|s)} \left[\mathbb{E}_{\psi(\theta|a,s)} \left[\mathbb{E}_{T_\theta(s'|a,s)} [R_{s,a}^{s'}] - \frac{1}{\beta} \log \frac{\psi(\theta|a,s)}{\mu(\theta|a,s)} \right] - \frac{1}{\alpha} \log \frac{\pi(a|s)}{\rho(a|s)} \right],$$

Equation (5.14) may be expressed in compact form: $T_{\pi,\psi}F = g_{\pi,\psi} + \gamma P_{\pi,\psi}F$. By applying the mapping $T_{\pi,\psi}$ an infinite number of times on an initial free-energy vector F , the free-energy vector $F_{\pi,\psi}$ of the policy-belief-pair (π, ψ) is obtained:

$$F_{\pi,\psi} := \lim_{i \rightarrow \infty} T_{\pi,\psi}^i F = \lim_{i \rightarrow \infty} \sum_{t=0}^{i-1} \gamma^t P_{\pi,\psi}^t g_{\pi,\psi} + \underbrace{\lim_{i \rightarrow \infty} \gamma^i P_{\pi,\psi}^i F}_{\rightarrow 0},$$

which does no longer depend on the initial F . It is straightforward to show that the quantity

$F_{\pi,\psi}$ is a fixed point of the operator $T_{\pi,\psi}$:

$$\begin{aligned}
T_{\pi,\psi}F_{\pi,\psi} &= g_{\pi,\psi} + \gamma P_{\pi,\psi} \lim_{i \rightarrow \infty} \sum_{t=0}^{i-1} \gamma^t P_{\pi,\psi}^t g_{\pi,\psi} \\
&= \gamma^0 P_{\pi,\psi}^0 g_{\pi,\psi} + \lim_{i \rightarrow \infty} \sum_{t=1}^i \gamma^t P_{\pi,\psi}^t g_{\pi,\psi} \\
&= \lim_{i \rightarrow \infty} \sum_{t=0}^{i-1} \gamma^t P_{\pi,\psi}^t g_{\pi,\psi} + \underbrace{\lim_{i \rightarrow \infty} \gamma^i P_{\pi,\psi}^i g_{\pi,\psi}}_{\rightarrow 0} = F_{\pi,\psi}.
\end{aligned}$$

Furthermore, $F_{\pi,\psi}$ is unique. Assume for this purpose an arbitrary fixed point F' such that $T_{\pi,\psi}F' = F'$, then $F' = \lim_{i \rightarrow \infty} T_{\pi,\psi}^i F' = F_{\pi,\psi}$. \square

Proposition 5.3.2. The optimal free-energy vector $F^* = \max_{\pi} \text{ext}_{\psi} F_{\pi,\psi}$ is a unique fixed point of Bellman's equation $F^* = BF^*$.

Proof. The proof consists of two parts where we assume $\text{ext} = \max$ in the first part and $\text{ext} = \min$ in the second part respectively. Let $\text{ext} = \max$ and $F^* = F_{\pi^*,\psi^*}$, where (π^*, ψ^*) denotes the optimal policy-belief-pair. Then

$$F^* = T_{\pi^*,\psi^*}F^* \leq \underbrace{\max_{\pi} \max_{\psi} T_{\pi,\psi}F^*}_{=BF^*} =: T_{\pi',\psi'}F^* \stackrel{\text{Induction}}{\leq} F_{\pi',\psi'},$$

where the last inequality can be straightforwardly proven by induction and exploiting the fact that $P_{\pi,\psi}(s, s') \in [0; 1]$. But by definition $F^* = \max_{\pi} \max_{\psi} F_{\pi,\psi} \geq F_{\pi',\psi'}$, hence $F^* = F_{\pi',\psi'}$ and therefore $F^* = BF^*$. Furthermore, F^* is unique. Assume for this purpose an arbitrary fixed point $F' = F_{\pi',\psi'}$ such that $F' = BF'$ with the corresponding policy-belief-pair (π', ψ') . Then

$$F^* = T_{\pi^*,\psi^*}F^* \geq T_{\pi',\psi'}F^* \stackrel{\text{Induction}}{\geq} F_{\pi',\psi'} = F',$$

and similarly $F' \geq F^*$, hence $F' = F^*$.

Let $\text{ext} = \min$ and $F^* = F_{\pi^*,\psi^*}$. By taking a closer look at Equation (5.13), it can be seen that the optimization over ψ does not depend on π . Then

$$F^* = T_{\pi^*,\psi^*}F^* \geq \min_{\psi} T_{\pi^*,\psi}F^* =: T_{\pi^*,\psi'}F^* \stackrel{\text{Induction}}{\geq} F_{\pi^*,\psi'}.$$

But by definition $F^* = \min_{\psi} F_{\pi^*,\psi} \leq F_{\pi^*,\psi'}$, hence $F^* = F_{\pi^*,\psi'}$. Therefore it holds that $BF^* = \max_{\pi} \min_{\psi} T_{\pi,\psi}F^* = \max_{\pi} T_{\pi,\psi^*}F^*$ and similar to the first part of the proof we obtain

$$F^* = T_{\pi^*,\psi^*}F^* \leq \underbrace{\max_{\pi} T_{\pi,\psi^*}F^*}_{=BF^*} =: T_{\pi',\psi^*}F^* \stackrel{\text{Induction}}{\leq} F_{\pi',\psi^*}.$$

But by definition $F^* = \max_{\pi} F_{\pi, \psi^*} \geq F_{\pi', \psi^*}$, hence $F^* = F_{\pi', \psi^*}$ and therefore $F^* = BF^*$. Furthermore, F_{π^*, ψ^*} is unique. Assume for this purpose an arbitrary fixed point $F' = F_{\pi', \psi'}$ such that $F' = BF'$. Then

$$F' = T_{\pi', \psi'} F' \leq T_{\pi', \psi^*} F' \stackrel{\text{Induction}}{\leq} F_{\pi', \psi^*} \stackrel{\text{Induction}}{\leq} T_{\pi', \psi^*} F^* \leq T_{\pi^*, \psi^*} F^* = F^*,$$

and similarly $F^* \leq F'$, hence $F^* = F'$. \square

Theorem 5.3.2. Let ϵ be a positive number satisfying $\epsilon < \frac{\eta}{1-\gamma}$ where $\gamma \in (0; 1)$ is the discount factor and where u and l are the bounds of the reward function $R_{s,a}^{s'}$ such that $l \leq R_{s,a}^{s'} \leq u$ and $\eta = \max\{|u|, |l|\}$. Suppose that the value iteration scheme from Algorithm 1 is run for $i = \lceil \log_{\gamma} \frac{\epsilon(1-\gamma)}{\eta} \rceil$ iterations with an initial free-energy vector $F(s) = 0$ for all s . Then, it holds that $\max_s |F^*(s) - B^i F(s)| \leq \epsilon$, where F^* refers to the unique fixed point from Theorem 5.3.1.

Proof. We start the proof by showing that the L_{∞} -norm of the difference vector between the optimal free-energy F^* and $B^i F$ exponentially decreases with the number of iterations i :

$$\begin{aligned} \max_s |F^*(s) - B^i F(s)| &=: |F^*(s^*) - B^i F(s^*)| \\ &\stackrel{\text{Eq. (5.9)}}{=} \left| \max_{\pi} \mathbb{E}_{\pi(a|s^*)} \left[\frac{1}{\beta} \log Z_{\beta}(a, s^*) - \frac{1}{\alpha} \log \frac{\pi(a|s^*)}{\rho(a|s^*)} \right] \right. \\ &\quad \left. - \max_{\pi} \mathbb{E}_{\pi(a|s^*)} \left[\frac{1}{\beta} \log Z_{\beta}^i(a, s^*) - \frac{1}{\alpha} \log \frac{\pi(a|s^*)}{\rho(a|s^*)} \right] \right| \\ &\leq \max_{\pi} \left| \mathbb{E}_{\pi(a|s^*)} \left[\frac{1}{\beta} \log Z_{\beta}(a, s^*) - \frac{1}{\beta} \log Z_{\beta}^i(a, s^*) \right] \right| \\ &\leq \max_a \left| \frac{1}{\beta} \log Z_{\beta}(a, s^*) - \frac{1}{\beta} \log Z_{\beta}^i(a, s^*) \right| \\ &=: \left| \frac{1}{\beta} \log Z_{\beta}(a^*, s^*) - \frac{1}{\beta} \log Z_{\beta}^i(a^*, s^*) \right| \\ &\stackrel{\text{Eq. (5.11)}}{=} \left| \text{ext}_{\psi} \mathbb{E}_{\psi(\theta|a^*, s^*)} \left[\mathbb{E}_{T_{\theta}(s'|a^*, s^*)} [R_{s,a}^{s'} + \gamma F^*(s')] \right] - \frac{1}{\beta} \log \frac{\psi(\theta|a^*, s^*)}{\mu(\theta|a^*, s^*)} \right| \\ &\quad - \left| \text{ext}_{\psi} \mathbb{E}_{\psi(\theta|a^*, s^*)} \left[\mathbb{E}_{T_{\theta}(s'|a^*, s^*)} [R_{s,a}^{s'} + \gamma B^{i-1} F(s')] \right] - \frac{1}{\beta} \log \frac{\psi(\theta|a^*, s^*)}{\mu(\theta|a^*, s^*)} \right| \\ &\leq \max_{\psi} \left| \mathbb{E}_{\psi(\theta|a^*, s^*)} \left[\mathbb{E}_{T_{\theta}(s'|a^*, s^*)} [\gamma F^*(s') - \gamma B^{i-1} F(s')] \right] \right| \\ &\leq \gamma \max_s |F^*(s) - B^{i-1} F(s)| \stackrel{\text{Recur.}}{\leq} \gamma^i \max_s |F^*(s) - F(s)| \leq \gamma^i \frac{\eta}{1-\gamma}, \end{aligned}$$

where we exploit the fact that $|\text{ext}_x f(x) - \text{ext}_x g(x)| \leq \max_x |f(x) - g(x)|$ and that the free-energy is bounded through the reward bounds l and u with $\eta = \max\{|u|, |l|\}$. For a convergence criterion $\epsilon > 0$ such that $\epsilon \geq \gamma^i \frac{\eta}{1-\gamma}$, it then holds that $i \geq \log_{\gamma} \frac{\epsilon(1-\gamma)}{\eta}$ presupposing that $\epsilon < \frac{\eta}{1-\gamma}$. \square

5.4 Experiments: Grid World

This section illustrates the proposed value iteration scheme with an intuitive example where an agent has to navigate through a grid-world. The agent starts at position $\mathbf{S} \in \mathcal{S}$ with the objective to reach the goal state $\mathbf{G} \in \mathcal{S}$ and can choose one out of maximally four possible actions $a \in \{\uparrow, \rightarrow, \downarrow, \leftarrow\}$ in each time-step. Along the way, the agent can encounter regular tiles (actions move the agent deterministically one step in the desired direction), walls that are represented as *gray tiles* (actions that move the agent towards the wall are not possible), holes that are represented as *black tiles* (moving into the hole causes a negative reward) and *chance tiles* that are illustrated as white tiles with a question mark (the transition probabilities of the chance tiles are unknown to the agent). Reaching the goal \mathbf{G} yields a reward $R = +1$ whereas stepping into a hole results in a negative reward $R = -1$. In both cases the agent is subsequently teleported back to the starting position \mathbf{S} . Transitions to regular tiles have a small negative reward of $R = -0.01$. When stepping onto a chance tile, the agent is pushed stochastically to an adjacent tile giving a reward as mentioned above. The true state-transition probabilities of the chance tiles are not known by the agent, but the agent holds the Bayesian belief

$$\mu(\boldsymbol{\theta}_{s,a}|a, s) = \text{Dirichlet}(\Phi_{s,a}^{s'_1}, \dots, \Phi_{s,a}^{s'_{N(s)}}) = \prod_{i=1}^{N(s)} (\theta_{s,a}^{s'_i})^{\Phi_{s,a}^{s'_i} - 1}$$

where transition model is denoted as $T_{\boldsymbol{\theta}_{s,a}}(s'|s, a) = \theta_{s,a}^{s'}$ and $\boldsymbol{\theta}_{s,a} = (\theta_{s,a}^{s'_1} \dots \theta_{s,a}^{s'_{N(s)}})$ and $N(s)$ is the number of possible actions in state s . The data is incorporated into the model as a count vector $(\Phi_{s,a}^{s'_1}, \dots, \Phi_{s,a}^{s'_{N(s)}})$ where $\Phi_{s,a}^{s'}$ represents the number of times that the transition (s, a, s') has occurred. The prior $\rho(a|s)$ for the actions at every state is set to be uniform. An important aspect of the model is that in the case of unlimited observational data, the agent will plan with the correct transition probabilities.

We conducted two experiments with discount factor $\gamma = 0.9$ and uniform priors $\rho(a|s)$ for the action variables. In the first experiment, we explore and illustrate the agent’s planning behavior under different degrees of computational limitations (by varying α) and under different model uncertainty attitudes (by varying β) with fixed uniform beliefs $\mu(\theta|a, s)$. In the second experiment, the agent is allowed to update its beliefs $\mu(\theta|a, s)$ and use the updated model to re-plan its strategy.

5.4.1 The Role of the Parameters α and β on Planning

Figure 5.1 shows the solution to the variational free energy problem that is obtained by iteration until convergence according to Algorithm 1 under different values of α and β . In particular, the first row shows the free energy function $F^*(s)$ (Eq. (5.8)). The second, third and fourth row show heat maps of the position of an agent that follows the optimal policy (Eq. (5.12)) according to the agent’s biased beliefs (plan) and to the actual transition probabilities in a friendly and unfriendly environment, respectively. In chance tiles, the most likely transitions in these two environments are indicated by arrows where the agent is teleported

with a probability of 0.999 into the tile indicated by the arrow and with a probability of 0.001 to a random other adjacent tile.

In the first column of Fig. 5.1 it can be seen that a stochastic agent ($\alpha = 3.0$) with high model uncertainty and optimistic attitude ($\beta = 400$) has a strong preference for the broad corridor in the bottom by assuming favorable transitions for the unknown chance tiles. This way the agent also avoids the narrow corridors that are unsafe due to the stochasticity of the low- α policy. In the second column of Fig. 5.1 with low $\alpha = 3$ and high model uncertainty with pessimistic attitude $\beta = -400$, the agent strongly prefers the upper broad corridor because unfavorable transitions are assumed for the chance tiles. The third column of Fig. 5.1 shows a very pessimistic agent ($\beta = -400$) with high precision ($\alpha = 11$) that allow the agent to safely choose the shortest distance by selecting the upper narrow corridor without risking any tiles with unknown transitions. The fourth column of Fig. 5.1 shows a very optimistic agent ($\beta = 400$) with high precision. In this case the agent chooses the shortest distance by selecting the bottom narrow corridor that includes two chance tiles with unknown transition.

5.4.2 Updating the Bayesian Posterior μ with Observations from the Environment

Similar to model identification adaptive controllers that perform system identification while the system is running (Åström and Wittenmark, 2013), we can use the proposed planning algorithm also in a reinforcement learning setup by updating the Bayesian beliefs about the MDP while executing always the first action and replanning in the next time step. During the learning phase, the exploration is governed by both factors α and β , but each factor has a different influence. In particular, lower α -values will cause more exploration due to the inherent stochasticity in the agent’s action selection, similar to an ϵ -greedy policy. If α is kept fixed through time, this will of course also imply a “suboptimal” (i.e. bounded optimal) policy in the long run. In contrast, the parameter β governs exploration of states with unknown transition-probabilities more directly and will not have an impact on the agent’s performance in the limit, where sufficient data has eliminated model uncertainty. We illustrate this with simulations in a grid-world environment where the agent is allowed to update its beliefs $\mu(\theta|a, s)$ over the state-transitions every time it enters a chance tile and receives observation data acquired through interaction with the environment—compare left panels in Figure 5.2. In each step, the agent can then use the updated belief-models for planning the next action.

Figure 5.2 (right panels) shows the number of data points acquired (each time a chance tile is visited) and the average reward depending on the number of steps that the agent has interacted with the environment. The panels show several different cases: while keeping $\alpha = 12.0$ fixed we test $\beta = (0.2, 5.0, 20.0)$ and while keeping $\beta = 0.2$ fixed we test $\alpha = (5.0, 8.0, 12.0)$. It can be seen that lower α leads to better exploration, but it can also lead to lower performance in the long run—see for example rightmost bottom panel. In contrast, optimistic β values can also induce high levels of exploration with the added advantage that in the limit no performance detriment is introduced. However, high β values can in general

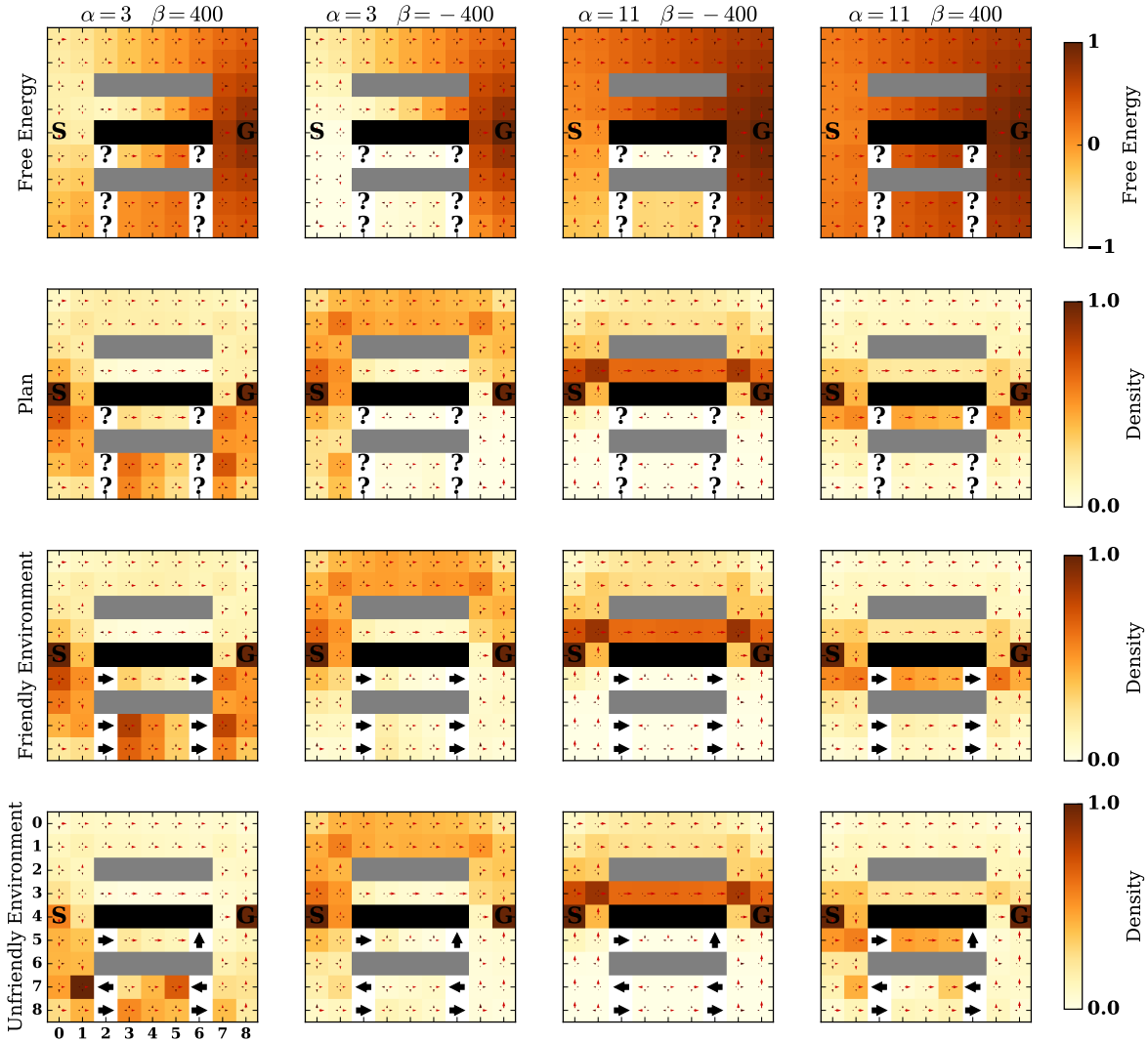


Figure 5.1. The four different rows show free energy values and heat-maps of planned trajectories according to the agent’s beliefs over state-transitions in chance tiles, heat-maps of real trajectories in a friendly environment and in an unfriendly environment respectively. The Start-position is indicated by **S** and the goal state is indicated by **G**. Black tiles represent holes with negative reward, gray tiles represent walls and chance tiles with a question mark have transition probabilities unknown to the agent. The white tiles with an arrow represent the most probable state-transition in chance tiles (as specified by the environment). Very small arrows in each cell encode the policy $\pi(a|s)$ (the length of each arrow encodes the probability of the corresponding action under the policy, highest probability action is indicated as a red arrow). The heat map is constructed by normalizing the number of visits for each state over 20000 steps, where actions are sampled from the agent’s policy and state-transitions are sampled according to one of three ways: the second row according to the agent’s belief over state-transitions $\psi(\theta|a, s)$, in the third and fourth row according to the actual transition probabilities of a friendly and an unfriendly environment respectively. Different columns show different α and β cases.

