# Risk Sensitivity under Partially Observable Markov Decision Processes

**Nikolas Höft (nikolas.hoeft@gmail.com)[1], Rong Guo (rong.guo@tu-berlin.de)[1], Vaios Laschos (vaios.laschos@tu-berlin.de)[1], Sein Jeung (sein.jeung@campus.tu-berlin.de)[2], Dirk Ostwald (dirk.ostwald@fu-berlin.de)[2], and Klaus Obermayer (klaus.obermayer@tu-berlin.de)[1]**

[1] *Institute of Software Engineering and Theoretical Computer Science, Technische Universität Berlin, Marchstrasse 23, 10587 Berlin, Germany*
[2] *Computational Cognitive Neuroscience, Freie Universität Berlin, Habelschwerdter Allee 45, 14195 Berlin, Germany*

## Abstract

**Many real-life decisions must be made in the face of risk that is due to uncertain information about the environment. Even facing the same environment, different people might behave differently due to their individual risk preferences. For instance, a risk-seeking gambler may overestimate the chance of favorable outcomes or the amount of money going to win in those cases and therefore prefers to gamble. In cognitive neuroscience, Bayesian inference is usually applied to model the objective perception of the unobservable state, under which risk-neutral decisions are made by solving a partially observable Markov decision process (POMDP). However, the subjective evaluation of such inferred state information, which leads to different individual risk preferences, and the underlying neurobiological process are still poorly understood. Hence, we derived a risk-sensitive POMDP method that models human choice behavior and response time in a simulated investment task. Our risk-sensitive POMDP model fits the experimental data considerably better than the risk-neutral model. The model's risk-sensitivity parameters explained subjects' individual risk preference under state uncertainty at the decision time. Our results may pave the way for understanding human risk-sensitive choice under perceptual uncertainty using a unified quantitative POMDP framework.**

**Keywords:** risk-sensitive POMDP; decision-making; perceptual uncertainty; reward

## Introduction

Many real-life decisions are made in the twilight of uncertainty, such as whether to invest in a risky asset. At least two types of uncertainty can impact the economic consequences of a choice and thus result in decision risk (Bach & Dolan, 2012): first, the uncertain consequences of the decision maker's choice, and second, the decision maker's uncertain knowledge about the state underlying the choice situation. While the neural mechanisms underlying economic risk processing are fairly well established (Niv, Edlund, Dayan, & O'Doherty, 2012), the risk preferences induced by perceptual uncertainty are less clear. In this study, we generalized the recent computational work on risk-sensitive Markov decision processes (MDPs) (Shen, Tobia, Sommer, & Obermayer, 2014) to the POMDP case and empirically validated this theoretical framework as a behavioral model for human response times (RT) and choices in a novel experiment, in which human subjects performed a simulated investment game. The behavioral task was designed to flexibly manipulate individual risk preferences induced by perceptual uncertainty. We identified subject groups of similar risk-preferences and demonstrated each group's belief update about the unobservable states using the risk-sensitive POMDP model.

## Method

### Experimental Paradigm: Human Risk-sensitive Choice under Perceptual Uncertainty

55 participants (32 female, mean age 25.44 $\pm$ 4.7 years old) performed a sequential decision task in which they imagined themselves as an investor in a simulated stock market. The market had two unobservable states, a "good" state with a high investment return and a "bad" state with a low investment return. The good (bad) state was indicated by the left (right) motion direction of the random dot kinematogram (RDK) (Britten, Shadlen, Newsome, & Movshon, 1993) stimulus that consisted of 180 frames with 1/60 seconds per frame. The stimulus switched its direction once within a trial at a random time point, which induced uncertain economic consequences of subject's actions. The perception about motion strength was induced by the probability that a particular dot would be displaced in the signal direction, which is typically referred to as coherence. At each frame, the participant chose between two possible actions, "sell" the stock or "wait". Selling in a good state led to a reward of 2.5 units, whereas in a bad state led to a reward of 1 unit. The wait action allowed to accumulate information at the expense of a small constant waiting cost per frame. The episode terminated either immediately after the subject chose to sell or automatically set to sell at the last frame if the subjects waited until the end of the episode. As an optimal strategy, a participant who believed to be in the good state should sell the stock as quickly as possible. On the other hand, when the participant believed to be in the bad state, they could either wait for the market state to switch to the good state for a larger profit or sell immediately to avoid further waiting costs. However, the accumulated waiting costs could be higher than the profit of the good state if the switch happened too late. The risk in this task arises as a consequence of the decision under perceptual uncertainty about the unobservable market state. Details about the 2 x 2 x 2 factorial experimental design with the order of stimulus states (good first , bad first), the coherence of both motion states (high, low), and the waiting cost (high, low) are shown in Figure 1. Each participant performed 60 trials under each of the eight

