Honda Research Institute EU
Honda Research Institute Europe GmbH
https://www.honda-ri.de/

Model averaging as a developmental outcome of reinforcement learning

Thomas H Weisswange, Constantin Rothkopf, Tobias Rodemann, Jochen Triesch

2010

Preprint:

This is an accepted article published in [Book Title / Conference / Journal]. The final authenticated version is available online at: https://doi.org/[DOI not available]

Model averaging as a developmental outcome of reinforcement learning.

Thomas H. Weisswange, Constantin A. Rothkopf, Tobias Rodemann and Jochen Triesch

To make sense of the world, humans have to rely on the information that they receive from their sensory systems. Due to noise on one side and redundancies on the other side, it is possible to improve estimates of the signal's causes by integrating over multiple sensors. In recent years it has been shown that humans do so in a way that can be matched by optimal Bayesian models (e.g. [1]). Such an integration is only beneficial for signals originating from a common source and there is evidence that human behavior takes into account the probability for a common cause [2]. For the case in which the signals can originate from one or two sources, it is so far unclear, whether human performance is best explained by model selection, model averaging, or probability matching [3]. Furthermore, recent findings show that young children are often not integrating different modalities [4,5], indicating that this has to be learned during development. But which mechanisms are involved and how interaction with the environment could determine this process remains unclear.

Here we show that a reinforcement learning algorithm develops behavior that corresponds to cue-integration and model-averaging. The reinforcement learning agent is trained to perform an audio-visual orienting task. Two signals originate from either one or two sources and provide noisy information about the position of these objects. The agent executes orienting actions and receives rewards that are exponentially decaying with the distance from the true position. The value function is represented through a non-linear basis function network. Positions in the two stimulus dimensions are coded through Gaussian tuning curves. The weights used in the computation of the orienting action are adapted during learning using gradient descent. Actions are selected probabilistically based on the current reward predictions using the softmax-function.

The agent quickly learns to act in a way that closely approximates the behavior of a Bayesian observer. It inherently learns the reliabilities of the cues and behaves differently depending on the probability for a single cause. The agent obtains more reward than Bayesian observers that always or never integrate cues.
When we test with signals for which the behavior of model selection and model averaging differ most, the agent obtains significantly more reward than a Bayesian model selecter and matches very closely the reward obtained by the Bayesian model averager. Furthermore, when a single object is the cause for both stimuli, the variance of the distribution of chosen actions is smaller than for actions based on either of the cues alone.

Our results show that a caching reinforcement learning agent can learn when and how to do cue integration, without explicitly computing with probability distributions. Moreover the performance of this an agent is matched best by a Bayesian observer that does model averaging. This suggests that reinforcement learning based mechanisms could at least support the development of such behavior.

References:
[1]Ernst&Banks (2002) Nature 6870
[2]Körding et al. (2007) PloS One 2(9)
[3]Shams&Beierholm (2009) in Proc. Cosyne09
[4]Nardini et al. (2008) Curr. Biol. 18(9)
[5]Gori et al. (2008) Curr. Biol. 18(9)