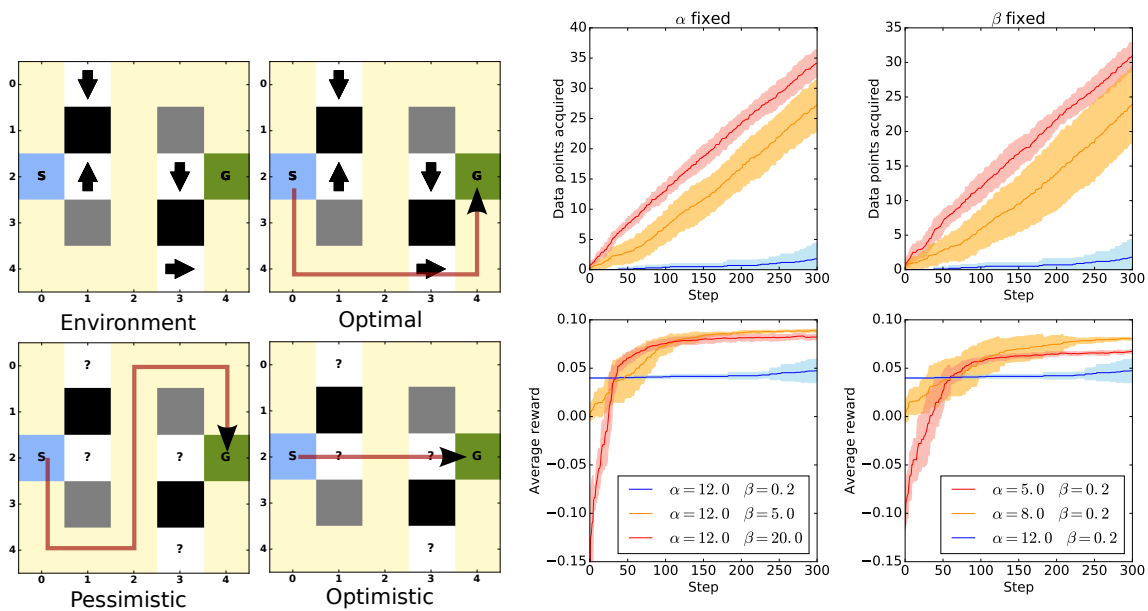


Figure 5.2. The effect of α and β when updating beliefs over 300 interaction steps with the environment. The four panels on the left show the grid-world environment and the pertaining optimal policy if the environment is known. The lower left panels show paths that the agent could take depending on its attitude towards model uncertainty. The panels on the right show the number of acquired data points, that is the number of times a chance tile is entered, and the average reward (bottom panels) for fixed α (varying β) or fixed β (varying α). The average reward at each step is computed as follows. Each time the agent observes a state-transition in a chance tile and updates its belief model, 10 runs of length 2000 steps are sampled (using the agent’s current belief model). The average reward (bold lines) and standard-deviation (shaded areas) across these 10 runs are shown in the figure.

also lead to a detrimental persistence with bad policies, as can be seen for example in the superiority of the low- β agent at the very beginning of the learning process.

5.5 Discussion and Conclusions

In this paper we are bringing two strands of research together, namely research on information-theoretic principles of control and decision-making and robustness principles for planning under model uncertainty. We have devised a unified recursion principle that extends previous generalizations of Bellman’s optimality equation and we have shown how to solve this recursion with an iterative scheme that is guaranteed to converge to a unique optimum. In simulations we could demonstrate how such a combination of information-theoretic policy and belief constraints that reflect model uncertainty can be beneficial for agents that act in partially unknown environments.

Most of the research on robust MDPs does not consider information-processing constraints on the policy, but only considers the uncertainty in the transition probabilities by specifying a set of permissible models such that worst-case scenarios can be computed in order to obtain a robust policy (Iyengar, 2005; Nilim and El Ghaoui, 2005). Recent extensions of these approaches include more general assumptions regarding the set properties of the permissible models and assumptions regarding the data generation process (Wiesemann et al., 2013). Our approach falls inside this class of robustness methods that use a restricted set of permissible models, because we extremize the biased belief $\psi(\theta|a, s)$ under the constraint that it has to be within some information bounds measured by the Kullback-Leibler divergence from a reference Bayesian posterior. Contrary to these previous methods, our approach additionally considers robustness arising from the stochasticity in the policy.

Information-processing constraints on the policy in MDPs have been previously considered in a number of studies (Kappen, 2005a; Peters et al., 2010; Rubin et al., 2012; Todorov, 2009), however not in the context of model uncertainty. In these studies a free energy value recursion is derived when restricting the class of policies through the Kullback-Leibler divergence and when disregarding separate information-processing constraints on observations. However, a small number of studies has considered information-processing constraints both for actions and observations. For example, Polani and Tishby (Tishby and Polani, 2011) and Ortega and Braun (Ortega and Braun, 2013) combine both kinds of information costs. The first cost formalizes an information-processing cost in the policy and the second cost constrains uncertainty arising from the state transitions directly (but crucially not the uncertainty in the latent variables). In both information-processing constraints the cost is determined as a Kullback-Leibler divergence with respect to a reference distribution. Specifically, the reference distribution in (Tishby and Polani, 2011) is given by the marginal distributions (which is equivalent to a rate distortion problem) and in (Ortega and Braun, 2013) is given by fixed priors. The Kullback-Leibler divergence costs for the observations in these cases essentially correspond to a risk-sensitive objective. While there is a relation between risk-sensitive and robust MDPs (Chow et al., 2015; Osogami, 2012; Shen et al., 2014), the innovation in our

approach is at least twofold. First, it allows combining information-processing constraints on the policy with model uncertainty (as formalized by a latent variable). Second, it provides a natural setup to study learning.

The algorithm presented here and Bayesian models in general (Duff, 2002) are computationally expensive as they have to compute possibly high-dimensional integrals depending on the number of allowed transitions for action-state pairs. However, there have been tremendous efforts in solving unknown MDPs efficiently, especially by sampling methods (Guez et al., 2012; Guez et al., 2013; Ross et al., 2011). An interesting future direction to extend our methodology would therefore be to develop a sampling-based version of Algorithm 1 to increase the range of applicability and scalability (Ortega et al., 2014). Moreover, such sampling methods might allow for reinforcement learning applications, for example by estimating free energies through TD-learning (Fox et al., 2015), or by Thompson sampling approaches (Ortega and Braun, 2010a; Ortega and Braun, 2010b) or other stochastic methods for adaptive control (Åström and Wittenmark, 2013).

Acknowledgments

This study was supported by the DFG, Emmy Noether grant BR4164/1-1. The code was developed on top of the RLPy library (Geramifard et al., 2015).

Chapter 6

Non-equilibrium Relations for Bounded Rational Decision-making in Changing Environments

This chapter is based on work *under review*.

Abstract

Living organisms from single cells to humans need to adapt and plan continuously to respond to changes in their environment. The process of behavioral adaptation and planning can be thought of as improving decision-making performance according to some utility function. Here we consider an abstract model of organisms as decision-makers with limited information-processing resources that trade off between maximization of utility and computational costs measured by a relative entropy, in a similar fashion to thermodynamic systems undergoing isothermal transformations. Such systems minimize the free energy to reach equilibrium states that balance internal energy and entropic cost. When there is a fast change in the environment these systems evolve in a non-equilibrium fashion because they are unable to follow the path of equilibrium distributions. Here we apply concepts from non-equilibrium thermodynamics to characterize decision-makers that adapt and plan in changing environments. This allows to quantify performance loss due to imperfect adaptation and planning in a general manner and, additionally, to find relations for decision-making similar to Crooks' fluctuation theorem and Jarzynski's equality. We provide simulations of several exemplary decision and inference problems in the discrete and continuous domains to illustrate the new relations.

Introduction

A number of recent studies have pointed out mathematical equivalences between thermodynamical systems described by statistical mechanics and information processing systems (Ortega and Braun, 2013; Tishby and Polani, 2011; D. H. Wolpert, 2006). In particular, it has been suggested that decision-makers with constrained information-processing resources can be described in analogy to closed physical systems in contact with a heat bath that seek to minimize energy (Ortega and Braun, 2013). In this analogy, decision-makers can be thought to act in a way that minimizes a cost function or, equivalently, maximizes a utility function in lieu of an energy function. Classic decision theory states that, given a set of actions \mathcal{X} and a set of observations \mathcal{O} , the perfectly rational decision maker should choose the best possible action $x^* \in \mathcal{X}$ that maximizes the expected utility $U(x)$ (Savage, 1954; Von Neumann and Morgenstern, 1944)

$$x^* = \operatorname{argmax}_x U(x) = \operatorname{argmax}_x \sum_{o \in \mathcal{O}} p(o|x)V(o), \quad (6.1)$$

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

Deviations from rational decision-making due to limited computational resources have been studied in the field of bounded rationality, originally propagated by Herbert Simon (Simon, 1955; Simon, 1979). A bounded-rational decision-maker is unable to choose the best possible action x^* due to a lack of computational resources. There are a number of different approaches to model bounded rational decision-making. Russell and colleagues have for example suggested a definition of bounded optimality that implies searching for the program that achieves the best utility performance on a particular machine (S. Russell, 1995; S. J. Russell and Subramanian, 1995). This definition has also spurred further work in cognitive science—see (Howes et al., 2009) for a review. Other approaches explicitly append computational costs to the utility function and optimize the total utility that includes the cost of reasoning (Horvitz, 1988). Decision-makers that explicitly reason about the cost of reasoning are said to perform meta-reasoning, for example by using anytime algorithms (Dean and Boddy, 1988; Zilberstein, 1996) that can be interrupted at any time deemed fit by the meta-reasoning level. In psychology, bounded rationality models have focused on describing deviations from expected utility theory in actual choice behavior by humans by relying both on optimality models such as prospect theory and heuristic approaches (Camerer, 2003; Gigerenzer and Goldstein, 1996; Daniel Kahneman, 2003).

Recently, new impulses for the development of bounded rationality theory have come from information-theoretic and thermodynamic perspectives on the general organization of perception-action-systems (Braun and Ortega, 2014; Braun et al., 2011b; Friston, 2010; Kappen et al., 2012; Ortega and Braun, 2011; Ortega and Braun, 2013; Rubin et al., 2012; Still, 2009; Still et al., 2012; Tishby and Polani, 2011; Todorov, 2009; Vijayakumar et al., 2012). In the economic and game-theoretic literature, these models have precursors that have

studied bounded rationality inspired by stochastic choice rules originally proposed by Luce, McFadden and others (Fudenberg and Levine, 1998; Luce, 1959; Mattsson and Jörgen W. Weibull, 2002; McFadden, 1980; McKelvey and Palfrey, 1995; Meginnis, 1976; Sims, 2003; D. H. Wolpert, 2006). In most of these models decision-makers face a trade-off between the attainment of maximum utility and the required information-processing cost measured as an entropy or relative entropy. The optimal solution to this trade-off usually takes the form of a Boltzmann-like distribution analogous to equilibrium distributions in statistical physics. The decision-making process can then be conceptualized as a change from a prior strategy distribution to a posterior strategy distribution, where the change is triggered by a change in the utility landscape. However, studying changes in equilibrium distributions neglects not only the time required for this change, but also the adaptation process itself.

The main contribution of this paper is to show that the analogy between equilibrium thermodynamics and bounded-rational decision-making (Ortega and Braun, 2013) can be extended to the non-equilibrium domain to provide new predictions that can be tested in experimental setups. The connection between the non-equilibrium and equilibrium domains is tied with the concept of dissipation and the derivation of a fluctuation theorem and a Jarzynski-like equality for decision-making, which are important recent results in non-equilibrium thermodynamics. The paper is organized as follows. In Section 6.1 we recapitulate the relation between bounded rational decision-making and equilibrium thermodynamics. In Section 6.2 we extend the relation to non-equilibrium processes and we include a derivation of the Jarzynski equality and Crooks' fluctuation theorem for decision-making. In Section 6.3 we provide simulations to illustrate the new relations in different decision-making scenarios. In Section 6.4 we discuss our results.

6.1 Equilibrium Thermodynamics and Decision-Making

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

On an abstract level, one can think about changes in the distribution of a random variable from a prior to a posterior distribution as the basis of information-processing. In Bayesian inference, for example, we update current prior beliefs $p_0(x)$ by means of a likelihood to obtain a posterior belief $p_1(x)$. Similarly, decision-making can be regarded as a process of changing a prior strategy $p_0(x)$ to a posterior strategy $p_1(x)$ by means of computation. With infinite computational resources the decision-maker retrieves the best action x^* with certainty.

Conversely, without any computational resources the best strategy is to simply stick to the prior strategy $p_1(x) = p_0(x)$. According to (Ortega and Braun, 2013) such decision-making behavior with limited resources can be formalized by optimizing the variational problem

$$p_1^{\text{eq}}(x) = \underset{p}{\operatorname{argmax}} \Delta F[p] \quad (6.2)$$

where

$$\Delta F[p] := \sum_x p(x) \Delta U(x) - \frac{1}{\beta} D_{\text{KL}}(p||p_0), \quad (6.3)$$

is a free energy functional, $\Delta U(x)$ is a change in utility function, the Kullback-Leibler divergence or relative entropy $D_{\text{KL}}(p||p_0) = \sum_x p(x) \log \frac{p(x)}{p_0(x)}$ quantifies the “information distance” between prior and posterior, and β can be interpreted as a resource or boundedness parameter. In contrast to the expected utility objective from Equation (6.1) this variational problem is different in three ways. First, the maximizing argument is not over an action x but over the distribution over actions $p(x)$ emphasizing the stochastic nature of the final strategy (Rieskamp, 2008). Second, instead of absolute utilities it considers changes in utility based on the current state similar to the notion of gains and losses in prospect theory (Daniel Kahneman, 2003). Third, it optimizes a trade-off between utility gains and computation costs, where the costs are quantified by a relative entropy term. In a physical system, Equation (6.3) evaluated at the optimum p_1^{eq} quantifies the (negative) *free energy* difference $\Delta F[p_1^{\text{eq}}]$ between the final state 1 and the initial state 0 assuming an isothermal process and a (negative) *energy* difference of $\Delta U = U_1 - U_0$.

The bounded rational decision-maker can be determined according to Equation (6.2) following the strategy

$$p_1^{\text{eq}}(x) = \frac{1}{Z_\beta} p_0(x) e^{\beta \Delta U(x)} \quad (6.4)$$

with partition function $Z_\beta = \sum_x p_0(x) e^{\beta \Delta U(x)}$. When inserting the optimal strategy $p_1^{\text{eq}}(x)$ into Equation (6.3), the certainty-equivalent value of strategy p_1^{eq} is determined by

$$\Delta F^{\text{eq}} := \Delta F[p_1^{\text{eq}}] = \frac{1}{\beta} \log Z_\beta. \quad (6.5)$$

For infinite resources ($\beta \rightarrow \infty$) the optimal strategy $p_1^{\text{eq}}(x)$ places all the probability mass on the maximum of $\Delta U(x)$ and the value of the strategy is $\lim_{\beta \rightarrow \infty} \Delta F[p_1^{\text{eq}}] = \max_x \Delta U(x)$. This models a perfectly rational decision-maker that can handpick the best action. For $\beta \rightarrow 0$ the cost of computation dominates and the optimal strategy is given by the prior strategy $p_1^{\text{eq}}(x) = p_0(x)$ with the value $\lim_{\beta \rightarrow 0} \Delta F[p_1^{\text{eq}}] = \langle \Delta U(x) \rangle_{p_0(x)}$. This models a decision-maker bare of computational resources. In the following we discuss a simple algorithm that acts according to the posterior distribution using a given prior.

An exemplary bounded rational decision-maker

The optimal distribution from Equation (6.4) can be implemented, for example, by a decision-maker that follows a probabilistic *satisficing* strategy with aspiration level $T \geq \max_x \Delta U(x)$.

Such a decision-maker optimizes the utility $\Delta U(x)$ by drawing samples from the prior distribution $x_s \sim p_0(x)$ and accepts with certainty the first sample x_s with utility $\Delta U(x_s) \geq T$ above the aspiration level T or any sample with utility below the aspiration level with acceptance probability $p_{\text{accept}} = \exp(\beta(\Delta U(x_s) - T))$. This particular version of the more general rejection sampling algorithm is shown in pseudo-code in Algorithm 2.

We can see the direct connection between β and computational resources when computing the average number of samples required until acceptance. It can be shown that the expected number of required samples from p_0 to obtain one accepted sample from p_1^{eq} is given by $\bar{n}_\beta = \exp(\beta T)/Z_\beta \geq \exp D_{\text{KL}}(p||p_0)$ (Ortega and Braun, 2014). In the no-resource limit $\beta \rightarrow 0$, the equilibrium posterior distribution is equal to the prior and the sampling complexity tends to $\bar{n}_{\beta \rightarrow 0} \rightarrow 1$. In the full-resource limit $\beta \rightarrow \infty$ the sampling complexity increases according to $\bar{n}_{\beta \rightarrow \infty} \rightarrow \exp(\beta T - \beta \Delta U(x_{\text{max}}))/p_0(x_{\text{max}})$ where $x_{\text{max}} = \text{argmax}_x \Delta U(x)$. Thus the higher the β , the higher the precision but also the higher the amount of resources spent, i.e. the amount of samples that will be required until acceptance. If the decision-maker's aspiration level is lower than the maximum utility, the same framework can be applied under a redefined utility function $\Delta V(x) = \min\{\Delta U(x), T\}$.

Algorithm 2: Rejection Sampling Algorithm

```

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

```

Characterization of the decision-maker's boundedness

In the previous example, the boundedness of the decision-maker consists in the fact that the average number of samples the decision-maker can afford during deliberation is limited. Implicitly, the available average number of samples determines the value of β in the acceptance step. The problem here is that it is not trivial to compute β from \bar{n}_β . Alternatively, we could imagine that we characterize the decision-maker by a fixed precision β that will then result in a particular average number of samples. While this is easy to implement, the problem here is that we do not know in advance the time that will be needed until acceptance. On a more abstract level the two characterizations of boundedness can be seen as follows.

Bounded rational decision-makers with fixed bit rate. In the first interpretation, the parameter β acts as a constraint on the maximum amount of information that can be processed in a certain time window Δt . In other words, β adopts the meaning of a Lagrange

multiplier for the constraint in the following optimization problem

$$\begin{aligned} & \operatorname{argmax}_p \quad \sum_x p(x) \Delta U(x) \\ & \text{subject to} \quad D_{\text{KL}}(p||p_0) \leq M. \end{aligned}$$

This corresponds to processing information at a fixed bit rate of $r = M/\Delta t$. In this case, the value of β is a function of M , $p_0(x)$ and $\Delta U(x)$. In such a scenario with fixed bit rate r the resource parameter β is different for different utility functions. In an economic interpretation β reflects the shadow price.

Example: Decision-maker XYZ with prior $p_0(x)$ has a bit rate r bits/s and time Δt seconds to optimize the utility $\Delta U_t(x)$. What is the bounded rational choice?

Answer: The precision $\beta(M)$ can be determined from the maximum KL-distance $M = r\Delta t$ as a Lagrange multiplier in Equation (6.4), such that the optimal strategy is given by

$$p_t^{\text{opt}}(x) = \frac{1}{Z} p_0(x) e^{\beta(M) \Delta U(x)}.$$

This decision-maker achieves a higher utility than any other decision-maker that chooses a different permissible precision $\beta \leq \beta(M)$. However, selecting the optimal β requires knowing the utility function in advance because in order to compute the KL-divergence we need to compute the posterior that depends on the utility function. While the bit rate might be a useful characterization of some systems (e.g. a von Neumann computer with fixed clock), it might be less useful for others (e.g. a particle undergoing a diffusion process).