experimental conditions (480 trials in total).

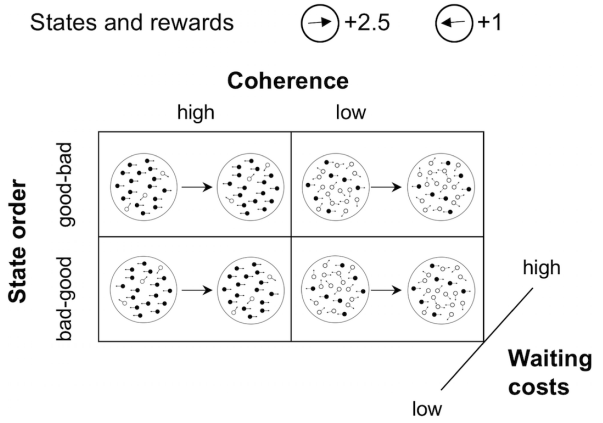


Figure 1: The 2 x 2 x 2 factorial experimental design of the simulated investment task.

## Computational Modeling: Risk-sensitive POMDP

The POMDP representation of the investment game is given by a tuple $(S, A, \Omega, T, O, R)$ with the following components:

- $S := \{good_{pre}, bad_{pre}, good_{post}, bad_{post}, sold\}$ is the unobservable state space [1]
- $A := \{wait, sell\}$ is the action space
- $\Omega$ is a set of observations given by the noisy RDK stimulus
- $T : S \times A \times S \to [0,1]$ is a state transition function
- $O : S \times A \times \Omega \to [0,1]$ is an observation function
- $R : S \times A \to \mathbb{R}$ is a reward function

The RDK stimulus represented the noisy observations that the agent received from the POMDP environment (displayed in Figure 2). In each trial, the duration of the RDK stimulus lasted at maximum N=180 time steps (frames).

The resulting belief-state MDP has a belief space $B$ which is the set of all probability distributions over the state space $S$. Bayesian inference is used to update the belief upon receiving each new observation:

$$\begin{aligned} b'(s') &= \frac{P(o|s',a)\sum_{s\in S} P(s'|a,s)b(s)}{P(o|a,b)} \\ &= \frac{O(s',a,o)\sum_{s\in S} T(s,a,s')b(s)}{P(o|a,b)} \end{aligned} \tag{1}$$

where $P(o|a,b) = \sum_{s'\in S} O(s',a,o)\sum_{s\in S} T(s,a,s')b(s)$.

[1] Due to the unique, single state transitioning in each episode, both the good and the bad state can be formally represented by a pre and a post state.

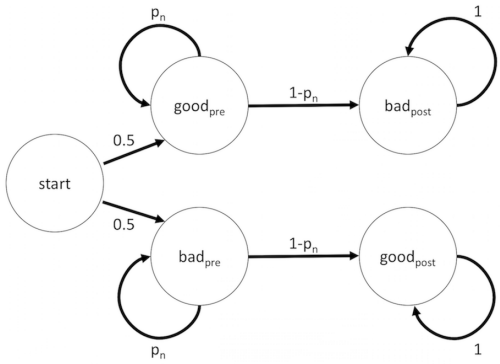


Figure 2: Transition dynamics of the unobservable state space in the experiment for the wait action.

**Risk-neutral model** A risk-neutral agent aims at maximizing the expected cumulative reward through a policy $\pi$:

$$J_N(\pi, b) := \max_{\pi\in\Pi} \mathbb{E}^{\pi}\left[\sum_{n=0}^{N} R_n | b_0 = b\right] \tag{2}$$

The optimal policy $\pi^* := \arg\max_{\pi\in\Pi} J_N(\pi, b)$ can be obtained using standard value iteration in the belief space $B$ with an appropriate approximation method (e.g. Spaan, 2012).

**Risk-sensitive model** To incorperate risk-sensitivity into the model, the agent is endowed with an exponential utility function, given by

$$U(x) = \begin{cases} \frac{1}{\lambda}(1 - e^{-\lambda x}), & \lambda \neq 0 \\ x, & else \end{cases} \tag{3}$$

where $\lambda$ controls the risk-preferences. When $\lambda < 0$ ($\lambda > 0$), $U$ is convex (concave) and therefore the agent will be risk-seeking (risk-averse). For risk-sensitive POMDPs, the state space is expanded by the agent's wealth $w \in W$ (i.e. its cumulative reward at any given time) (Bäuerle & Rieder, 2017; Marecki & Varakantham, 2010). The objective of the risk-sensitive agent is then given by:

$$J_N(\pi, b) := \max_{\pi\in\Pi} \mathbb{E}^{\pi}\left[U(w_0 + \sum_{n=0}^{N} R_n | b_0 = b)\right] \tag{4}$$

Value iteration can be performed by recursively calculating $V_U^n(b, w)$ starting at the final decision epoch $N$ where it must hold that:

$$\begin{aligned} V_U^N(b, w) &= U(w) \\ &= U((N-1)c + r_s) \end{aligned} \tag{5}$$

for all $b \in B$. Here, $r_s \in \{1, 2.5\}$ is short-hand for the reward for selling in state $s$ and $c$ denotes the waiting costs.

Accordingly, the optimal state-action values can be calculated via backward recursion by:

$$
\begin{aligned}
V_U^n(b,w) &= \max_{a\in A}\Big\{\sum_{o\in\Omega} P(o|b,a)V_U^{n+1}(b',w+R(b,a))\Big\} \\
&= \max\Big\{\sum_{o\in\Omega} P(o|b,a=wait)V_U^{n+1}(b',nc), \\
&\qquad \sum_{s\in S} b(s)U((n-1)c+r_s)\Big\}
\end{aligned} \tag{6}
$$

where $b'$ is the updated belief and $R(b,a)$ is the expected reward under action $a$ with respect to the current belief b. An approximation to the optimal value function was obtained using grid-based approximation with nearest-neighbor interpolation (Hauskrecht, 2000). Exploratory analysis of a risk-sensitive agent's choice behavior showed that negative values of $\lambda$ corresponded to quicker responses at the cost of higher state uncertainty at decision time, whereas positive values of $\lambda$ induced longer evidence accumulation and thus longer RTs on average.

### Model-based Analysis

Both risk-neutral and risk-sensitive models are fitted to the subjects' behavioral RT. We identified subject groups with similar risk-preferences by applying k-Means clustering to their RT quantiles. Goodness of fit was determined by measuring similarity between the humans' and the model agent's RT distributions based on the Euclidean distance between the quantiles. For sets of candidate values, $\lambda \in \Lambda, coh_{low}, coh_{high} \in C$ we fit parameters by the following procedure:

Let $\Theta := \Lambda \times C \times C$ denote the parameter space. Furthermore $q^{low}$ and $q^{high}$ denote the vectorial data representation of the RT distribution quantiles corresponding to the experimental conditions with low and high coherence, respectively. The optimal set of parameters, $\theta^*$ is then given by [2]:

$$
\begin{aligned}
\theta^* = \arg\min_{\theta\in\Theta} &\left\| q^{low}_{subjects} - q^{low}_{agent(\lambda,coh_{low})} \right\| + \\
&\left\| q^{high}_{subjects} - q^{high}_{agent(\lambda,coh_{high})} \right\|
\end{aligned} \tag{7}
$$

The analysis was performed on both, group level and for individual subjects [3].