Bounded rational decision-makers with fixed precision. In the second interpretation, β acts as a fixed precision parameter that quantifies by how much two different levels of utilities can be told apart. The precision β translates utilities into units of information (nats or bits), such that with increasing β the magnitude of the informational difference $\Delta I := -\log p_1^{\text{eq}}(x_2) + \log p_1^{\text{eq}}(x_1) = \beta (\Delta U(x_1) - \Delta U(x_2)) + \log p_0(x_1)/p_0(x_2)$ between two different utility levels increases, that is they become more distinguishable. In such scenario when the precision β is fixed the bit rate $r = D_{\text{KL}}(p||p_0)/\Delta t$ is different for different utility functions. In an economic interpretation β reflects a fixed market price.

Example: Decision-maker XYZ with prior $p_0(x)$ has to pay a price of β bits/utile when optimizing the utility $\Delta U_t(x)$. What is the bounded rational choice?

Answer: The optimal choice is

$$p_t^{\text{opt}}(x) = \frac{1}{Z} p_0(x) e^{\beta^* \Delta U(x)}.$$

However, while in the fixed bit rate case we know how much time it is needed, here this different problem formulation with fixed precision does not directly relate to time, that is, it is unknown how long it takes to go from $p_0(x)$ to $p(x)$.

For a single fixed utility function $\Delta U(x)$ both characterizations of a bounded rational decision-maker (with fixed precision and with fixed bit-rate) are equivalent and can be mapped into each other. However, when considering changing utility functions, the question arises which of the two features (β or M) stays invariant across time and therefore best characterizes the decision-maker. In the remainder of the paper we study the second characterization of bounded rational decision-making with invariant precision β in relation to temporal processes.

6.2 Non-equilibrium Thermodynamics and Decision-Making

The problem of relating bounded rational decision-making with fixed precision to the time domain can be illustrated with the aforementioned rejection sampling scheme. Even though the precision β implies an *expected* number of samples \bar{n}_β and we can assume that generating each sample requires time δt , it is unknown whether after a particular time Δt^* a sample has been accepted or not. We only know the average time for acceptance. In order to make sure that a sample is accepted we would have to allow for more time. In particular, to know with certainty, we would have to allow for an infinite amount of time. A finite time answer would require a computational process that can be interrupted at any time to deliver an answer, i.e. an anytime process model. We begin our study of non-equilibrium decision-making with an anytime process example.

Example. In an anytime version of rejection sampling, the decision-maker is allowed a particular time Δt^* to produce a sample x_s . This limits the number of samples that can be drawn to a maximum k —see pseudocode in Algorithm 3. The probability of *not* accepting a sample after k tries is given by

$$q_k = \left(1 - \frac{Z(\beta)}{\exp(\beta T)}\right)^k.$$

In this case the sample x_s will be distributed according the prior distribution $p_0(x)$. The probability of accepting a sample that is distributed according to $p_1^{\text{eq}}(x)$ after k tries is given by $1 - q_k$. Accordingly, the action at time k is a mixture distribution of the form

$$p_k^{\text{neq}}(x) = (1 - q_k)p_1^{\text{eq}}(x) + q_k p_0(x). \quad (6.6)$$

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

6.2.1 Non-equilibrium Thermodynamics

In thermodynamics, non-equilibrium processes are often modeled in the presence of an external parameter $\lambda(t) \in [0; 1]$ that determines how the energy function $E_\lambda(x)$ changes over time—for example, when switching on a potential in a linear fashion, the energy would be $E_\lambda(x) =$

Algorithm 3: Rejection Sampling Algorithm with fixed number of samples

```

for  $i = 1 \dots k$  do
   $x \sim p_0$ 
   $u \sim \text{Uniform}[0, 1]$ 
  if  $u \leq \exp(\alpha(\Delta U(x) - T))$  then accepted, return  $x$ ;
end
return  $x$ 

```

$E_0(x) + \lambda(E_1(x) - E_0(x))$. When the change in the parameter λ is done infinitely slowly (quasi-statically), the system's probability distribution follows exactly the path of equilibrium distributions (for any λ) $p_\lambda(x) = \frac{1}{Z_\lambda} e^{-\beta E_\lambda(x)}$. Importantly, when the switching of the external parameter λ is done in finite time, the trajectory in phase space of the evolving thermodynamic system can potentially be very different from the quasi-static case. In particular, the non-equilibrium path of probability distributions is going to be, in general, different from the equilibrium path. We define the trajectory of an evolving system as a finite sequence of states $\mathbf{x} := (x_0, x_1, \dots, x_N)$ at times t_0, t_1, \dots, t_N , and the probability of the trajectory as $p(\mathbf{x}) := p(x_0|t_0) \prod_{n=1}^N p(x_n|x_{n-1}, t_n)$ that follows Markovian dynamics. Since λ is then a function of time $\lambda(t_n)$, we can effectively consider the energy as a function of state and time $E(x_n, t_n) := E_{\lambda(t_n)}(x_n)$. Accordingly, the internal energy of the system can change in two ways depending on changes in the two variables t_n and x_n . Assuming discrete time steps, an energy change due to a change in the external parameter is defined as the work

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

and an energy change due to an internal state change is defined as the heat

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

In an entire process x_0, x_1, \dots, x_N measured at times t_0, t_1, \dots, t_N the extracted work is $W(\mathbf{x}) = -\sum_{n=1}^N w(x_{n-1}, t_{n-1} \rightarrow t_n)$ and the heat transferred to the environment by relaxation steps is $Q(\mathbf{x}) = -\sum_{n=1}^N q(x_{n-1} \rightarrow x_n, t_n)$. The sum of work and heat is the total energy difference $\Delta E(\mathbf{x}) := -(E(x_N, t_N) - E(x_0, t_0)) = W(\mathbf{x}) + Q(\mathbf{x})$. In expectation with respect to $p(\mathbf{x})$ we define the average work $W := \langle W(\mathbf{x}) \rangle_{p(\mathbf{x})}$, the average heat $Q := \langle Q(\mathbf{x}) \rangle_{p(\mathbf{x})}$ and the average energy change $\Delta E := \langle \Delta E(\mathbf{x}) \rangle_{p(\mathbf{x})}$. With these averaged quantities we obtain *the first law of thermodynamics* in its usual form

$$\begin{aligned} \Delta E &= W + Q \\ &= W + T\Delta S + W^{\text{diss}} \end{aligned} \tag{6.7}$$

with the temperature T , the entropy difference $\Delta S = -(S(t_N) - S(t_0))$ and the average dissipation W^{diss} . The entropy flow ΔS captures the reversible entropy exchange with the environment, whereas the dissipation captures the irreversible entropy change. By identifying

the equilibrium free energy difference with $\Delta F := -(F(t_N) - F(t_0)) = \Delta E - T\Delta S$, we can then write the first law as

$$W = \Delta F - W^{\text{diss}}. \quad (6.8)$$

In case of a quasi-static process the extracted work W exactly coincides with the equilibrium free energy difference (thus $W^{\text{diss}} = 0$) that is trajectory independent. In case of a finite time process we can express the average dissipated work as (Gomez-Marin et al., 2008; Christopher Jarzynski, 2011; Roldán, 2014)

$$W^{\text{diss}} := \left\langle W^{\text{diss}}(\mathbf{x}) \right\rangle_{p(\mathbf{x})} = \Delta F - W = \frac{1}{\beta} D_{\text{KL}}(p(\mathbf{x}) || p^\dagger(\mathbf{x})) \quad (6.9)$$

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

$$W \leq \Delta F. \quad (6.10)$$

Crooks' Fluctuation Theorem. Equation (6.9) can be given in a more general form without averages. It is possible to relate the reversibility of a process with its dissipation at the trajectory level. Given a protocol $\Lambda = (\lambda_0, \lambda_1, \dots, \lambda_N)$ i.e. a sequence of external parameters, the probability $p(\mathbf{x})$ of observing a trajectory of the system in phase space compared with its time-reversal conjugate $p^\dagger(\mathbf{x})$ (when using the time-reversal protocol $\Lambda^\dagger = (\lambda_N, \lambda_{N-1}, \dots, \lambda_0)$) depends on the dissipation of the trajectory in the forward direction according to the following expression

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

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

Jarzynski equality. Additionally, another relation of interest in non-equilibrium thermodynamics has recently been found transforming the inequality of Equation 6.10 into an equality, the so-called Jarzynski equality (C. Jarzynski, 1997)

$$\left\langle e^{\beta W(\mathbf{x})} \right\rangle_{p(\mathbf{x})} = e^{\beta \Delta F} \quad (6.11)$$

where the angle brackets denote an average over all possible trajectories \mathbf{x} of a process that drives the system from an equilibrium state at $\lambda = 0$ to another state at $\lambda = 1$. Specifically,

the above equality says that, no matter how the driving process is implemented, we can determine equilibrium quantities from work fluctuations in the non-equilibrium process. Or in other words, this equality connects non-equilibrium thermodynamics with equilibrium thermodynamics. In the following, we are interested in the question whether there exist similar relations such as the Jarzynski equality or Crooks' fluctuation theorem and similar underlying concepts such as dissipation and time reversibility for the case of decision-making.

6.2.2 Non-equilibrium Decision-Making

In contrast to physical processes where the system is forced to react to its environment by a slow adaptation process, intelligent decision-makers might be able to anticipate changes in the environment due to prediction or very fast planning processes that happen on a much shorter time scale than the occurrence of changes in the environment. However, such planning or prediction processes are expensive and we assume in the following that the Kullback-Leibler divergence is an appropriate measure of this computational expense, as outlined in the introduction.

In the following we consider decision-makers facing a sequence of decision problems expressed by the utility functions $\Delta U_1(x), \dots, \Delta U_n(x), \dots, \Delta U_N(x)$ with $\Delta U_n(x) := \Delta U(x, t_{n-1} \rightarrow t_n)$. In particular, we distinguish decision-makers that plan and decision-makers that do not plan. Decision-makers that do not plan have to act before realizing the change in utility

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

and choose action x_{n-1} when faced with ΔU_n . In contrast, decision-makers that plan can consider the change in utility

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

in their action x_n when faced with ΔU_n . We describe the decision-maker's behavior by a vector $\mathbf{x} := (x_0, \dots, x_N)$ and the probability of the trajectory as $p(\mathbf{x}) := p(x_0|t_0) \prod_{n=1}^N p(x_n|x_{n-1}, t_n)$ with $p(x_0|t_0) = p_0(x_0|t_0)$, assuming that the initial strategy is a bounded rational equilibrium strategy. Note that in the no-planning scenario the last decision x_N can be ignored, as it does not contribute to the utility, and similarly in the planning scenario, x_0 does not constitute a decision.

In Figure 6.1 we illustrate the difference between the two scenarios in an exemplary one-step decision problem $\Delta U(x, t_0 \rightarrow t_1)$ (with behavior vector $\mathbf{x} = (x_0, x_1)$). An instantaneous change in the environment occurs at time t_0 represented by a vertical jump from λ_0 to λ_1 in the upper panels that translates directly in a change in free energy difference represented by ΔF in the lower panels. The system's previous state at t_0 is given by $p_0(x)$ i.e. the equilibrium distribution for U_0 . The new equilibrium is given by $p_1(x)$ i.e. the equilibrium distribution for U_1 . In the no-planning scenario the utility $\mathcal{U}^{\text{net}} = \sum_x p_0(x_0) \Delta U(x_0, t_0 \rightarrow t_1)$ extracted from the system is exactly given at t_0 and the dissipation is $\mathcal{U}^{\text{diss}} = \Delta F - \mathcal{U}^{\text{net}}$. In the planning scenario, the utility is given just after the deliberation time at $t_1 = t_0 + \Delta t^*$.

During deliberation the decision-maker has changed the strategy distribution from $p_0(x_0)$ to a non-equilibrium distribution $\tilde{p}(x_1)$ (that corresponds to Equation (6.6) for the rejection sampling scheme) spending in the process a certain amount of resources and achieving an average net utility of $\mathcal{U}^{\text{net}} = \Delta F[\tilde{p}(x)]$ (compare Equation (6.3)). In such a scenario with a single decision-problem we define, in analogy with non-equilibrium thermodynamics, the average dissipated utility as (Gaveau and Schulman, 1997; Still et al., 2012)

$$\begin{aligned}\mathcal{U}^{\text{diss}} &:= \Delta F - \mathcal{U}^{\text{net}} \\ &= \frac{1}{\beta} D_{\text{KL}}(\tilde{p}(x) || p_1^{\text{eq}}(x)).\end{aligned}\tag{6.12}$$

where it easily follows from the positivity of the relative entropy $D_{\text{KL}}(p||q) \geq 0$ that

$$\mathcal{U}^{\text{net}} \leq \Delta F\tag{6.13}$$

with equality when $\tilde{p}(x) = p_1^{\text{eq}}(x)$ corresponding to an infinite amount of available samples with $k \rightarrow \infty$. This inequality shows that we cannot obtain more utility than the equilibrium free energy difference.

6.2.2.1 Non-equilibrium decision-making without planning

With no planning capabilities the decision-maker faces an instant utility switch ΔU_n at each time point and chooses the action x_{n-1} according to the strategy $p(x_{n-1}|x_{n-2}, t_{n-1})$ that is lagging behind the utility changes. However, once the action is chosen the decision-maker can adapt its behavior to the experienced utility $\Delta U_n(x_{n-1})$ before the utility changes again in the next time point. This adaptation corresponds to a physical relaxation process and implies a strategy change between x_{n-1} and x_n . The utility $\Delta U_n(x_{n-1})$ gained by the decision-maker at time point t_{n-1}

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

parallels the concept of work in physics. For a whole trajectory we define the total utility gain due to changes in the environment as $\mathcal{U}(\mathbf{x}) = \sum_{n=1}^N \Delta U(x_{n-1}, t_{n-1} \rightarrow t_n)$. Similarly to Equation (6.8), the *first law for decision-making*

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

then states that the total average utility $\mathcal{U} := \langle \mathcal{U}(\mathbf{x}) \rangle_{p(\mathbf{x})}$ is the difference between the bounded optimal utility (following the equilibrium strategy with precision β) expressed by the equilibrium free energy difference ΔF and the dissipated utility $\mathcal{U}^{\text{diss}}$. The dissipation for a trajectory $\mathcal{U}^{\text{diss}}(\mathbf{x}) := \Delta F - \mathcal{U}(\mathbf{x})$ measures the amount of utility loss due to the inability of the decision-maker to act according to the equilibrium distribution for each decision problem. This is because the decision-maker is forced to act without planning and cannot anticipate the changes in the environment. At most, the decision-maker could act according to the

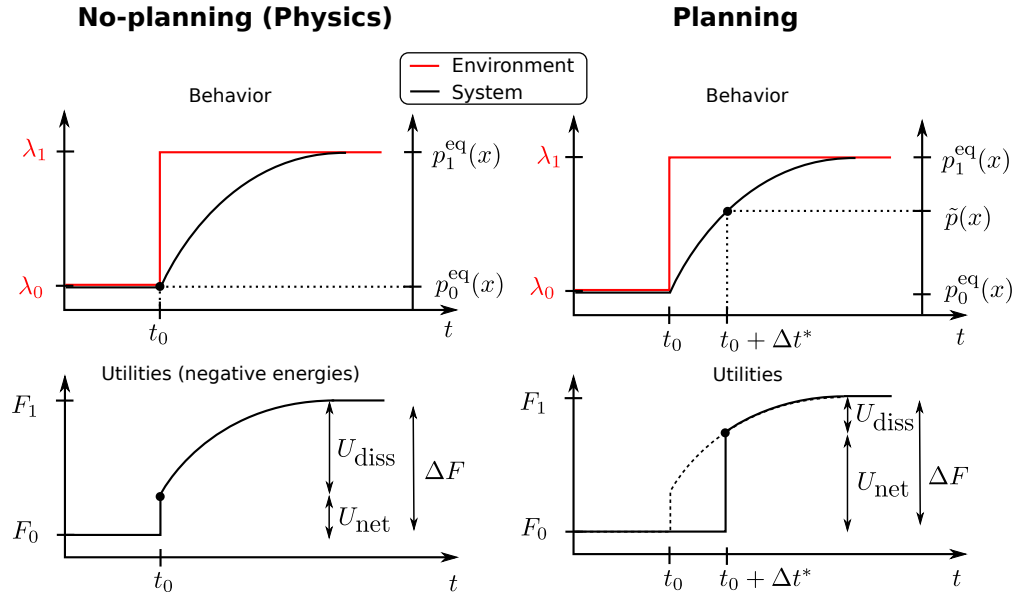


Figure 6.1. Difference between no-planning and planning scenarios in one-step decision problem. An instantaneous change in the environment occurs at time t_0 represented by a vertical jump from λ_0 to λ_1 in the upper panels that translates directly in a change in free energy difference represented by ΔF in the lower panels. The system's previous state at t_0 is given by $p_0^{\text{eq}}(x)$ i.e. the equilibrium distribution for U_{λ_0} , and the posterior equilibrium is given by $p_1^{\text{eq}}(x)$ i.e. the equilibrium distribution for U_{λ_1} . **Left:** No-planning. In the no-planning scenario the decision-maker is unable to plan and therefore acts according to the previous strategy behavior $p_0^{\text{eq}}(x)$ at time t_0 . On average with such strategy the utility gained is $\mathcal{U}^{\text{net}} = \sum_x p_0(x_0) \Delta U(x)$ at t_0 and the dissipation is $\mathcal{U}^{\text{diss}} = \Delta F - \mathcal{U}^{\text{net}}$. **Right:** In the planning scenario, the decision-maker is able to compute a better strategy $\tilde{p}(x)$ after deliberation at time $t_1 = t_0 + \Delta t^*$. In this case the net utility is $\mathcal{U}^{\text{net}} = \sum_x \tilde{p}(x) \Delta U - \frac{1}{\beta} D_{\text{KL}}(\tilde{p}(x) || p_0(x))$.

equilibrium distributions of the previous environment. Thus even with full adaptation the decision-maker will always lag behind one time-step and will therefore always dissipate.

Due to an equivalent version of Equation (6.9), we can also state the *second law for decision-making* $\mathcal{U}^{\text{diss}} \geq 0$, which implies that a purely adaptive decision-maker can gain a maximum utility that cannot be larger than the free energy difference

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

Similarly we can obtain equivalent relationships to the Crooks-Fluctuation Theorem

$$\frac{p(\mathbf{x})}{p^\dagger(\mathbf{x})} = e^{\beta \mathcal{U}^{\text{diss}}(\mathbf{x})}, \quad (6.14)$$

and the Jarzynski Equality

$$\left\langle e^{\beta \mathcal{U}(\mathbf{x})} \right\rangle_{p(\mathbf{x})} = e^{\beta \Delta F} \quad (6.15)$$

which both have the same implications as in the physical scenario and can be derived in the same way as in the physical counterpart (Crooks, 1998). In summary, we can say that an adaptive decision-maker without planning follows the same laws as a thermodynamic physical system that is lagging behind the equilibrium.

6.2.2.2 Non-equilibrium decision-making with planning

In contrast to an agent without planning capabilities, an agent that plans will be able to act according to a different distribution than the prior strategy. This means that when facing the decision problem ΔU_n at time t_n the agent chooses the action x_n sampled from the posterior strategy, contrary to an agent without planning that chooses x_{n-1} sampled from the prior strategy. The planning process incurs a computational cost that is measured—in a similar fashion to stochastic thermodynamics (Seifert, 2005) and previous formulations of bounded rationality given in the introduction—with the difference between the *conditional* stochastic entropies from prior to posterior

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

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

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

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

which again parallels the concept of work. The total net utility $\mathcal{U}^{\text{net}}(\mathbf{x}) = \sum_{n=1}^N u(x_n, t_{n-1} \rightarrow t_n)$ takes the form of a non-equilibrium free energy at trajectory level

$$\mathcal{U}^{\text{net}}(\mathbf{x}) = \sum_{n=1}^N \Delta U(x_n, t_{n-1} \rightarrow t_n) - \frac{1}{\beta} \sum_{n=1}^N \log \frac{p(x_n|x_{n-1}, t_n)}{p(x_n|x_{n-1}, t_{n-1})}. \quad (6.16)$$

Similarly to Equation (6.8), the *first law for decision-making with planning costs* is

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

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

$$\mathcal{U}^{\text{diss}}(\mathbf{x}) := \Delta F - \mathcal{U}^{\text{net}}(\mathbf{x}) \quad (6.17)$$

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

To investigate the counterpart of the second law, we need to determine whether $\mathcal{U}^{\text{diss}} \geq 0$ holds. This can be achieved, for example, by first deriving the counterpart of Crooks fluctuation theorem or the counterpart of the Jarzynski equation with subsequent application of Jensen's inequality.