## Results

The group-wise cumulative RT distributions are shown in Figure 3 for each experimental condition. The clearly distinct group patterns supported the assumption that the risk-sensitivity towards perceptual uncertainty guides subjects' choice behavior. For example, *Group 1* prefers to accumulate a lot of evidence in order to reduce perceptual uncertainty at decision time. Conversely, *Group 2* sells very quickly on average, thus avoiding waiting costs at the expense of higher perceptual uncertainty.

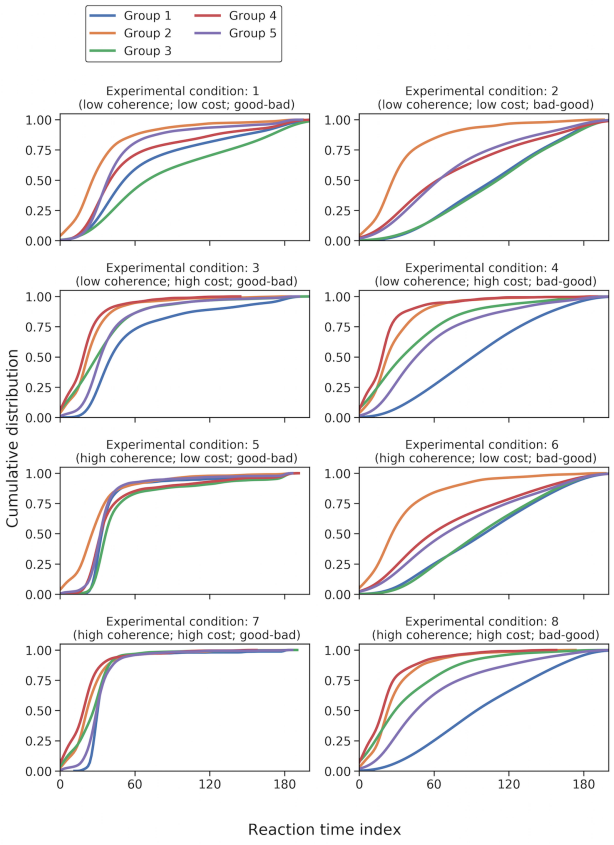


Figure 3: Group RT distributions for human subjects. The groups were determined by k-Means clustering.

The group-wise goodness-of-fit and best-fitting risk-sensitivity parameters are shown in Table 1. The overall lower values of the fitting criterion from the risk-sensitive model showed that it explained the behavioral data better than the risk-neutral model. This was particularly evident for groups favoring rather extreme choice strategies, either risk-averse (e.g., Group 1) or risk-seeking (e.g., Group 2). However, subjects' choice response time distribution depended on the combined manipulation of waiting cost, coherence level and state orders. The simuluated response time were parameterized by both the risk-sensitive parameter $\lambda$ and the Gaussian coherence paramter. Therefore, larger (smaller) $\lambda$ alone does not necessarily lead to earlier selling (longer waiting). We further visualized each group's choice behavior by the fraction of selling in the good state across all experiemntal conditions. The corresponding risk-sensitive agents that were fitted to every group replicated the choice behavior closely (Figure 4).

Subject-wise fitting scores under the risk-neutral vs. the

[2] Coherence was modeled by the degree of overlap of two Gaussian observation distributions. The means were placed symmetrically around zero and overlap was controlled by the standard deviation, which was fitted to the data.

[3] For calculating the agent's RTs, we used a reduced waiting cost of c = 0.009 in the high-cost conditions, because the belief-action value under Bayes-optimality at the first decision epoch was always higher for the sell action

risk-sensitive models are visualized in Figure 5. The results from the individual analysis further demonstrated that the risk-sensitive model fitted the experimental data considerably better than the risk-neutral model.

| Group | Size | λ | RT quantile distance |
|---|---|---|---|
| 1 | 12 | 2.0 (0) | 452.5 (515.2) |
| 2 | 9 | -1.0 (0) | 401.9 (635.3) |
| 3 | 12 | 0.1 (0) | 484.7 (495.8) |
| 4 | 7 | -0.3 (0) | 292.0 (304.0) |
| 5 | 15 | -0.5 (0) | 400.8 (450.1) |

Table 1: The best-fitting λ and corresponding fitting criterion of the risk-sensitive vs. risk-neutral (in brackets) model by each RT group.