Theorem 6.2.1. Crook's Fluctuation Theorem for decision-making with planning costs states that

$$\frac{p(\mathbf{x})}{p^\dagger(\mathbf{x})} = e^{\beta \mathcal{U}^{\text{diss}}[\mathbf{x}]} \quad (6.18)$$

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

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

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

where in the second line we have used the identity

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

from local detailed balance, and we assumed that in the backward process the decision-maker starts also using the equilibrium strategy $p^\dagger(x_N|t_N) = \frac{1}{Z_N} e^{\beta U(x_N, t_N)}$. \square

Although at first sight Equation (6.18) looks the same as the previous Crooks' relation for the no-planning case (6.14), it is not the same. Here the net utility is defined by Equation (6.16) which takes into account both, the gain in utility *and* the computational costs of planning.

Theorem 6.2.2. The Jarzynski equality for decision-making with planning costs states that

$$\left\langle e^{\beta \mathcal{U}^{\text{net}}(\mathbf{x})} \right\rangle_{p(\mathbf{x})} = e^{\beta \Delta F} \quad (6.19)$$

Proof.

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

In (1.) we unfold the expression. In (2.), we cancel the trajectory probabilities $\prod_{n=1}^N \tilde{p}(x_n | t_n, x_{n-1})$ and then take one term out of the two remaining products. In (3.) first we use the equivalence $\exp(\beta U(x_1, t_0)) = Z_0 p_{\text{eq}}(x_1 | t_0)$ (because at time t_0 the decision-maker acting according to the equilibrium distribution) that allows to cancel with $\tilde{p}(x_1 | t_0, x_0) = p_{\text{eq}}(x_1 | t_0)$, and second, we sum over x_0 with the only term that depends on it being $p(x_0 | t_0)$. In (4.) we take one term of the second product and perform the sum over x_1 to obtain by detailed balance $\exp(\beta U(x_2, t_1))$ that will allow to cancel with the term in the denominator of the first product. We perform steps (3.) and (4.) repeatedly until obtaining the last equivalence that proves the theorem. \square

Again we note that previously proved Jarzynski relation from Equation (6.19) is not the same equation as in the no-planning case (6.15). In the planning case the definition of the net utility is different and takes into account both, the utility gain and the computational cost of planning.

We can now state the *second law of decision-making with planning costs* as

$$\left\langle \mathcal{U}^{\text{diss}}(\mathbf{x}) \right\rangle_{p(\mathbf{x})} = \frac{1}{\beta} D_{\text{KL}}(p(\mathbf{x}) || p^\dagger(\mathbf{x})) \geq 0 \quad (6.20)$$

from Equation (6.18) by rearranging and taking expectations. The same inequality can be obtained from Equation (6.19) by applying Jensen's inequality $\langle \exp x \rangle \geq \exp \langle x \rangle$ to recover $\langle \mathcal{U}^{\text{net}}(\mathbf{x}) \rangle_{p(\mathbf{x})} \leq \Delta F$. Equation (6.19) connects finite with infinite time decision-making. That is, there is a relation between the equilibrium free-energy differences that is the maximum attainable net utility with unlimited computation time and the net utility obtained by decision-makers with limited computation time. In section 6.3 we will provide examples of how to use these relations to extract useful information from decision-making processes.

6.3 Application to Exemplary Learning and Planning Systems

In this section we illustrate the applicability of our results in a series of simulations for different decision-making scenarios.

- **No-planning.** For the no-planning scenario we study two model classes: the first one contains simple one-step lag models of adaptation where equilibrium is always reached with one time step delay, and the second one contains more complex models of adaptation that do not necessarily equilibrate after one time step. In the first model class we can easily study the relation between dissipation and the rate of information-processing, whereas in the second class of models we can study more complex non-equilibrium phenomena such as learning hysteresis.
- **Planning.** For the planning scenario we illustrate the novel Jarzynski equality and Crooks Theorem for decision-making in two cases: the first case is a discrete decision-making scenario with clearly defined independent episodes, the second case is a continuous planning problem.

The four example sections therefore are (a) No-Planning: dissipation and information-processing rate, (b) No-Planning: dissipation and learning hysteresis, (c) Planning: Jarzynski and Crooks for episodic decision-making, (d) Planning: Jarzynski and Crooks for continuous decision-making.

6.3.1 No-Planning: Dissipation and Information-Processing Rate

Human decision-makers have to make decisions typically under delayed information. For example, when trying to avoid an obstacle while driving, optimal actions are delayed due to a minimum reaction time to notice the obstacle. Here we show how both, idealized information-processing systems with delay and Bayesian inference schemes that can also be seen as delayed systems, are subject to thermodynamic interpretation.

6.3.1.1 One-step lag models of adaptation

Consider a learner that is adapted to their environment such that their behavior can be described by the equilibrium distribution $p_0(x)$. For this idealized scenario we assume that

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

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

in the first time step which results directly from Equation (6.12) when replacing the non-equilibrium distribution $\tilde{p}(x)$ by $p_0(x)$. In all subsequent time steps no more utility is wasted, assuming the environment does not change anymore. In the second case, the utility function can be written as $U_t(x) = U_0(x) + \frac{t}{N} \Delta U(x)$ for $t \in \mathbb{N} : 0 \leq t \leq N$. To compute the dissipated utility we need to compare the learner's behavior in time step t to the bounded optimal behavior which is

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

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

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

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

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

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

6.3.1.2 Bayesian inference as a one-step lag process

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

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

$$p(\theta|\mathcal{D}) = \operatorname{argmax}_{\tilde{p}} \Delta F[\tilde{p}] = \operatorname{argmax}_{\tilde{p}} \int \tilde{p}(\theta|\mathcal{D}) \log p(\mathcal{D}|\theta) d\theta - \frac{1}{\beta} \int \tilde{p}(\theta|\mathcal{D}) \log \frac{\tilde{p}(\theta|\mathcal{D})}{p_0(\theta)} d\theta \quad (6.21)$$

$$= \frac{1}{Z} p_0(\theta) e^{\beta \log p(\mathcal{D}|\theta)} = \frac{1}{Z} p_0(\theta) p(\mathcal{D}|\theta)^\beta \quad (6.22)$$

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

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

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

where the equivalence of dissipation corresponds to

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

when processing all the data at once and to

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

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

In Figure 6.2 we show how number of data chunks have an effect on the overall surprise and dissipation. In particular, we have a dataset $\mathcal{D} = (x_1, \dots, x_T)$ consisting of $T = 100$ data points Gaussian distributed $x \sim \mathcal{N}(x; \mu_d, \sigma_d^2)$ that we divide in batches of different size $b \in \{100, 50, 25, 20, 10, 5, 2, 1\}$. The decision maker has prior belief $p_0(\theta)$ about the mean $\theta = \mu_d$ and incorporates the data of every batch of data according to bayes' rule until all the data is incorporated. For $b = 100$ the Bayesian learner processes all data at once, having thus high surprise and for $b = 1$ it incorporates the data in T/b updates with an overall lesser surprise. In Figure 6.2 we show for different batch sizes the free energy optimum $\Delta F = \log \int p_0(\theta) p(\mathcal{D}|\theta)$, the surprise \mathcal{S} and the dissipation $\mathcal{U}^{\text{diss}} = \Delta F - \mathcal{S}$. It can be seen that when acquiring the data in small chunks the surprise of the decision-maker and the dissipation is lower.

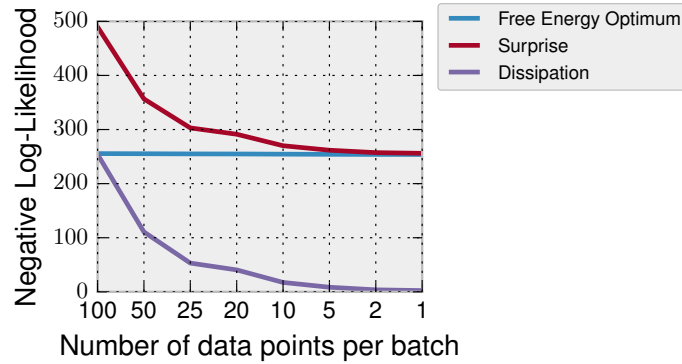


Figure 6.2. Surprise, dissipation and free energy optimum as a function of the number of data points per batch in a Bayesian inference task. We see that when the decision-maker processes all the data at the same time has maximum surprise and dissipation. However, when incorporating the data slowly the surprise and dissipation is minimal. The free energy optimum is only a function of the data independent of how it is incorporated.

6.3.2 No-Planning: Dissipation and Learning Hysteresis

A common paradigm to study how humans learn is through adaptation tasks where subjects are exposed to changes in an environmental variable that they can counteract by changing an internal variable. Sensorimotor adaptation in humans has been extensively studied in these error-based paradigms for example, where subjects have to adapt their hand position (internal variable) to change virtual hand movements (external variable) represented by a dot on a screen. When trying to reach a target, decision-makers must reduce this mismatch by adapting their movements trial by trial in order to reduce errors. After many trials they completely adapted to the mismatch and produce no errors (other than motor noise).

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

Metropolis-Hastings Algorithm. In a decision-theoretic context, the Metropolis-Hastings algorithm (Chib and Greenberg, 1995), can be considered an anytime decision-making process that converges over time to the equilibrium distribution. The Metropolis-Hastings decision-maker uses a Markov chain of states that starts at some initial location $x_0 \sim p_0(x)$ reflecting

the prior of the decision-maker and then adapts its behavior according to the transition probabilities $p(x'|x) = g(x'|x)\alpha(x'|x)$ where $g(x'|x)$ is the proposal probability distribution and

$$\alpha(x'|x) = \min \left(\frac{e^{\beta U(x')} g(x|x')}{e^{\beta U(x)} g(x'|x)}, 1 \right)$$

is the acceptance probability for each decision. In this way the decision-maker proposes a new choice $x' \sim g(x'|x)$ and decides at each time point whether to accept it with probability $\alpha(x'|x)$. After a long time the chain of states converges to the optimal equilibrium distribution.

Crooks Theorem and Hysteresis Effects in Adaptation Tasks

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

In the two left panels of Figure 6.3 we show a simulated trajectory of actions composed of 80 trials for an adaptation task using the Metropolis-Hastings algorithm with $\beta = 22.5$ and a Gaussian proposal $g(x'|x) = \mathcal{N}(x'; \mu = x, \sigma_p = 0.1)$ when changing the environmental variable from $\mu_0 = 0.0$ to $\mu_1 = 1.0$. In blue we show the trajectory for the forward-in-time perturbation which converges after a few dozen trials to the new equilibrium. In brown we show the trajectory for the reversed perturbation where the process starts with the last trial (80) and ends with the initial trial (0). In the left panel the perturbation is made instantaneously in one step at trial 40 and in the right panel in multiple steps ($N = 23$). The hysteresis effect is clearly seen in the instantaneous perturbation where the path of actions followed by the decision-maker in the forward perturbation is clearly different from a typical trajectory of actions taken when applying the reversed perturbation. When the perturbation is made in multiple steps both typical backward and typical forward trajectories become more similar denoting a smaller hysteresis effect. In this way hysteresis effects are tightly connected to the concept of dissipation.

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

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

The probability of observing a trajectory of accepted actions $\mathbf{x} = (x_0, x_1, \dots, x_T)$ for the Metropolis-Hastings algorithm is easily computed with $p(\mathbf{x}) = p(x_0) \prod_{t=1}^T g(x_t|x_{t-1})\alpha(x_t|x_{t-1})$. Similarly, the probability of observing the same trajectory in the backward protocol is $p(\mathbf{x}^\dagger) =$

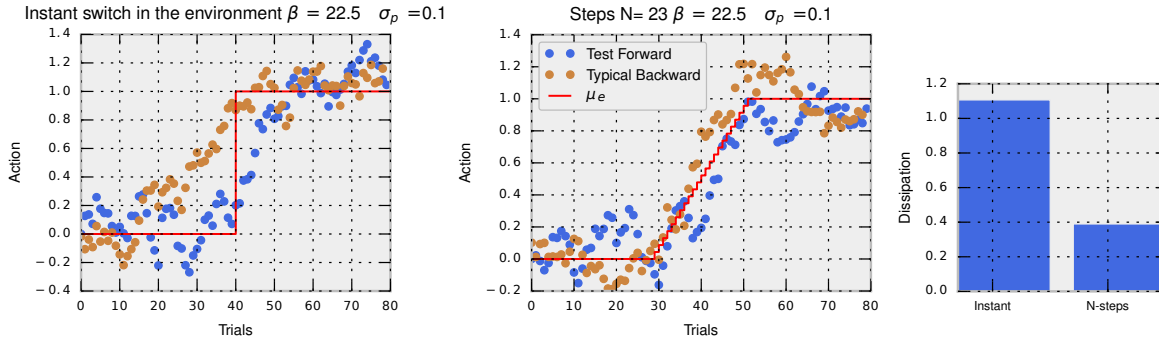


Figure 6.3. We show the trajectories of actions from the Metropolis-Hastings algorithm with $\beta = 22.5$ and proposal standard deviation $\sigma_p = 0.1$ in a forward (blue) or backward (brown) protocol for an instant change in the environment (first panel) and for a slow change in the environment (second panel). In both cases the total change in the environment is $\mu_e = 0$ to $\mu_e = 1$. The last panels shows the dissipation for the forward protocol (blue) in both, the instant or the slow change in the environment. The difference in probability densities of forward and backward trajectories relates directly to dissipation and to hysteresis effects.

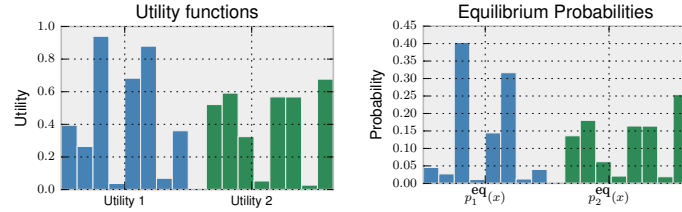
$p_{\text{eq}}(x_T) \prod_{t=1}^T g(x_{T-t}|x_{T-t+1})\alpha(x_{T-t}|x_{T-t+1})$. The dissipated utility is $\mathcal{U}^{\text{diss}} = \Delta F - U_{\text{tot}}$ where the free energy difference is computed between the final $p_1(x) = \frac{1}{Z}e^{\beta U_1(x)}$ and initial equilibrium distributions $p_0(x) = \frac{1}{Z}e^{\beta U_0(x)}$, and the total utility gained U_{tot} is the sum of the utilities $\Delta U(x, t_n \rightarrow t_{n+1})$ at each environmental change at time t_n . In the third panel of Figure 6.3 we show that the protocol with the instantaneous perturbation has higher dissipation (related to higher hysteresis) compared to the protocol with multiple small perturbations.

6.3.3 Planning: Jarzynski and Crooks Relations for Episodic Decision-Making

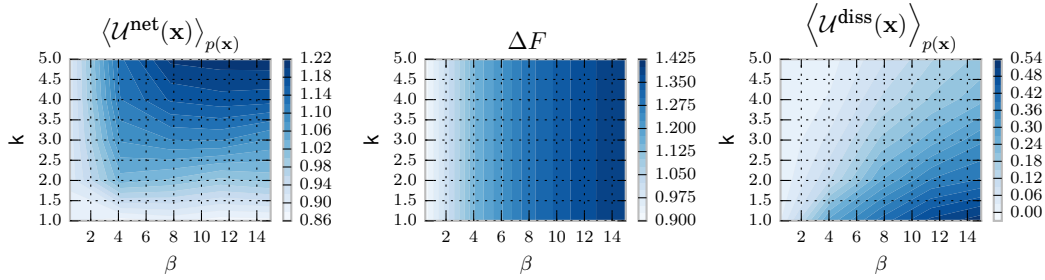
Choice-reaction-time experiments aimed to study information-processing in humans typically consider episodic tasks. Here, we take a variation of Hicks episodic task with discrete action space, commonly used in the decision-making literature. In our variation of Hicks task, the decision-maker is shown a set of eight light bulbs $|\mathcal{X}| = 8$ that are turned off. Then all light bulbs are turned on with different light intensities for a limited amount of time in which the decision-maker must choose the brightest light associated with high utility. When given enough time a decision-maker with arbitrary prior $p_0(x)$ chooses its actions according to the equilibrium distribution from Equation (6.4). In such equation, the precision β specifies the how well the light intensities can be told apart e.g. a human would be constrained by the precision or the density of its photoreceptors. The choice task is repeated N times, each time with different light intensities. For simplicity we set $N = 2$ enough to analyze the non-equilibrium behavior of such decision-maker.

In Figure 6.4A we show the utility values and the corresponding equilibrium distributions for a certain precision $\beta = 4$, reached only when having unlimited time. However, when having

A



B



C

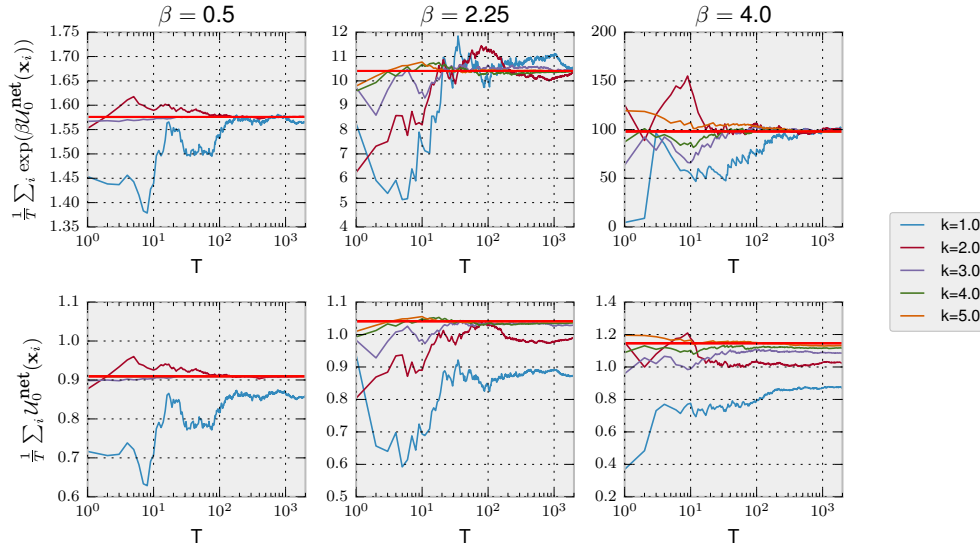


Figure 6.4. Episodic decision-making with planning. **A:** Utility functions and equilibrium distributions for the two decision problems. **B:** We show for different β and k (left) the average net utility, (middle) the free energy difference and (right) the average dissipated utility. **C** Top panels: Convergence of the empirical Jarzynski estimate depending on the number of trajectories T using different β and different number of available samples k . Bottom panels: the associated expected net utility gain which in the limit $T \rightarrow \infty$ is lower than the free energy difference (horizontal light red line).

limited time the equilibrium distribution is not reached and instead the decision-maker acts according to a non-equilibrium distribution. We model this non-equilibrium choice strategy

with the rejection sampling algorithm with limited samples from the introduction, where the non-equilibrium probability distribution are described by Equation (6.6). In this kind of episodic task the decision-maker always starts with the same prior $p_0(x)$ over the possible choices x . The probability of a trajectory of decisions \mathbf{x} is defined as $p(\mathbf{x}) := \prod_{n=1}^N p(x_n|t_n)$ for each episode n , and the net utility for a trajectory is

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

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

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

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

In the first panel of Figure 6.4B we show that, as expected, the more samples k the higher the average net utility $\langle \mathcal{U}_0^{\text{net}} \rangle_{p(\mathbf{x})}$. In the second panel, we show the equilibrium free energy difference invariant with respect to k and increasing for higher precision denoted by β . Lastly, in the third panel we plot the average dissipated utility $\langle \mathcal{U}^{\text{diss}} \rangle_{p(\mathbf{x})}$ that measures how much utility is lost due to the limited number of available samples. The highest dissipation occurs for high β and few samples k because such a high-precision decision-maker can potentially obtain high utility but the limited amount of samples restrain it. In the following we derive both a Jarzynski-like relation and a fluctuation theorem valid for fixed prior.

Jarzynski equality for decision-making with fixed prior p_0 . For fixed prior it is trivial to show that the following relation is valid

$$\left\langle e^{\beta \mathcal{U}_0^{\text{net}}(\mathbf{x})} \right\rangle_{p(\mathbf{x})} = e^{\beta \Delta F}. \quad (6.23)$$

To empirically test the validity of Equation (6.23) we simulated a decision-maker that faces T times the same two decision-problems from Figure 6.4A. We can estimate the left hand side of Equation (6.23) with the empirical average $\frac{1}{T} \sum_i \exp(\beta \mathcal{U}_0^{\text{net}}(\mathbf{x}_i))$ with the T trajectories of decisions, where $\mathbf{x}_i \sim p(\mathbf{x})$. In the top row of Figure 6.4 we show the empirical average converging to $\exp(\beta \Delta F)$ depending on the number of simulated trajectories T and precision β , empirically validating Equation (6.23). In the bottom row we show how the second law for decision-making is fulfilled as the average net utility is less than the equilibrium free energy thus satisfying inequality (6.13).

Equation (6.23) could be tested straightforwardly in an experiment with human subjects. We would make sure that the statistics of the light bulbs induce a uniform prior to the decision-maker, and we would test them in an infinite time condition (where the choice probabilities

should correspond to stable equilibrium distributions allowing to compute in the first place β and in the second the equilibrium free energies ΔF) and in a finite time condition (allowing to compute $\mathcal{U}_0^{\text{net}}$ from the observed non-equilibrium behavior $p(\mathbf{x})$).

Crooks' fluctuation Theorem for decision-making with fixed prior p_0 . For fixed priors it easy to show that the following fluctuation relation holds

$$\frac{\tilde{p}(\mathbf{x})}{p^{\text{eq}}(\mathbf{x})} = e^{\beta(\Delta F - \mathcal{U}_0^{\text{net}}(\mathbf{x}))} = e^{\beta \mathcal{U}^{\text{diss}}(\mathbf{x})} \quad (6.24)$$

where $p^{\text{eq}}(\mathbf{x}) := \prod_{n=1}^N p^{\text{eq}}(x_n|t_n)$ is the optimal equilibrium distribution over trajectories \mathbf{x} . Note in this case the probability distribution of the backward process $p^\dagger(\mathbf{x})$ coincides with the optimal equilibrium distribution $p^\dagger(\mathbf{x}) = p^{\text{eq}}(\mathbf{x})$ because of the independence of the decision-problems. More specifically, the original Crooks theorem for decision-making from Equation (6.18) is valid only when the backward process start in equilibrium. In our episodic task all decision problems are independent which makes the starting equilibrium distributions for all the backward processes coincide with the posterior equilibrium distributions.

The fluctuation relation (6.24) for episodic tasks adopts a different meaning than the conventional relation. Specifically, the ratio between probabilities is now between the probability of observing a trajectory of actions when having finite time to make a decision (a sequence of non-equilibrium probabilities) and the probability of observing the same trajectory when having infinite time (a sequence of equilibrium probabilities). This ratio is governed by the exponential of the dissipated utility $\mathcal{U}^{\text{diss}}(\mathbf{x})$ similarly to the original Crooks equation.

Equation (6.24) can be rewritten by re-arranging the terms and averaging over $p(\mathbf{x})$ as

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

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

6.3.4 Planning: Jarzynski and Crooks Relations for Continuous Decision-Making

Since many decision-tasks are in continuous setups (such as sensorimotor tasks) here we consider such continuous state space problems. In particular, we validate our Jarzynski equation in the continuous domain with a non-episodic task.

In many optimization problems extracting gradient information from the cost function is crucial for the optimization process. Here we use a diffusion process, modeled by Langevin dynamics, that uses gradient information to reach equilibrium. In particular, we will employ

quadratic utility functions that will allow for a close form solution of the non-equilibrium probability density that changes over time.

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

$$\frac{\partial x}{\partial t} = \alpha \frac{\partial U(x)}{\partial x} + \alpha \xi(t) \quad (6.25)$$

where $\xi(t)$ is white Gaussian noise with mean $\langle \xi(t) \rangle = 0$ and correlation $\langle \xi(t)\xi(t') \rangle = 2D\delta(t-t')$, $D = \frac{\sigma^2}{2}$ and σ^2 is the variance in $(t-t')$ time. In physics, when $U(x)$ is an energy function, this equation would correspond to the general equation of motion for Langevin dynamics $m \frac{\partial^2 x}{\partial t^2} = -\frac{m}{\alpha} \frac{\partial x}{\partial t} - \frac{\partial U(x)}{\partial x} + m\xi(t)$ in the limit of strong friction $|\gamma \frac{\partial x}{\partial t}| \gg |m \frac{\partial^2 x}{\partial t^2}|$ and mass $m = 1$. Note that Equation (6.25) is closely related to learning algorithms that use gradient information such as for example Stochastic Gradient Descent (SGD). These algorithms find the minimum of a cost function by taking steps in the state space in the opposite direction of the gradient. In here we see that the learning rate corresponds to the parameter α that, in contrast with plain GD, not only multiplies the gradient but also the noise term.

Equation (6.25) gives the micro-dynamics of the decision-making process, however, the evolution of the macro-dynamics $p(x, t)$ are described by the following (well-known) Fokker-Planck equation (Garcia-Palacios, 2007)

$$\frac{\partial p(x, t)}{\partial t} = -\alpha p(x, t) \frac{\partial^2 U(x)}{\partial x^2} + \alpha \frac{\partial U(x)}{\partial x} \frac{\partial p(x, t)}{\partial x} + D\alpha^2 \frac{\partial^2 p(x, t)}{\partial x^2}. \quad (6.26)$$

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

$$p(x, t) = \frac{1}{\sqrt{2\pi\sigma^2(t)}} e^{-\frac{(x-\mu(t))^2}{2\sigma^2(t)}} \quad (6.27)$$

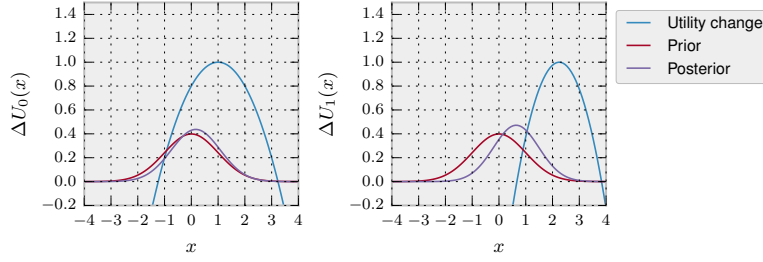
with

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

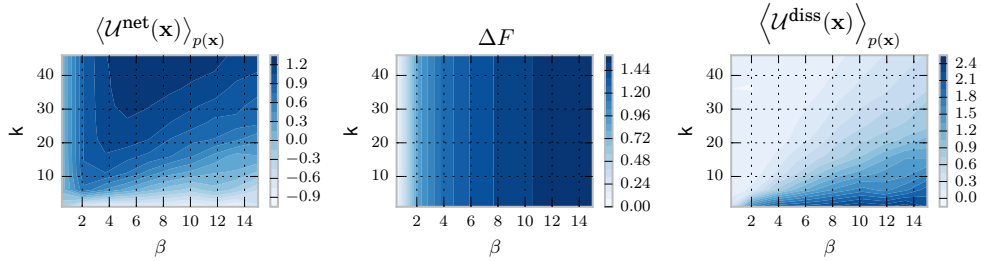
where $c = 2\alpha a_1$, and we assumed that the prior strategy is Gaussian distributed with mean μ_0 and variance σ_0^2 . The precision parameter relates to the other parameters with the relation $\beta = \frac{2\alpha}{D}$, which means that the higher the α the more we take into account the gradient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

Following a similar approach from the previous section we expose a decision-maker to two utility function ΔU_1 and ΔU_2 which are shown in Figure 6.5A. In Figure 6.5B we show the net utility, equilibrium free-energy differences and dissipated utility (according to Equations (6.16)

A



B



C

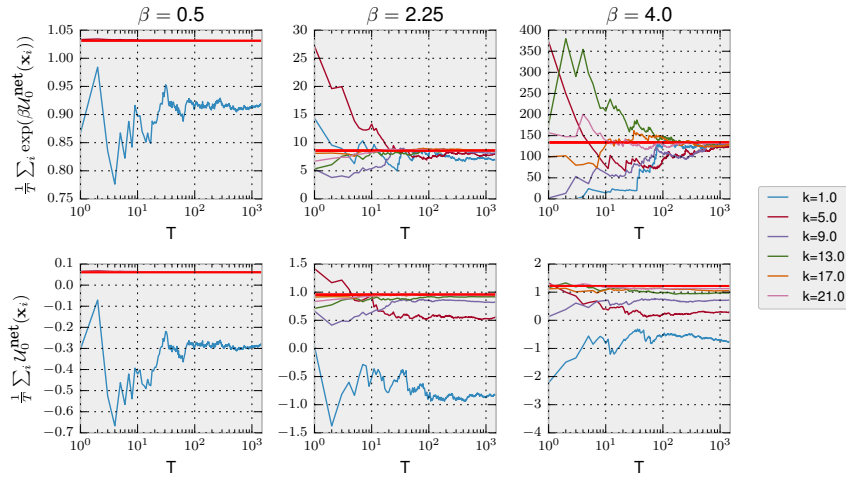


Figure 6.5. Langevin Dynamics simulations. **A** In blue the different utility changes ΔU_1 and ΔU_2 , in red the prior p_0 , and in purple the posterior for $\beta = 0.5$ **B**: We show for different β and time $t = k\Delta t$ directly depending on k , (left) the average net utility, (middle) the free energy difference and (right) the average dissipated utility. **C** Top panels: Convergence of the empirical Jarzynski estimate depending on the number of trajectories T using different β and different number of update steps k in the discrete Langevin equation. Bottom panels: the associated expected net utility gain which in the limit $T \rightarrow \infty$ is lower than the free energy difference (horizontal light red line). With this simulations we validate Equation (6.19).

and (6.17)) for different values of β and number of steps k —corresponding to time $t = k\Delta t$ in Equation (6.27) for a given reference Δt . In Figure 6.5C we show the convergence of the Jarzynski term towards the true equilibrium free energy difference term depending on the number of trajectories to make the estimation. We can see on the bottom row that the second law for decision making represented by the inequality (6.13) is fulfilled.

6.4 Discussion

In this paper we have used both precision and time to describe the behavior of decision-makers with computational limitations in changing environments. We highlighted the similarities with non-equilibrium thermodynamics in the case of agents that plan and agents that do not plan. Additionally, we derived a novel Jarzynski equality and a Crooks fluctuation theorem for decision-making scenarios with planning. We have shown how to use Jarzynski’s and Crooks’ equations in different scenarios to extract relevant variables of the decision-making process such as the equilibrium free energy difference, the average dissipated utility and the action-path probabilities for both, equilibrium posterior distributions and distributions of the backward-in-time protocol. We have provided a few examples for the no-planning and planning scenario, such as one-step lag dynamics, discrete choice tasks and continuous decision-making tasks.

The trade-off between expected utility and computational cost is a fundamental aspect of decision-making that matters not only to the field of artificial intelligence, but also to neuroscience, biology, psychology and cognitive science. Recently, there has been an increasing interest in modeling decision-making with computational constraints (Gershman et al., 2015; Parkes and Wellman, 2015). As pointed out in (Gershman et al., 2015) the computational costs can be various, for example, expenditures arising from delays in time-critical settings or the number of available samples when using Monte Carlo methods. Our modeling assumptions take the relative entropy between a prior and a posterior distribution after deliberation as an average computation cost. As a result of having limited time, decision-makers have an additional sub-optimality which takes the form of a dissipated utility in analogy with non-equilibrium thermodynamics.

The idea of using the relative entropy as a computational cost is not new (Mattsson and Jörgen W. Weibull, 2002; Ortega and Braun, 2010b; D. H. Wolpert, 2006). In (Mattsson and Jörgen W. Weibull, 2002) and similarly in (Ortega and Braun, 2011) the authors derive the relative entropy as a control cost from an information-theoretic point of view, under axioms of monotonicity and invariance under relabeling and decomposition. In other fields such as robotics the relative entropy has also been used as a control cost (Braun et al., 2011b; Kappen et al., 2012; Peters et al., 2010; Todorov, 2009) to regularize the behavior of the controller by penalizing controls that are far from the uncontrolled dynamics of the system. Generally, the relative entropy as a regularizer for utility maximization is convenient and general as it is independent of the parametrization of the probability distributions.

The mutual information has also been considered as a candidate to quantify the compu-

tation costs. We note, however, that the mutual information can be recovered from a free energy when assuming a multi-task scenario. In particular, when assuming a probability distribution $p(w)$ for an environment w , it is possible to show that the mutual information is an average relative entropy (Genewein et al., 2015). This approach to quantify computational resources has been employed in (Rubin et al., 2012; Still and Precup, 2012; Tishby and Polani, 2011) to characterize decision-makers with information constraints. Additionally, theoretical results (Bernardo, 1997; Lindley, 1997) directly relate the connection between the number of samples needed in an experiment to perform inference with the mutual information, by realizing that both have logarithmic form and thus indicating its correspondence to the sample size.

Stochastic optimal control theory is closely related to the sequential decision-problems shown here when considering trajectories as elementary states (Braun et al., 2011b). Its objective is to compute an optimal set of actions that minimize future cost. Classical approaches to solve this problem is through the Bellman equation using a dynamic programming argument where the reward in the present time step also depends on the value function of the next time step. A possible approach to solve such class of problems is through a value iteration scheme as for example in (Rubin et al., 2012). This type of value iteration scheme has been recently extended to the case where the model of the environment is unknown (Grau-Moya et al., 2016a). Another possible approach to solve the Bellman recursion is through path integral control that to consider a specific class of stochastic optimal control problems where the cost of control is also a relative entropy. In path integral control one can simply sample from uncontrolled dynamics to estimate the cost to go and act optimally (Kappen, 2005b; Kappen et al., 2012; Theodorou et al., 2010). This directly connect to our planning scenario when considering trajectories as single actions. In such case the utility function for a trajectory is the sum of utility functions for every step, the prior distribution corresponding to the uncontrolled dynamics is fixed and can be sampled from (e.g. brownian motion), and the trajectories form posterior distribution correspond to the controlled dynamics.

The algorithms presented here exploit randomness to compute optimal solutions by sampling. These algorithms belong to the class of Monte Carlo methods. There is growing evidence that the human brain might exploit also randomness to perform computations. Instantiations of sampling procedures in humans have been shown experimentally in visual perception (Moreno-Bote et al., 2011), sentence processing (R. P. Levy et al., 2009) and inference (Griffiths and Tenenbaum, 2006; Sanborn et al., 2010). Approximate inference with sampling methods have been able to explain some behavioral biases in decision-making (Lieder et al., 2012), and have been shown to be efficient when using few samples (Vul et al., 2014). Evidence that learning and perception are probabilistic and tied together has been found in behavioral and neural experiments (Fiser et al., 2010). From the theoretical point of view sampling procedures have been shown to be compatible with neural architectures that perform such statistical computations (Buesing et al., 2011). In some neuroscientific studies, the brain is seen as a statistical machine that minimizes surprise (Friston, 2010) also by means of a free energy principle. These theoretical insights together with the experimental evidence

of stochastic computations in the brain reinforces the view of the brain as statistical machine that can be subjected to thermodynamic analysis.

The results presented here are novel since only some connections between non-equilibrium thermodynamics and decision-making have been reported in a handful of papers in the literature. For example, regarding the connection between predictive power and dissipation, (Still et al., 2012) has found that non-predictive systems have been shown theoretically to be highly dissipative. This directly connects to our approach in the following way. If the decision-maker cannot plan because it does not know the future utility function it will employ a non-equilibrium strategy that leads to high inefficiency in terms of dissipated utility. However, having full predictive power will allow it to produce actions from the equilibrium distribution which means no dissipation. Jarzynski-like and Crooks-like relations have been found in the economics literature in gambling scenarios (Hirono and Hidaka, 2015) and when studying the arrow of time for decision-making (Mlodinow and Brun, 2014; Roldán et al., 2015) respectively. We reported preliminary results for the one-step delayed decision-making in (Grau-Moya and Braun, 2013; Grau-Moya et al., 2013). At the machine learning level generalized fluctuation theorems have been used in (Hayakawa and Aoyagi, 2015) to train artificial neural networks with efficient exploration. In general, fluctuation theorems and Jarzynski equalities allow to estimate free energy differences which are very important in decision-making because the free energy directly relates to the value function which is a central concept in control and reinforcement learning.

In conclusion, the results presented here bring the fields of stochastic thermodynamics and decision-making closer together. It can be useful to study decision-making systems as statistical systems just like in thermodynamics. The energy functions in physics correspond utility functions in decision-making with a certain caveats. The first is that decision-maker have limitations in computational power expressed by slow adaptation or limited time for planning, whereas in thermodynamic system have relaxation rates that can be potentially slow. Importantly, the statistical ensembles of both, decisions and physical states, can be conceptualized as non-equilibrium ensembles that reach equilibrium after a certain time.

Acknowledgments

This study was supported by the DFG, Emmy Noether grant BR4164/1-1 and by the DFG grant KR 3844/2-1.

Appendix: Fokker-Planck solution for unknown initial state

Although we have found solutions in the literature of the Fokker-Planck equation for known initial state x_0 (Risken, 1984) we have not for the case of unknown initial state. Here we derive the solution when the initial state is Gaussian distributed.

Consider the following dynamics:

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

where $A(x, t) = \alpha \frac{\partial U_1}{\partial x}$, $B(x, t) = \alpha$ and $U_1(x) = U_0(x) + \Delta U(x)$, $U_0(x) = \frac{1}{\beta} \log p_0(x) + \frac{1}{\beta} \log Z_0$ and thus

$$\frac{\partial U_1}{\partial x} = \frac{\partial \Delta U(x)}{\partial x} - \frac{1}{\beta} \frac{1}{p_0} \frac{\partial p_0(x)}{\partial x}.$$

When imposing Gaussian distributions in the potential $U_y(x) = -(a_y x^2 + b_y x)$ for any y ,

$$\frac{\partial U_1}{\partial x} = \frac{\partial \Delta U(x)}{\partial x} - \frac{1}{\beta} \frac{(x - \mu_0)}{\sigma_0^2}$$

Then the associated Fokker Planck Equation is

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

We will solve this equation by doing first the Fourier transform and then solve for the method of characteristics. The Fourier transform is

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

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

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

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

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

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

$$\begin{aligned} \frac{d\hat{P}}{\hat{P}} &= dt \left(c_2 i s_0 e^{ct} - \frac{\alpha^2 D}{2} s_0^2 e^{2ct} \right) \\ \log \hat{P} \Big|_{\hat{P}(s_0, t=0)}^{\hat{P}(s, t')} &= \frac{c_2 i s_0}{c} e^{ct} - \frac{\alpha^2 D}{4c} s_0^2 e^{2ct} \Big|_{t=0}^{t=t'} \end{aligned}$$

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

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

Then the solution in frequency space is

$$\begin{aligned} \hat{P}(s, t) &= \exp \left\{ -\frac{\alpha^2 D}{4c} s^2 (1 - e^{-2ct}) - \sigma_0^2 s^2 e^{-2ct} + i s \frac{b_1}{2a_1} (1 - e^{-ct}) - i s e^{-ct} \right\} \\ &= \exp \{ s^2 f_1(t) - i s f_2(t) \} \end{aligned}$$

with $f_1(t) = -\frac{\alpha^2 D}{4c} (1 - e^{-2ct}) - \frac{\sigma_0^2}{2} e^{-2ct}$ and $f_2(t) = e^{-ct} \mu_0 - \frac{b_1}{2a_1} (1 - e^{-ct})$. Transforming back to the signal domain we obtain

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

Chapter 7

Discussion

7.1 Summary

In the introduction of this thesis I presented an existing information-theoretic framework to model decision-making under both, bounded rationality and model uncertainty. Bounded rational behavior was described by a posterior distribution over actions that depended on the decision-maker's prior distribution and on its available computational resources tuned by the resource parameter α . Decision-making under model uncertainty was obtained by computing certainty equivalents from free energy optimizations. In this way we obtained risk-sensitive behavior given a prior model of the environment $p(y|x)$ with a particular risk-sensitive parameter β . Additionally, we also obtained ambiguity-sensitive behavior given a prior probability $\mu(\theta)$ on the possible world models represented by parameters θ and an ambiguity-sensitive parameter γ .