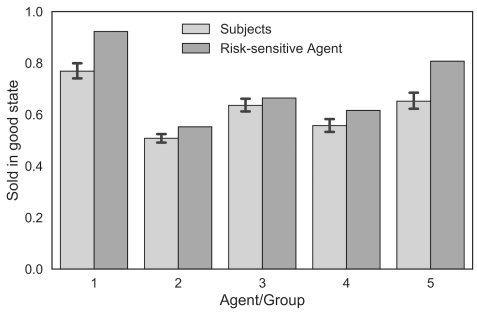


Figure 4: Fraction of selling in the good state, aggregated across all trials. Bar and error bar represent the mean and the standard deviation of the group members, respectively.

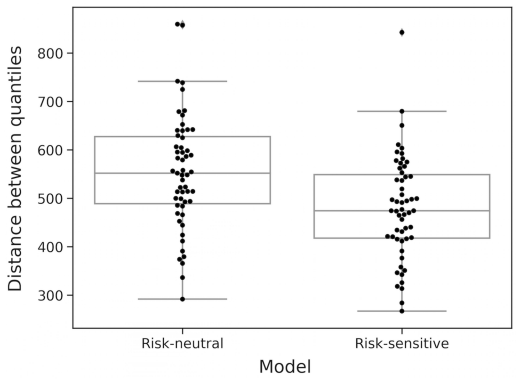


Figure 5: Distribution of subject-wise Euclidean distances of RT quantiles for risk-neutral and risk-sensitive agents.

## Discussions

Our results provide evidence favoring the risk-sensitive POMDPs for modeling choice behavior compared to the risk-neutral model. Risk-preferences under perceptual uncertainty are reflected in the parameters of the best-fitting risk-sensitive POMDP model. Individual risk-preferences were identified by their differential RT distributions. The RT distribution of the risk-sensitive model agents with varying parameters resembled the subjects' choice behavior, especially with respect to the policies of waiting long (accumulating evidence) or selling quickly (avoiding waiting costs). In summary, our study demonstrated that the concepts derived for risk-sensitive planning under economic uncertainty can be carried over to perceptual uncertainty at the within-trial level for describing behavioral RT. A follow-up study at the between-trial level with both behavioral and neuroimaging experiments could yield further insights into whether the neural correlates of risk-sensitive reinforcement learning (Niv et al., 2012; Shen et al., 2014) are also involved in the acquisition of risk-sensitive decision policies under perceptual uncertainty.

## References


Bach, D. R., & Dolan, R. J. (2012). Knowing how much you don't know: a neural organization of uncertainty estimates. *Nature Reviews Neuroscience*, *13*(8), 572–586.

Britten, K. H., Shadlen, M. N., Newsome, W. T., & Movshon, J. A. (1993). Responses of neurons in macaque MT to stochastic motion signals. *Visual Neuroscience*, *10*(6), 1157–1169.

Bäuerle, N., & Rieder, U. (2017). Partially Observable Risk-Sensitive Markov Decision Processes. *Mathematics of Operations Research*, *42*(4), 1180–1196.

Hauskrecht, M. (2000). Value-Function Approximations for Partially Observable Markov Decision Processes. *Journal of Artificial Intelligence Research*, *13*, 33–94.

Marecki, J., & Varakantham, P. (2010). Risk-Sensitive Planning in Partially Observable Environments. *In Proc. of AAMAS*, 1357–1368.

Niv, Y., Edlund, J. A., Dayan, P., & O'Doherty, J. P. (2012). Neural Prediction Errors Reveal a Risk-Sensitive Reinforcement-Learning Process in the Human Brain. *Journal of Neuroscience*, *32*(2), 551–562.

Shen, Y., Tobia, M. J., Sommer, T., & Obermayer, K. (2014). Risk-sensitive Reinforcement Learning. *Neural Computation*, *26*(7), 1298–1328.

Spaan, M. T. J. (2012). Partially Observable Markov Decision Processes. In M. Wiering & M. van Otterlo (Eds.), *Reinforcement Learning: State-of-the-Art* (pp. 387–414). Berlin, Heidelberg: Springer Berlin Heidelberg.