In the following three chapters we studied experimentally if human decision-makers were acting according to the previous information-theoretic models of decision-making. In particular, in Chapter 2 we studied human risk-sensitivity in a Bayesian sensorimotor integration task; in Chapter 3 we studied human ambiguity-sensitivity in two tasks, one involving uncertain choices between urns and another involving uncertain sensorimotor choices; in Chapter 4 we studied in a two-player game how model-uncertainty coming from both players affected cooperation. In the remaining two chapters, we focused on extending some theoretical aspects of the theory. In particular, in Chapter 5 we developed a generalized value iteration algorithm that can describe sequential decision-making behavior under both, bounded rationality and model uncertainty; in Chapter 6 we studied the inefficiencies of decision-makers when implementing sub-optimal bounded rational policies due to limited time and drew clear analogies with non-equilibrium thermodynamics.

In the introduction we introduced several question that I tried to answer with the results presented in this thesis. In the following I briefly answer these questions and then proceed to a more extended general discussion.

Experiments in Human Decision-Making *Are humans making choices according to the proposed information-theoretic models of bounded rationality, risk and ambiguity in laboratory experiments (Chapter 2, 3 and 4)?* Yes, however with some caveats that will be discussed in the following section.

- *Is the human sensorimotor system subject to risk-sensitivity in an estimation task (Chapter 2)?* Yes, we showed that at for linear utility functions the sensorimotor system acts in a risk-sensitive fashion for Bayesian estimation tasks.
- *Do the same humans have different ambiguity attitudes in different situations? What are the important factors that determine these ambiguity attitudes (Chapter 3)?* Yes, our experiments conclude that humans have different ambiguity attitudes in different tasks and that the important factor regarding this ambiguity attitude is how the uncertainty is visualized and not if the task is within a sensorimotor context.
- *In two-player games, how are cooperative solutions affected by model uncertainty coming from different players (Chapter 4)?* We performed simulations and tested human subjects in two-player games with model uncertainty and conclude that the risk-sensitive parameter of the opponent plays a crucial role in driving behavior towards cooperative plays.

Theoretical Advancements of Information-Theoretic Approaches to Decision-Making

- *How can we extend the theory in a sequential decision-making scenario to take into account bounded rationality and model uncertainty simultaneously when planning into the future (Chapter 5)?* We proposed an objective function to take into account not only future rewards but also future information costs coming from bounded rationality and model uncertainty. This new objective function has an analytic solution which could be exploited to derive a novel generalized value iteration algorithm. We tested the algorithm in a grid world environment.
- *Given that the free energy is a concept from statistical physics, how does bounded rational decision-making relate to non-equilibrium statistical physics? Can we make novel predictions when importing concepts from physics to decision-making (Chapter 6)?* We conclude that we can establish clear relationships with non-equilibrium thermodynamics by realizing that bounded rational decision-makers must spend time to compute solutions from the posterior equilibrium distribution. In this way, suboptimal behavior due to employing a non-equilibrium distribution can be quantified by the amount of dissipation in analogy with thermodynamics. Thanks to the concept of dissipation we derived novel relations that could be the basis for novel interesting predictions in the context of decision-making with limited resources.

In the following we discuss the results presented in this thesis and provide future interesting research directions.

7.2 Discussion and Outlook

A general question regarding the experimental part of this thesis was if human subjects make decisions according to the information-theoretic models of bounded rationality, risk-sensitivity and ambiguity in laboratory experiments. We conclude that the answer is yes, however, with some caveats that we discuss in the following sections. Additionally, regarding the theoretical part, we discuss how our algorithms can be improved for better scalability and how the theoretical models could be tested in human experiments.

Statistical complexity and utility functions.

In our first experiment, human subjects had to estimate a latent variable by producing a motor response. By adding sensorimotor costs we were able to induce risk-sensitive behavior. We modeled this behavior with an information-theoretic model that predicted a modulation of behavior depending on both, the uncertainty and the cost of the motor command. Importantly, the cost function employed in our experiments was simply a linear function in the space of motor commands. We observed that under this simple cost function subjects' behavior is consistent with the risk-sensitive predictions. However, we doubt that this will be the case in more complex scenarios with highly non-linear utility functions or complex statistical data. In fact, recently, complex internal representations in sensorimotor decision-making have been studied in the context of Bayesian Decision Theory (Acerbi, 2015). In particular, deviations from optimal Bayesian inference are investigated in complex scenarios with non-Gaussian statistical data. They find that the origins of suboptimal Bayesian behavior is due to the difficulty to learn such complex statistics (or priors) and not because of suboptimal computations of posterior distributions (Acerbi et al., 2014). We hypothesize that in the context of risk-sensitive Bayesian integration similar suboptimal behavior could in principle be found due to similar reasons, that is, by inaccurate learning of complex priors. However, in analogy with their results the computations involved to express risk-sensitivity might be optimal. In the context of our remaining experiments (Chapters 3 and 4), we also used simple utility functions and small decision-spaces. This allowed to verify that the information-theoretic model was consistent with the behavioral data, but it is not clear if these models will hold valid for complex decision-making processes. It would be interesting to test whether our predictions remain valid in more complex decision-making processes. For example, the information-theoretic model of risk-sensitive behavior could be tested using more complex non-linear cost functions or more complex statistical distributions involving latent variables.

Intricacies of bounded rational behavior.

In all our experiments subjects were showing instances of bounded rational behavior. However, due to our experimental design subjects learned flat priors over the distributions of actions. When assuming flat a priors, the posterior becomes simply a soft-max distribution in the bounded rational model. For example, in Chapter 3 and 4, subjects had a flat prior

over possible actions which induced a soft-max posterior as can be seen in Equations (4.4) and (4.4). In this sense we only partially validated the consistency of human behavior with the information-theoretic model of bounded rationality. In the case of our experiment in risk-sensitivity (Chapter 2), the prior was a Gaussian distribution over the latent variable. Importantly, we were focusing in explaining risk-sensitivity and in our predictions we disregarded the bounded rational part of the model by assuming perfectly rational agents i.e. assuming an infinite value for the rationality parameter. Due to this our predictions were delta distributions and not probability distributions over the possible actions. For this reason we could not test whether our subjects were acting according to the information-theoretic bounded rational model at a probabilistic level. However, when inspecting the data it seems this is likely to be the case as the sensorimotor actions provided by subjects were not distributed according to a delta distribution—see typical subject data in Figure 2.2.

Although in this thesis we focused on how human behavior is affected by model uncertainty, it could be an interesting to focus more on the bounded rationality part. For example, experiments could be done in Bayesian sensorimotor integration by limiting the allowed decision time in order to implicitly change the bounded rationality parameter of human decision-makers. Regarding the experiments' statistical properties, variations could be made in order to induce non-flat priors that make interesting predictions ready to be tested in similar experimental paradigms proposed in this thesis.

Partial ambiguity reversal.

In our second experiment subjects had to make choices under ambiguity in two tasks—an urn task and a sensorimotor task—under different framing such as sensorimotor framing and visual framing. Although in both tasks of Experiment 1 all subjects were acting in accordance with the information-theoretic ambiguity model, they did not do so in the control experiment—compare blue bars in Figure 3.4C. In this experiment the utility function was reversed to test whether subjects' choice behavior can be explained by genuinely ambiguity sensitivity or it is a consequence of idiosyncratic biases. The experimental data reveals a trend in choice behavior where subjects prefer the risky option when the ambiguous option has high ambiguity. Strangely, subjects show a reversal in choice behavior by adopting an increased preference for the ambiguous option only for intermediate levels of ambiguity. At the time of publication we were not aware of fMRI research on ambiguity that has similar results for partial ambiguity conditions. In particular, in (Lopez Paniagua and Seger, 2013) found that certain brain regions associated with the processing of ambiguity, the so-called 'fronto-parietal-striatal system', showed an increased activity for partial ambiguity conditions compared to no uncertainty or full uncertainty conditions. They hypothesize that this high activation in partial ambiguity conditions could be an indicator for the brain to search for useful information that is greater for trials with intermediate levels of ambiguity. They argue that in the case of risk (no-ambiguity) there is no need to search for additional information whereas in the case of full ambiguity it might be too difficult to search for useful information. In the context of our experiment, the behavioral data suggests that there is a perceptual bias

for partial ambiguity conditions, however, after reviewing the results from (Lopez Paniagua and Seger, 2013) the underlying reason of different behavior in partial ambiguity is still unclear. In principle it could be attributed not to a perceptual bias but to a more profound reason only valid to partial ambiguity decision-making.

Scalability in high dimensional decision-spaces.

In Chapter 5 we developed a generalized value iteration scheme for planning under both, bounded rationality and model uncertainty. The results provided are interesting and highlight the fact that we can design agents that can be robust in at least two different ways, due to the stochasticity in their actions and due to pessimism about the possible models of the environment. Additionally, we argued that exploration can also be achieved in two-ways, by being stochastic in the policy or by optimism about the world models. These ways to achieve robustness or exploration have been studied in the literature before but in separate ways. For example, exploration due to stochasticity has been induced by Boltzmann distributions or e-greedy approaches, whereas exploration through optimism has been done by Upper Confidence Bounds (UCB) algorithms where actions are valued optimistically depending on the underlying uncertainty. The algorithm presented in Chapter 5 that solves the planning problem in partially unknown MDP is computationally expensive and does not scale well with respect to the state and latent spaces. For this reason it would be interesting to implement sampling approaches that scale well. For example, the authors in (Guez et al., 2012; Guez et al., 2013) have explored Monte Carlo sampling techniques which seem a promising direction to follow in problems with high-dimensional state spaces. An alternative approach could be to design deep neural architectures for the approximation of the free energy function (that acts as a value or Q-function). For example, recently in (Mnih et al., 2015) it is shown that it is possible to learn such complex Q-functions by means of a deep neural network, achieving state-of-the-art performance in Atari games comparable to human professional players. It would be interesting to test our ideas to design agents that take into account model uncertainty and bounded rationality in more complex environments such as the Atari games. These agents would show interesting and diverse behavior such as aggressive or cautious behavior.

Validating the novel theoretical contributions by testing human behavior.

In Chapter 6 we explored the analogies between non-equilibrium thermodynamics and bounded rational decision-making under limited time. We derived new interesting relations that connect bounded optimal behavior with sub-optimal behavior due to limited time (the newly proposed counterparts of Jarzynski and Crooks relations for the decision-making case). There are only a few research articles drawing clear analogies between these two fields, some of them were discussed in the conclusions of Chapter 6. Possible future research in this direction would come from experiments in human decision-making that test whether the proposed relations hold and explain human behavioral data. For example, a possible paradigm to test these ideas would be in visuomotor adaptation tasks where one can clearly interpret non-equilibrium be-

havior as the adaptation dynamics. Additionally, in Chapter 5 we also proposed an interesting model of sequential decision-making behavior. It would be interesting to test whether human behavior can be explained with our model. In particular, it would be interesting to design a Markov Decision Problem in where humans solve sequential decision making tasks under model uncertainty and learn about the world. This could be easily modeled as our model provides a way to study learning in unknown environments and stochastic policies.

Setting the parameters.

In our experimental and theoretical work the parameters of the models that we used to explain or model behavior were set arbitrarily for the theoretical part or fitted in the experimental part. How to set and adapt these parameters depending on the nature of the decision-problems or the incoming data of the tasks at hand is still an open problem. However, interesting work has been done in scheduling the bounded rationality parameter. In (Fox et al., 2015), a temporal difference method is used to learn the free energy function in a way that the bounded rationality parameter is increased over time from zero to a very high value. This method allows for fast learning at early stages when the estimate of the free energy function is inaccurate and accurate estimates at later stages when the value of the rationality parameters is high. Thus with this scheduling, the algorithm transitions from learning a free energy value function with small rationality parameter to learning a classic Q-function when the rationality parameter is high. Another possible approach would be, for example, when employing a rejection sampling scheme (such as in Chapter 6) to track the number of rejections. One could design an scheduling of the rationality parameter in such a way that when there are many rejections the parameter decreases—thus simplifying the sampling problem—and when there are too many acceptances the parameter increases—to improve performance in the decision-making problem. Regarding the parameters for model uncertainty, the tuning of, for example, the risk-sensitive parameter is still an open problem.

Extending the information theoretical models

Our planning algorithm takes into account bounded rationality and model uncertainty. An extension of this algorithm could be a version of it that also includes risk-sensitivity. In this way, when the model of the environment is learned and there is no uncertainty about the parameters the decision-maker could still show risk-sensitive behavior. I believe that adding risk-sensitivity to the algorithm would be straightforward and interesting in its own right, but it could also amplify the computational demands. Another type of extension could be done by taking into account more intermediate variables. For example in (Genewein et al., 2015) the authors model a decision-maker with an internal stochastic representation (or abstraction) in order to maximize utility subject to information-theoretical constraints. Our model could be combined with their model by unified principle for temporal abstraction in the sequential decision-making case. This extension might not be trivial though.

7.3 Conclusions

In conclusion the goals of this thesis are met by providing behavioral evidence supporting information-theoretic models of decision-making under risk-sensitivity and ambiguity-sensitivity. Additionally, we extended the theoretical model for sequential decision-making under bounded rationality and model uncertainty that could be used for robust or explorative behavior in artificial agents. Moreover, the thermodynamic analogies drawn in this thesis open a novel approach to study the evolution of the decision-making process under limited time and precision. Finally, in our discussion we provide hints on how to continue investigations in this interesting line of research about information-theoretic decision-making. Overall, the combination of information theory and decision-making seems a promising theoretical framework to describe and model, in a unified way, several important characteristics of the decision-making process such as computational resources, robustness and exploration. Although, the ideas presented in this thesis are certainly not common in the decision-making community, I have no doubt the elegance of the theory and the experimental evidence will make them appealing for researchers in the field.

Bibliography

- Acerbi, Luigi (2015). *Complex internal representations in sensorimotor decision making: a Bayesian investigation. PhD Thesis*. The University of Edinburgh.
- Acerbi, Luigi, Sethu Vijayakumar, and Daniel M Wolpert (2014). “On the origins of suboptimality in human probabilistic inference”. In: *PLoS Computational Biology* 10.6, e1003661.
- Allais, Maurice (1953). “Le comportement de l’homme rationnel devant le risque: Critique des postulats et axiomes de l’école Américaine.” In: *Econometrica*.
- Anastasio, Thomas J, Paul E Patton, and Kamel Belkacem-Boussaid (2000). “Using Bayes’ rule to model multisensory enhancement in the superior colliculus”. In: *Neural Computation* 12.5, pp. 1165–1187.
- Arrow, Kenneth J. (1965). *Aspects of the Theory of Risk-Bearing*. Helsinki: Yrjö Jahansson Foundation.
- Arrow, Kenneth J and Leonid Hurwicz (1972). “An optimality criterion for decision-making under ignorance”. In: *Uncertainty and expectations in economics*, pp. 1–11.
- Åström, Karl J and Björn Wittenmark (2013). *Adaptive control*. Courier Corporation.
- Avenanti, Alessio, Angela Sirigu, and Salvatore M Aglioti (2010). “Racial bias reduces empathic sensorimotor resonance with other-race pain.” eng. In: *Current Biology* 20.11, pp. 1018–1022.
- Bach, Dominik R and Raymond J Dolan (2012). “Knowing how much you don’t know: a neural organization of uncertainty estimates”. In: *Nature Reviews Neuroscience* 13.8, pp. 572–586.
- Bach, Dominik R, Ben Seymour, and Raymond J Dolan (2009). “Neural activity associated with the passive prediction of ambiguity and risk for aversive events”. In: *The Journal of Neuroscience* 29.6, pp. 1648–1656.
- Bartra, Oscar, Joseph T McGuire, and Joseph W Kable (2013). “The valuation system: a coordinate-based meta-analysis of BOLD fMRI experiments examining neural correlates of subjective value”. In: *Neuroimage* 76, pp. 412–427.
- Beers, Robert J van, Patrick Haggard, and Daniel M Wolpert (2004). “The role of execution noise in movement variability.” eng. In: *Journal of Neurophysiology* 91.2, pp. 1050–1063.
- Beers, Robert J van, Anne C Sittig, and Jan J Denier van Der Gon (1999). “Integration of proprioceptive and visual position-information: An experimentally supported model”. In: *Journal of neurophysiology* 81.3, pp. 1355–1364.

- Bellman, Richard (1957). *Dynamic Programming*. 1st ed. Princeton, NJ, USA: Princeton University Press.
- Berger, Ulrich (2005). “Fictitious play in 2 x n games”. In: *Journal of Economic Theory* 120, pp. 139–154.
- (2007). “Brown’s original fictitious play.” In: *Journal of Economic Theory* 135.1, pp. 572–578.
- Bernardo, José M (1997). “Statistical inference as a decision problem: the choice of sample size”. In: *The Statistician*, pp. 151–153.
- Bernoulli, Daniel (1954). “Exposition of a new theory on the measurement of risk”. In: *Econometrica: Journal of the Econometric Society*, pp. 23–36.
- Bertsekas, Dimitri P and John N Tsitsiklis (1996). *Neuro-Dynamic Programming*. Athena Scientific.
- Blakemore, Sarah J, Susan J Goodbody, and Daniel M Wolpert (1998). “Predicting the consequences of our own actions: the role of sensorimotor context estimation”. In: *The Journal of Neuroscience* 18.18, pp. 7511–7518.
- Braun, Daniel A, Ad Aertsen, Daniel M Wolpert, and Carsten Mehring (2009a). “Learning optimal adaptation strategies in unpredictable motor tasks”. eng. In: *Journal of Neuroscience* 29.20, pp. 6472–8.
- Braun, Daniel A, Arne J Nagengast, and Daniel Wolpert (2011a). “Risk-sensitivity in sensorimotor control”. In: *Frontiers in human neuroscience* 5, p. 1.
- Braun, Daniel A and Pedro A Ortega (2014). “Information-Theoretic Bounded Rationality and ε -Optimality”. In: *Entropy* 16.8, pp. 4662–4676.
- Braun, Daniel A, Pedro A Ortega, Evangelos Theodorou, and Stefan Schaal (2011b). “Path integral control and bounded rationality”. In: *Adaptive Dynamic Programming And Reinforcement Learning (ADPRL), 2011 IEEE Symposium on*. IEEE, pp. 202–209.
- Braun, Daniel A, Pedro A Ortega, and Daniel M Wolpert (2009b). “Nash equilibria in multi-agent motor interactions.” eng. In: *PLoS Computational Biology* 5.8, e1000468.
- (2011c). “Motor coordination: when two have to act as one”. In: *Experimental brain research* 211.3-4, pp. 631–641.
- Brodlie, Ken, Rodolfo Allendes Osorio, and Adriano Lopes (2012). “A review of uncertainty in data visualization”. In: *Expanding the Frontiers of Visual Analytics and Visualization*. Springer, pp. 81–109.
- Broek, Bart van den, Wim Wiegierinck, and Hilbert J. Kappen (2010). “Risk Sensitive Path Integral Control”. In: *Uncertainty in Artificial Intelligence*.
- Brown, Lawrence D. (1981). “A Complete Class Theorem for Statistical Problems with Finite Sample Spaces”. English. In: *The Annals of Statistics* 9.6,
- Buesing, Lars, Johannes Bill, Bernhard Nessler, and Wolfgang Maass (2011). “Neural dynamics as sampling: a model for stochastic computation in recurrent networks of spiking neurons”. In: *PLoS Computational Biology* 7.11, e1002211.

-
- Buschman, Timothy J, Markus Siegel, Jefferson E Roy, and Earl K Miller (2011). “Neural substrates of cognitive capacity limitations”. In: *Proceedings of the National Academy of Sciences* 108.27, pp. 11252–11255.
- Büyükboyacı, Mürüvvet (2014). “Risk attitudes and the stag-hunt game”. In: *Economics Letters* 124.3, pp. 323–325.
- Camerer, Colin (2003). *Behavioral game theory: Experiments in strategic interaction*. Princeton University Press.
- Camerer, Colin and Martin Weber (1992). “Recent developments in modeling preferences: Uncertainty and ambiguity”. In: *Journal of Risk and Uncertainty* 5.4, pp. 325–370.
- Chakravarty, Sujoy and Jaideep Roy (2009). “Recursive expected utility and the separation of attitudes towards risk and ambiguity: an experimental study”. English. In: *Theory and Decision* 66.3, pp. 199–228.
- Chen-Harris, Haiyin, Wilsaan M Joiner, Vincent Ethier, David S Zee, and Reza Shadmehr (2008). “Adaptive control of saccades via internal feedback.” eng. In: *Journal of Neuroscience* 28.11, pp. 2804–2813.
- Chib, Siddhartha and Edward Greenberg (1995). “Understanding the metropolis-hastings algorithm”. In: *The american statistician* 49.4, pp. 327–335.
- Chow, Yinlam, Aviv Tamar, Shie Mannor, and Marco Pavone (2015). “Risk-Sensitive and Robust Decision-Making: a CVaR Optimization Approach”. In: *Advances in Neural Information Processing Systems*, pp. 1522–1530.
- Christopoulos, George I, Philippe N Tobler, Peter Bossaerts, Raymond J Dolan, and Wolfram Schultz (2009). “Neural correlates of value, risk, and risk aversion contributing to decision making under risk”. In: *The Journal of Neuroscience* 29.40, pp. 12574–12583.
- Chumbley, Justin R., Guillaume Flandin, Dominik R. Bach, Jean Daunizeau, Ernst Fehr, Raymond J. Dolan, and Karl J. Friston (2012). “Learning and Generalization under Ambiguity: An fMRI Study”. In: *PLoS Comput Biol* 8.1, e1002346.
- Cisek, Paul (2007). “Cortical mechanisms of action selection: the affordance competition hypothesis”. In: *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 362.1485, pp. 1585–1599.
- Coricelli, Giorgio and Rosemarie Nagel (2009). “Neural correlates of depth of strategic reasoning in medial prefrontal cortex”. In: *Proceedings of the National Academy of Sciences* 106.23, pp. 9163–9168.
- Cosmides, Leda, John Tooby, and Robert Kurzban (2003). “Perceptions of race.” eng. In: *Trends Cogn Sci* 7.4, pp. 173–179.
- Costantini, Stefania (2002). “Meta-reasoning: a survey”. In: *Computational Logic: Logic Programming and Beyond*. Springer, pp. 253–288.
- Crooks, Gavin E (1998). “Nonequilibrium measurements of free energy differences for microscopically reversible Markovian systems”. In: *Journal of Statistical Physics* 90.5-6, pp. 1481–1487.
- De Finetti, Bruno (1937). “La prévision: ses lois logiques, ses sources subjectives”. In: *Annales de l’institut Henri Poincaré*. Vol. 7. 1, pp. 1–68.
-

- Dean, Thomas L and Mark S Boddy (1988). “An Analysis of Time-Dependent Planning.” In: *AAAI*. Vol. 88, pp. 49–54.
- Diedrichsen, Jörn (2007). “Optimal task-dependent changes of bimanual feedback control and adaptation.” eng. In: *Current Biology* 17.19, pp. 1675–1679.
- Diedrichsen, Jörn and Noreen Dowling (2009). “Bimanual coordination as task-dependent linear control policies.” eng. In: *Human Movement Science* 28.3, pp. 334–347.
- Diedrichsen, Jörn, Reza Shadmehr, and Richard B Ivry (2010). “The coordination of movement: optimal feedback control and beyond”. In: *Trends in cognitive sciences* 14.1, pp. 31–39.
- Doya, Kenji, Shin Ishii, Alexandre Pouget, and Rajesh PN Rao, eds. (2007). *Bayesian Brain: Probabilistic Approaches to Neural Coding*. The MIT Press.
- Duff, Michael O’Gordon (2002). “Optimal Learning: Computational procedures for Bayes-adaptive Markov decision processes”. PhD thesis. University of Massachusetts Amherst.
- Ellsberg, Daniel (1961). “Risk, Ambiguity, and the Savage Axioms”. In: *The Quarterly Journal of Economics* 75.4, pp. 643–669.
- Ernst, Marc O and Martin S Banks (2002). “Humans integrate visual and haptic information in a statistically optimal fashion.” eng. In: *Nature* 415.6870, pp. 429–433.
- Etner, Johanna, Meglena Jeleva, and Jean-Marc Tallon (2012). “Decision theory under ambiguity”. In: *Journal of Economic Surveys* 26.2, pp. 234–270.
- Faisal, A Aldo, Luc PJ Selen, and Daniel M Wolpert (2008). “Noise in the nervous system”. In: *Nature reviews neuroscience* 9.4, pp. 292–303.
- Feltovich, Nick, Atsushi Iwasaki, and Sobei H. Oda (2012). “Payoff levels, loss avoidance, and equilibrium selection in games with multiple equilibria: An experimental study”. In: *Economic Inquiry* 50.4, pp. 932–952.
- Fiser, József, Pietro Berkes, Gergő Orbán, and Máté Lengyel (2010). “Statistically optimal perception and learning: from behavior to neural representations”. In: *Trends in cognitive sciences* 14.3, pp. 119–130.
- Flanagan, J Randall and Alan M Wing (1997). “The role of internal models in motion planning and control: evidence from grip force adjustments during movements of hand-held loads”. In: *The Journal of Neuroscience* 17.4, pp. 1519–1528.
- Föllmer, Hans and Alexander Schied (2011). *Stochastic finance: an introduction in discrete time*. Walter de Gruyter.
- Fox, Roy, Ari Pakman, and Naftali Tishby (2015). “G-Learning: Taming the Noise in Reinforcement Learning via Soft Updates”. In: *arXiv preprint arXiv:1512.08562*.
- Friston, Karl (2009). “The free-energy principle: a rough guide to the brain?” In: *Trends in cognitive sciences* 13.7, pp. 293–301.
- (2010). “The free-energy principle: a unified brain theory?” In: *Nature Reviews Neuroscience* 11.2, pp. 127–138.
- Friston, Karl, Spyridon Samothrakis, and Read Montague (2012). “Active inference and agency: optimal control without cost functions”. English. In: *Biological Cybernetics* 106.8-9, pp. 523–541.

-
- Fudenberg, Drew and David K Levine (1998). *The theory of learning in games*. Vol. 2. MIT press.
- Garcia-Palacios, JL (2007). “Introduction to the theory of stochastic processes and Brownian motion problems”. In: *arXiv preprint cond-mat/0701242*.
- Gaveau, Bernard and L.S. Schulman (1997). “A general framework for non-equilibrium phenomena: the master equation and its formal consequences”. In: *Physics Letters A* 229.6, pp. 347–353.
- Genewein, Tim, Felix Leibfried, Jordi Grau-Moya, and Daniel Alexander Braun (2015). “Bounded Rationality, Abstraction, and Hierarchical Decision-Making: An Information-Theoretic Optimality Principle”. In: *Frontiers in Robotics and AI* 2, p. 27.
- Geramifard, Alborz, Christoph Dann, Robert H Klein, William Dabney, and Jonathan P How (2015). “RLPy: A Value-Function-Based Reinforcement Learning Framework for Education and Research”. In: *Journal of Machine Learning Research* 16, pp. 1573–1578.
- Gershman, Samuel J, Eric J Horvitz, and Joshua B Tenenbaum (2015). “Computational rationality: A converging paradigm for intelligence in brains, minds, and machines”. In: *Science* 349.6245, pp. 273–278.
- Ghirardato, Paolo, Fabio Maccheroni, and Massimo Marinacci (2004). “Differentiating ambiguity and ambiguity attitude”. In: *Journal of Economic Theory* 118.2, pp. 133–173.
- Gigerenzer, Gerd and Daniel G Goldstein (1996). “Reasoning the fast and frugal way: models of bounded rationality.” In: *Psychological review* 103.4, p. 650.
- Gigerenzer, Gerd and Reinhard Selten (2002). *Bounded rationality: The adaptive toolbox*.
- Gilboa, Itzhak and Massimo Marinacci (2011). *Ambiguity and the Bayesian Paradigm*. Tech. rep. IGIER (Innocenzo Gasparini Institute for Economic Research), Bocconi University.
- Gilboa, Itzhak and David Schmeidler (1989). “Maxmin expected utility with non-unique prior”. In: *Journal of mathematical economics* 18.2, pp. 141–153.
- Gino, Francesca, Zachariah Sharek, and Don A Moore (2011). “Keeping the illusion of control under control: Ceilings, floors, and imperfect calibration”. In: *Organizational Behavior and Human Decision Processes* 114.2, pp. 104–114.
- Girshick, Ahna R and Martin S Banks (2009). “Probabilistic combination of slant information: weighted averaging and robustness as optimal percepts.” eng. In: *Journal of Vision* 9.9, pp. 8.1–820.
- Gomez-Marin, Alex, Juan MR Parrondo, and Christian Van den Broeck (2008). “Lower bounds on dissipation upon coarse graining”. In: *Physical Review E* 78.1, p. 011107.
- Grabenhorst, Fabian and Edmund T Rolls (2011). “Value, pleasure and choice in the ventral prefrontal cortex”. In: *Trends in cognitive sciences* 15.2, pp. 56–67.
- Grau-Moya, Jordi and Daniel A Braun (2013). “Bounded rational decision-making in changing environments”. In: *arXiv preprint arXiv:1312.6726*.
- Grau-Moya, Jordi, Eduard Hez, Giovanni Pezzulo, and Daniel A. Braun (2013). “The effect of model uncertainty on cooperation in sensorimotor interactions”. In: *Journal of The Royal Society Interface* 10.87.
-

- Grau-Moya, Jordi, Felix Leibfried, Tim Genewein, and Daniel A. Braun (2016a). “Planning with Information-Processing Constraints and Model Uncertainty in Markov Decision Processes”. In: *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2016, Riva del Garda, Italy, September 19-23, 2016, Proceedings, Part II*. Ed. by Paolo Frasconi, Niels Landwehr, Giuseppe Manco, and Jilles Vreeken. Cham: Springer International Publishing, pp. 475–491.
- Grau-Moya, Jordi, Pedro A. Ortega, and Daniel A. Braun (2012). “Risk-Sensitivity in Bayesian Sensorimotor Integration”. In: *PLoS Comput Biol* 8.9, e1002698.
- Grau-Moya, Jordi, Pedro A Ortega, and Daniel A Braun (2016b). “Decision-Making under Ambiguity Is Modulated by Visual Framing, but Not by Motor vs. Non-Motor Context. Experiments and an Information-Theoretic Ambiguity Model”. In: *PloS one* 11.4, e0153179.
- Gregersen, Nils Petter (1996). “Young drivers’ overestimation of their own skill—an experiment on the relation between training strategy and skill”. In: *Accident Analysis and Prevention* 28.2, pp. 243–250.
- Griffiths, Thomas L and Joshua B Tenenbaum (2006). “Optimal predictions in everyday cognition”. In: *Psychological Science* 17.9, pp. 767–773.
- Guez, Arthur, David Silver, and Peter Dayan (2012). “Efficient Bayes-adaptive reinforcement learning using sample-based search”. In: *Advances in Neural Information Processing Systems*, pp. 1025–1033.
- (2013). “Scalable and efficient Bayes-adaptive reinforcement learning based on Monte-Carlo tree search”. In: *Journal of Artificial Intelligence Research*, pp. 841–883.
- Hanany, Eran and Peter Klibanoff. “Updating ambiguity averse preferences”. In: *The BE Journal of Theoretical Economics* 9.1.
- Hansen, Lars Peter. and Thomas J. Sargent (2008). *Robustness*. Princeton University Press.
- Harris, Adam J. L. and Ulrike Hahn (2011). “Unrealistic optimism about future life events: A cautionary note.” In: *Psychological Review* 118.1, pp. 135–154.
- Harris, Christopher M and Daniel M Wolpert (1998). “Signal-dependent noise determines motor planning”. In: *Nature* 394.6695, pp. 780–784.
- Hayakawa, Takashi and Toshio Aoyagi (2015). “Learning in neural networks based on a generalized fluctuation theorem”. In: *Physical Review E* 92.5, p. 052710.
- Hick, William E (1952). “On the rate of gain of information”. In: *Quarterly Journal of Experimental Psychology* 4.1, pp. 11–26.
- Hirono, Yuji and Yoshimasa Hidaka (2015). “Jarzynski-type equalities in gambling: role of information in capital growth”. In: *arXiv preprint arXiv:1505.06216*.
- Horvitz, Eric (1988). “Reasoning under Varying and Uncertain Resource Constraints.” In: *AAAI*. Vol. 88, pp. 111–116.
- Howard, Ian S, James N Ingram, and Daniel M Wolpert (2009). “A modular planar robotic manipulandum with end-point torque control”. eng. In: *Journal of Neuroscience Methods* 181.2, pp. 199–211.

-
- Howard, Ian S., James N. Ingram, and Daniel M. Wolpert (2009). “A modular planar robotic manipulandum with end-point torque control”. In: *Journal of Neuroscience Methods* 181.2, pp. 199–211.
- Howes, Andrew, Richard L Lewis, and Alonso Vera (2009). “Rational adaptation under task and processing constraints: implications for testing theories of cognition and action.” In: *Psychological review* 116.4, p. 717.
- Hsu, Ming, Meghana Bhatt, Ralph Adolphs, Daniel Tranel, and Colin F Camerer (2005). “Neural systems responding to degrees of uncertainty in human decision-making”. In: *Science* 310.5754, pp. 1680–1683.
- Huettel, Scott A, C Jill Stowe, Evan M Gordon, Brent T Warner, and Michael L Platt (2006). “Neural signatures of economic preferences for risk and ambiguity”. In: *Neuron* 49.5, pp. 765–775.
- Hyman, Ray (1953). “Stimulus information as a determinant of reaction time.” In: *Journal of experimental psychology* 45.3, p. 188.
- Iani, Cristina, Filomena Anelli, Roberto Nicoletti, Luciano Arcuri, and Sandro Rubichi (2011). “The role of group membership on the modulation of joint action.” eng. In: *Exp Brain Res* 211.3-4, pp. 439–445.
- Inukai, Keigo and Taiki Takahashi (2009). “Decision under ambiguity: effects of sign and magnitude”. In: *International Journal of Neuroscience* 119.8, pp. 1170–1178.
- Iyengar, Garud N (2005). “Robust dynamic programming”. In: *Mathematics of Operations Research* 30.2, pp. 257–280.
- Izawa, Jun, Tushar Rane, Opher Donchin, and Reza Shadmehr (2008). “Motor adaptation as a process of reoptimization.” eng. In: *Journal of Neuroscience* 28.11, pp. 2883–2891.
- Jarzynski, C. (1997). “Nonequilibrium Equality for Free Energy Differences”. In: *Phys. Rev. Lett.* 78, pp. 2690–2693.
- Jarzynski, Christopher (2011). “Equalities and inequalities: irreversibility and the second law of thermodynamics at the nanoscale”. In: *Annu. Rev. Condens. Matter Phys.* 2.1, pp. 329–351.
- Johnson, Chris (2004). “Top Scientific Visualization Research Problems”. In: *IEEE Comput. Graph. Appl.* 24.4, pp. 13–17.
- Kahneman, D and A Tversky (1979). “Prospect Theory: An analysis of decision under risk”. In: *Econometrica* 47.2, pp. 263–291.
- Kahneman, Daniel (2003). “Maps of bounded rationality: Psychology for behavioral economics”. In: *American economic review*, pp. 1449–1475.
- Kahneman, Daniel and Amos Tversky (1979). “Prospect theory: An analysis of decision under risk”. In: *Econometrica: Journal of the econometric society*, pp. 263–291.
- (1984). “Choices, values, and frames.” In: *American psychologist* 39.4, p. 341.
- Kappen, Hilbert J (2005a). “Linear theory for control of nonlinear stochastic systems”. In: *Physical review letters* 95.20, p. 200201.
- (2005b). “Path integrals and symmetry breaking for optimal control theory”. In: *Journal of statistical mechanics: theory and experiment* 2005.11, P11011.
-

- Kappen, Hilbert J, Vicenç Gómez, and Manfred Opper (2012). “Optimal control as a graphical model inference problem”. In: *Machine learning* 87.2, pp. 159–182.
- Kawato, Mitsu (1999). “Internal models for motor control and trajectory planning”. In: *Current opinion in neurobiology* 9.6, pp. 718–727.
- Kepecs, Adam, Naoshige Uchida, Hatim A Zariwala, and Zachary F Mainen (2008). “Neural correlates, computation and behavioural impact of decision confidence”. In: *Nature* 455.7210, pp. 227–231.
- Keren, Gideon and Léonie E.M. Gerritsen (1999). “On the robustness and possible accounts of ambiguity aversion”. In: *Acta Psychologica* 103.1-2, pp. 149–172.
- Kitano, Hiroaki (2004). “Biological robustness”. In: *Nat Rev Genet* 5.11, pp. 826–837.
- Klibanoff, Peter, Massimo Marinacci, and Sujoy Mukerji (2005). “A smooth model of decision making under ambiguity”. In: *Econometrica* 73.6, pp. 1849–1892.
- Knight, Frank H (1921). *Risk, Uncertainty and Profit*. Boston, MA: Houghton Mifflin.
- Knill, David C and Alexandre Pouget (2004). “The Bayesian brain: the role of uncertainty in neural coding and computation”. In: *Trends in Neurosciences* 27.12, pp. 712–719.
- Koechlin, Etienne, Jean Luc Anton, and Yves Burnod (1999). “Bayesian inference in populations of cortical neurons: a model of motion integration and segmentation in area MT”. In: *Biological cybernetics* 80.1, pp. 25–44.
- Koopmans, Tjalling C et al. (1951). *Activity analysis of production and allocation*. 13. Wiley New York.
- Körding, Konrad (2007). “Decision Theory: What ”Should” the Nervous System Do?” In: *Science* 318.5850, pp. 606–610.
- Körding, Konrad P and Daniel M Wolpert (2004). “Bayesian integration in sensorimotor learning”. In: *Nature* 427.6971, pp. 244–247.
- (2006). “Bayesian decision theory in sensorimotor control.” eng. In: *Trends in Cognitive Sciences* 10.7, pp. 319–326.
- Krain, Amy L., Amanda M. Wilson, Robert Arbuckle, F. Xavier Castellanos, and Michael P. Milham (2006). “Distinct neural mechanisms of risk and ambiguity: A meta-analysis of decision-making”. In: *NeuroImage* 32.1, pp. 477–484.
- Krishna, Vijay and Tomas Sjostrom (1998). “On the convergence of fictitious play”. In: *Mathematics of Operations Research*, pp. 479–511.
- Kruger, Justin (1999). “Lake Wobegon be gone! The “below-average effect” and the egocentric nature of comparative ability judgments.” In: *Journal of Personality and Social Psychology* 77.2, pp. 221–232.
- Kruger, Justin and David Dunning (1999). “Unskilled and unaware of it: How difficulties in recognizing one’s own incompetence lead to inflated self-assessments.” In: *Journal of Personality and Social Psychology* 77.6, pp. 1121–1134.
- Langer, Ellen J (1975). “The illusion of control.” In: *Journal of personality and social psychology* 32.2, p. 311.

-
- Levy, Ifat, Jason Snell, Amy J Nelson, Aldo Rustichini, and Paul W Glimcher (2010). “Neural representation of subjective value under risk and ambiguity”. In: *Journal of neurophysiology* 103.2, pp. 1036–1047.
- Levy, Roger P, Florencia Reali, and Thomas L Griffiths (2009). “Modeling the effects of memory on human online sentence processing with particle filters”. In: *Advances in neural information processing systems*, pp. 937–944.
- Lieder, Falk, Thomas Griffiths, and Noah Goodman (2012). “Burn-in, bias, and the rationality of anchoring”. In: *Advances in neural information processing systems*, pp. 2690–2798.
- Lindley, Dennis V (1997). “The choice of sample size”. In: *The Statistician*, pp. 129–138.
- Lopez Paniagua, Dan and Carol Seger (2013). “Coding of level of ambiguity within neural systems mediating choice”. In: *Frontiers in neuroscience* 7, p. 229.
- Losin, Elizabeth A Reynolds, Marco Iacoboni, Alia Martin, Katy A. Cross, and Mirella Dapretto (2012). “Race modulates neural activity during imitation.” eng. In: *Neuroimage* 59.4, pp. 3594–3603.
- Luce, Duncan R (1959). *Individual Choice Behavior*. Wiley.
- Ma, Wei J., Jeffrey M. Beck, Peter E. Latham, and Alexandre Pouget (2006). “Bayesian inference with probabilistic population codes”. In: *Nature Neuroscience* 9.11, pp. 1432–1438.
- Ma, Wei Ji, Jeffrey M Beck, and Alexandre Pouget (2008). “Spiking networks for Bayesian inference and choice.” eng. In: *Curr Opin Neurobiol* 18.2, pp. 217–222.
- Maccheroni, Fabio, Massimo Marinacci, and Aldo Rustichini (2006). “Ambiguity Aversion, Robustness, and the Variational Representation of Preferences”. In: *Econometrica* 74.6, pp. 1447–1498.
- Machina, Mark and W Kip Viscusi (2013). *Handbook of the Economics of Risk and Uncertainty*. Newnes.
- MacKay, David JC (1998). “Introduction to monte carlo methods”. In: *Learning in graphical models*. Springer, pp. 175–204.
- Mannor, Shie, Duncan Simester, Peng Sun, and John N Tsitsiklis (2007). “Bias and variance approximation in value function estimates”. In: *Management Science* 53.2, pp. 308–322.
- Marewski, Julian N, Wolfgang Gaissmaier, and Gerd Gigerenzer (2010). “Good judgments do not require complex cognition”. In: *Cognitive processing* 11.2, pp. 103–121.
- Markowitz, Harry (1952). “Portfolio selection”. In: *The journal of finance* 7.1, pp. 77–91.
- Marois, René and Jason Ivanoff (2005). “Capacity limits of information processing in the brain”. In: *Trends in cognitive sciences* 9.6, pp. 296–305.
- Marx, Vivien (2013). “Data visualization: ambiguity as a fellow traveler”. In: *Nature methods* 10.7, pp. 613–615.
- Mattsson, Lars-Göran and Jörgen W. Weibull (2002). “Probabilistic choice and procedurally bounded rationality”. In: *Games and Economic Behavior* 41.1, pp. 61–78.
- McFadden, Daniel (1980). “Econometric Models for Probabilistic Choice Among Products”. English. In: *The Journal of Business* 53.3,
-

- McKelvey, Richard D and Thomas R Palfrey (1995). “Quantal response equilibria for normal form games”. In: *Games and economic behavior* 10.1, pp. 6–38.
- Medina, José Ramón, Dongheui Lee, and Sandra Hirche (2012). “Risk-sensitive optimal feedback control for haptic assistance”. In: *Robotics and Automation (ICRA), 2012 IEEE International Conference on*. IEEE, pp. 1025–1031.
- Medina, José Ramón, Tamara Lorenz, and Sandra Hirche (2015). “Synthesizing Anticipatory Haptic Assistance Considering Human Behavior Uncertainty”. In: *IEEE Transactions on Robotics* 31.1, pp. 180–190.
- Meginnis, J.R. (1976). “A new class of symmetric utility rules for gambles, subjective marginal probability functions, and a generalized Bayes rule.” In: *Proceedings of the American Statistical Association, Business and Economic Statistics Section*, pp. 471–476.
- Mehta, Biren and Stefan Schaal (2002). “Forward models in visuomotor control”. eng. In: *J Neurophysiol* 88.2, pp. 942–53.
- Mlodinow, Leonard and Todd A Brun (2014). “Relation between the psychological and thermodynamic arrows of time”. In: *Physical Review E* 89.5, p. 052102.
- Mnih, Volodymyr et al. (2015). “Human-level control through deep reinforcement learning”. In: *Nature* 518.7540, pp. 529–533.
- Moreno-Bote, Rubén, David C Knill, and Alexandre Pouget (2011). “Bayesian sampling in visual perception”. In: *Proceedings of the National Academy of Sciences* 108.30, pp. 12491–12496.
- Nagengast, Arne J, Daniel A Braun, and Daniel M Wolpert (2009). “Optimal control predicts human performance on objects with internal degrees of freedom”. eng. In: *PLoS Computational Biology* 5.6, e1000419.
- (2010). “Risk-sensitive optimal feedback control accounts for sensorimotor behavior under uncertainty”. In: *PLoS Comput Biol* 6.7, e1000857.
- Nagengast, Arne J, Daniel A. Braun, and Daniel M. Wolpert (2011a). “Risk sensitivity in a motor task with speed-accuracy trade-off”. In: *Journal of Neurophysiology* 105, pp. 2668–2674.
- Nagengast, Arne J, Daniel A Braun, and Daniel M Wolpert (2011b). “Risk-sensitivity and the mean-variance trade-off: decision making in sensorimotor control”. In: *Proceedings of the Royal Society B: Biological Sciences* 278.1716, pp. 2325–2332.
- Neumann, Thomas and Bodo Vogt (2009). *Do Players’ Beliefs or Risk Attitudes Determine The Equilibrium Selections in 2x2 Coordination Games?* Tech. rep. Otto-von-Guericke University Magdeburg, Faculty of Economics and Management.
- Nilim, Arnab and Laurent El Ghaoui (2005). “Robust control of Markov decision processes with uncertain transition matrices”. In: *Operations Research* 53.5, pp. 780–798.
- O’Neill, Martin and Wolfram Schultz (2010). “Coding of reward risk by orbitofrontal neurons is mostly distinct from coding of reward value”. In: *Neuron* 68.4, pp. 789–800.
- Ortega, Pedro A and Daniel A Braun (2010a). “A Bayesian Rule for Adaptive Control based on Causal Interventions”. In: *3d Conference on Artificial General Intelligence (AGI-2010)*. Atlantis Press.

-
- (2010b). “A minimum relative entropy principle for learning and acting”. In: *Journal of Artificial Intelligence Research*, pp. 475–511.
 - (2011). “Information, utility and bounded rationality”. In: *Lecture notes on artificial intelligence*. Vol. 6830, pp. 269–274.
 - (2013). “Thermodynamics as a theory of decision-making with information-processing costs”. In: *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Science* 469.2153.
 - (2014). “Generalized Thompson sampling for sequential decision-making and causal inference”. In: *Complex Adaptive Systems Modeling* 2.1, p. 2.
- Ortega, Pedro A, Daniel A Braun, Justin Dyer, Kee-Eung Kim, and Naftali Tishby (2015). “Information-Theoretic Bounded Rationality”. In: *arXiv preprint arXiv:1512.06789*.
- Ortega, Pedro A, Daniel A Braun, and Naftali Tishby (2014). “Monte Carlo methods for exact & efficient solution of the generalized optimality equations”. In: *Robotics and Automation (ICRA), 2014 IEEE International Conference on*. IEEE, pp. 4322–4327.
- Osborne, Martin J and Ariel Rubinstein (1999). *A Course in Game Theory*. MIT Press.
- Ozogami, Takayuki (2012). “Robustness and risk-sensitivity in Markov decision processes”. In: *Advances in Neural Information Processing Systems*, pp. 233–241.
- Pang, Alex T., Craig M. Wittenbrink, and Suresh K. Lodh (1996). “Approaches to Uncertainty Visualization”. In: *The Visual Computer* 13, pp. 370–390.
- Parkes, David C and Michael P Wellman (2015). “Economic reasoning and artificial intelligence”. In: *Science* 349.6245, pp. 267–272.
- Peters, Jan, Katharina Mülling, and Yasemin Altün (2010). “Relative entropy policy search”. In: *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence*. AAAI Press, pp. 1607–1612.
- Pezzulo, Giovanni and Haris Dindo (2011). “What should I do next? Using shared representations to solve interaction problems”. In: *Experimental Brain Research* 211.3, pp. 613–630.
- Platt, Michael L and Scott A Huettel (2008). “Risky business: the neuroeconomics of decision making under uncertainty”. In: *Nature neuroscience* 11.4, pp. 398–403.
- Pouget, Alexandre, Jeffrey M Beck, Wei Ji Ma, and Peter E Latham (2013). “Probabilistic brains: knowns and unknowns”. In: *Nature neuroscience* 16.9, pp. 1170–1178.
- Pratt, John W (1964). “Risk Aversion in the Small and in the Large”. In: *Econometrica* 32.1/2, pp. 122–136.
- Pratt, John W. (1964). “Risk Aversion in the Small and in the Large”. In: *Econometrica* 32.1/2.
- Preuschoff, Kerstin, Steven R Quartz, and Peter Bossaerts (2008). “Human insula activation reflects risk prediction errors as well as risk”. In: *The Journal of neuroscience* 28.11, pp. 2745–2752.
- Pulford, Briony D and Andrew M Coleman (2008). “Ambiguity Aversion in Ellsberg Urns with Few Balls”. In: *Experimental Psychology* 55.1, pp. 31–37.

- Ramezani, Vahid Reza and Steven I Marcus (2005). “Risk-sensitive probability for Markov chains”. In: *Systems & control letters* 54.5, pp. 493–502.
- Ramsey, Frank P (1926). “Truth and probability”. In: *The foundations of mathematics and other logical essays (1931)* 156-198.
- Rangel, Antonio and Todd Hare (2010). “Neural computations associated with goal-directed choice”. In: *Current opinion in neurobiology* 20.2, pp. 262–270.
- Ray, Paramesh (1973). “Independence of irrelevant alternatives”. In: *Econometrica: Journal of the Econometric Society*, pp. 987–991.
- Rieskamp, Jörg (2008). “The probabilistic nature of preferential choice.” In: *Journal of Experimental Psychology: Learning, Memory, and Cognition* 34.6, p. 1446.
- Risken, Hannes (1984). “Fokker-planck equation”. In: *The Fokker-Planck Equation*. Springer, pp. 63–95.
- Roldán, Édgar (2014). *Irreversibility and dissipation in microscopic systems*. Springer.
- Roldán, Édgar, Izaak Neri, Meik Dörpinghaus, Heinrich Meyr, and Frank Jülicher (2015). “Decision making in the arrow of time”. In: *Physical review letters* 115.25, p. 250602.
- Ross, Stéphane, Joelle Pineau, Brahim Chaib-draa, and Pierre Kreitmann (2011). “A Bayesian approach for learning and planning in partially observable Markov decision processes”. In: *The Journal of Machine Learning Research* 12, pp. 1729–1770.
- Rubin, Jonathan, Ohad Shamir, and Naftali Tishby (2012). “Trading value and information in MDPs”. In: *Intelligent Systems Reference Library* 28, pp. 57–74.
- Russell, Stuart (1995). “Rationality and intelligence”. In: *Proceedings of the 14th international joint conference on Artificial intelligence-Volume 1*. Morgan Kaufmann Publishers Inc., pp. 950–957.
- Russell, Stuart J and Devika Subramanian (1995). “Provably bounded-optimal agents”. In: *Journal of Artificial Intelligence Research* 2, pp. 575–609.
- Sacheli, Lucia Maria, Matteo Candidi, Enea Francesco Pavone, Emmanuele Tidoni, and Salvatore Maria Aglioti (2012). “And yet they act together: interpersonal perception modulates visuo-motor interference and mutual adjustments during a joint-grasping task.” eng. In: *PLoS One* 7.11, e50223.
- Saida, Masao, José Ramón Medina, and Sandra Hirche (2012). “Adaptive attitude design with risk-sensitive optimal feedback control in physical human-robot interaction”. In: *2012 IEEE RO-MAN: The 21st IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, pp. 955–961.
- Sanborn, Adam N, Thomas L Griffiths, and Daniel J Navarro (2010). “Rational approximations to rational models: alternative algorithms for category learning.” In: *Psychological review* 117.4, p. 1144.
- Savage, Leonard J (1954). *The foundations of statistics*. John Wiley & Sons.
- Schmeidler, David (1989). “Subjective probability and expected utility without additivity”. In: *Econometrica: Journal of the Econometric Society*, pp. 571–587.
- Sebanz, Natalie, Harold Bekkering, and Guenther Knoblich (2006). “Joint action: bodies and minds moving together.” In: *Trends Cogn Sci* 10.2, pp. 70–76.

-
- Seifert, Udo (2005). “Entropy production along a stochastic trajectory and an integral fluctuation theorem.” In: *Physical review letters* 95.4, p. 040602.
- Shadmehr, Reza and John W Krakauer (2008). “A computational neuroanatomy for motor control”. In: *Experimental Brain Research* 185.3, pp. 359–381.
- Shadmehr, Reza and Ferdinando A Mussa-Ivaldi (1994). “Adaptive representation of dynamics during learning of a motor task”. In: *The Journal of Neuroscience* 14.5, pp. 3208–3224.
- Shapley, Lloyd (1964). “Advances in Game Theory”. In: ed. by L.S. Shapley M. Dresher and A.W. Tucker. Princeton University Press. Chap. Some Topics in Two-Person Games.
- Shen, Yun, Michael J Tobia, Tobias Sommer, and Klaus Obermayer (2014). “Risk-sensitive reinforcement learning”. In: *Neural computation* 26.7, pp. 1298–1328.
- Shenhav, Amitai, Matthew M Botvinick, and Jonathan D Cohen (2013). “The expected value of control: an integrative theory of anterior cingulate cortex function”. In: *Neuron* 79.2, pp. 217–240.
- Simon, Herbert A (1955). “A behavioral model of rational choice”. In: *The quarterly journal of economics*, pp. 99–118.
- (1979). “Rational decision making in business organizations”. In: *The American economic review*, pp. 493–513.
- Sims, Christopher A (2003). “Implications of rational inattention”. In: *Journal of monetary Economics* 50.3, pp. 665–690.
- Slovic, Paul and Amos Tversky (1974). “Who accepts Savage’s axiom?” In: *Behavioral science* 19.6, pp. 368–373.
- Smith, Kip, John Dickhaut, Kevin McCabe, and José V Pardo (2002). “Neuronal substrates for choice under ambiguity, risk, gains, and losses”. In: *Management Science* 48.6, pp. 711–718.
- Start, K. B. (1963). “Overestimation of Personal Abilities and Success at First-Year University Examinations”. In: *The Journal of Social Psychology* 59.2, pp. 337–345.
- Stevenson, Ian H., Hugo L. Fernandes, Iris Vilares, Kunlin Wei, and Konrad P. Körding (2009). “Bayesian Integration and Non-Linear Feedback Control in a Full-Body Motor Task”. In: *PLoS Comput Biol* 5.12, e1000629.
- Still, Susanne (2009). “An information-theoretic approach to interactive learning”. In: *Europhysics Letters* 85, p. 28005.
- Still, Susanne and Doina Precup (2012). “An information-theoretic approach to curiosity-driven reinforcement learning”. In: *Theory in Biosciences* 131.3, pp. 139–148.
- Still, Susanne, David A Sivak, Anthony J Bell, and Gavin E Crooks (2012). “Thermodynamics of prediction”. In: *Physical Review Letters* 109.12, p. 120604.
- Strehl, Alexander L, Lihong Li, and Michael L Littman (2009). “Reinforcement learning in finite MDPs: PAC analysis”. In: *The Journal of Machine Learning Research* 10, pp. 2413–2444.
- Sutton, Richard S and Andrew G Barto (1998). *Reinforcement learning: An introduction*. MIT press.
-

- Szita, István and András Lőrincz (2008). “The many faces of optimism: a unifying approach”. In: *Proceedings of the 25th international conference on Machine learning*. ACM, pp. 1048–1055.
- Szita, István and Csaba Szepesvári (2010). “Model-based reinforcement learning with nearly tight exploration complexity bounds”. In: *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pp. 1031–1038.
- Theodorou, Evangelos, Jonas Buchli, and Stefan Schaal (2010). “A generalized path integral control approach to reinforcement learning”. In: *The Journal of Machine Learning Research* 9999, pp. 3137–3181.
- Thrun, Sebastian, John Langford, and Vandi Verma (2002). “Risk Sensitive Particle Filters”. In: *Advances in neural information processing systems 14*. MIT Press.
- Tin, Chung and Chi-Sang Poon (2005). “Internal models in sensorimotor integration: perspectives from adaptive control theory”. In: *Journal of Neural Engineering* 2.3, S147.
- Tishby, Naftali and Daniel Polani (2011). “Information theory of decisions and actions”. In: *Perception-action cycle*. Springer, pp. 601–636.
- Tobler, Philippe N, John P O’Doherty, Raymond J Dolan, and Wolfram Schultz (2007). “Reward value coding distinct from risk attitude-related uncertainty coding in human reward systems”. In: *Journal of neurophysiology* 97.2, pp. 1621–1632.
- Todorov, Emanuel (2004). “Optimality principles in sensorimotor control.” eng. In: *Nature Neuroscience* 7.9, pp. 907–915.
- (2005). “Stochastic optimal control and estimation methods adapted to the noise characteristics of the sensorimotor system.” eng. In: *Neural Computation* 17.5, pp. 1084–1108.
- (2006). “Linearly-solvable Markov decision problems”. In: *Advances in neural information processing systems*, pp. 1369–1376.
- (2009). “Efficient computation of optimal actions”. In: *Proceedings of the national academy of sciences* 106.28, pp. 11478–11483.
- Todorov, Emanuel and Michael I Jordan (2002). “Optimal feedback control as a theory of motor coordination”. In: *Nature neuroscience* 5.11, pp. 1226–1235.
- Trommershäuser, Julia, Laurence T Maloney, and Michael S Landy (2003a). “Statistical decision theory and the selection of rapid, goal-directed movements”. eng. In: *J Opt Soc Am A* 20.7, pp. 1419–33.
- (2003b). “Statistical decision theory and trade-offs in the control of motor response”. In: *Spatial vision* 16.3, pp. 255–275.
- (2008). “Decision making, movement planning and statistical decision theory”. In: *Trends in cognitive sciences* 12.8, pp. 291–297.
- Turnham, Edward J. A., Daniel A. Braun, and Daniel M. Wolpert (2011). “Inferring Visuo-motor Priors for Sensorimotor Learning”. In: *PLoS Comput Biol* 7.3, e1001112.
- Tversky, Amos and Daniel Kahneman (1975). “Judgment under uncertainty: Heuristics and biases”. In: *Utility, probability, and human decision making*. Springer, pp. 141–162.
- (1981). “The framing of decisions and the psychology of choice”. In: *Science* 211.4481, pp. 453–458.
-

-
- (1992). “Advances in prospect theory: Cumulative representation of uncertainty”. In: *Journal of Risk and uncertainty* 5.4, pp. 297–323.
- Al-Ubaydli, O., G. Jones, and J. Weel (2011). *Patience, Cognitive Skill and Coordination in the Repeated Stag-Hunt*. Tech. rep. George Mason University.
- Valdesolo, Piercarlo, Jennifer Ouyang, and David DeSteno (2010). “The rhythm of joint action: Synchrony promotes cooperative ability”. In: *Journal of Experimental Social Psychology* 46.4, pp. 693–695.
- Vesper, Cordula, Robrecht P R D. van der Wel, Guenther Knoblich, and Natalie Sebanz (2013). “Are you ready to jump? Predictive mechanisms in interpersonal coordination.” eng. In: *J Exp Psychol Hum Percept Perform* 39.1, pp. 48–61.
- Vijayakumar, Konrad Rawlik, Marc Toussaint, and Sethu (2012). “On Stochastic Optimal Control and Reinforcement Learning by Approximate Inference”. In: *Proceedings of Robotics: Science and Systems*. Sydney, Australia.
- Von Neumann, John and Oskar Morgenstern (1944). “Theory of games and economic behavior.” In:
- Vul, Edward, Noah Goodman, Thomas L Griffiths, and Joshua B Tenenbaum (2014). “One and done? Optimal decisions from very few samples”. In: *Cognitive science* 38.4, pp. 599–637.
- Wald, Abraham (1945). “Statistical decision functions which minimize the maximum risk”. In: *Annals of Mathematics*, pp. 265–280.
- Weibull, Jörgen W (1997). *Evolutionary game theory*. MIT press.
- Welford, Alan Traviss (1952). “The “psychological refractory period” and the timing of high-speed performance—a review and a theory”. In: *British Journal of Psychology. General Section* 43.1, pp. 2–19.
- Whittle, Peter (1981). “Risk-sensitive linear/quadratic/Gaussian control”. In: *Advances in Applied Probability*, pp. 764–777.
- (1990). “Risk-sensitive optimal control”. In: *John Wiley & Sons*.
- Wiesemann, Wolfram, Daniel Kuhn, and Berç Rustem (2013). “Robust Markov decision processes”. In: *Mathematics of Operations Research* 38.1, pp. 153–183.
- Wolpert, Daniel M, Zoubin Ghahramani, and Michael I Jordan (1995). “An internal model for sensorimotor integration”. In: *Science* 269.5232, p. 1880.
- Wolpert, Daniel M. and Michael S. Landy (2012). “Motor control is decision-making.” In: *Current opinion in neurobiology* 22.6, pp. 996–1003.
- Wolpert, David H (2006). “Information Theory-The Bridge Connecting Bounded Rational Game Theory and Statistical Physics”. In: *Complex Engineered Systems*. Springer, pp. 262–290.
- Wu, Shih-Wei, Mauricio R Delgado, and Laurence T Maloney (2009). “Economic decision-making compared with an equivalent motor task”. In: *Proceedings of the National Academy of Sciences* 106.15, pp. 6088–6093.
- Yoshida, Wako, Ray J. Dolan, and Karl J. Friston (2008). “Game Theory of Mind”. In: *PLoS Comput Biol* 4.12, e1000254+.
-

- Zilberstein, Shlomo (1996). “Using anytime algorithms in intelligent systems”. In: *AI magazine* 17.3, p. 73.
- (2008). “Metareasoning and bounded rationality”. In: *Metareasoning: Thinking about Thinking*, MIT Press, *forthcoming*.