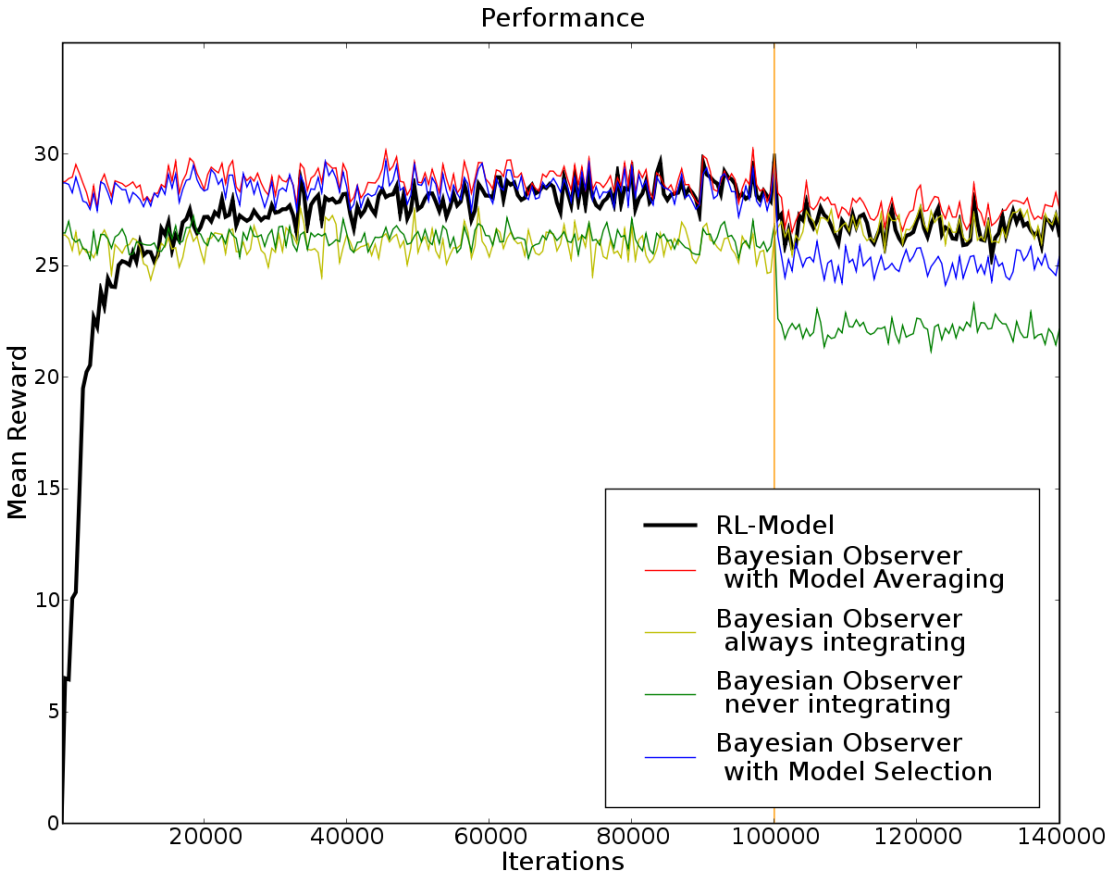

Fig. 1: Performance (averaged over 500 trials) of the reinforcement learning agent and the optimal average reward received by different Bayesian models. The performance behind the orange vertical line shows a test case restricted on inputs for which model averaging and model selection differ most.

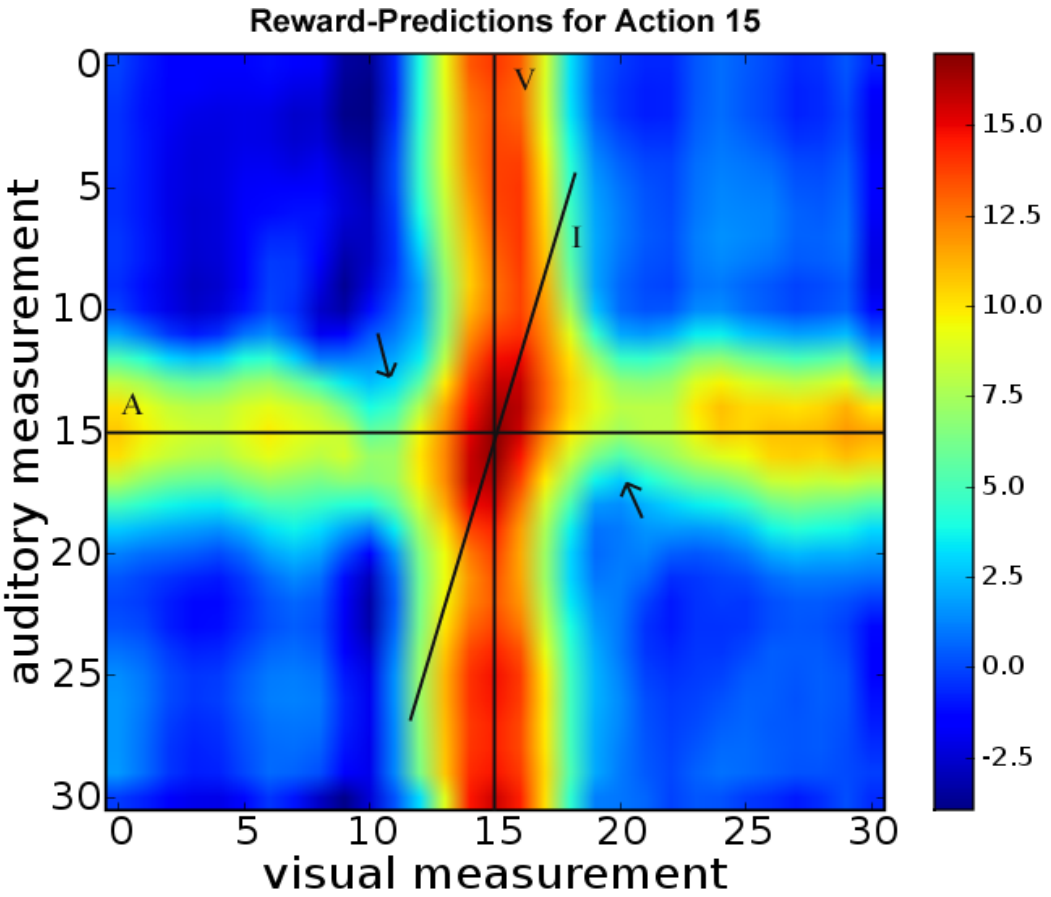

Fig. 2: Reward predictions of an exemplary action for all combinations of visual and auditory input measurements. Line "A" highlights cases where the auditory measurement is equal to the output action, line "V" does the same for the visual modality. Along line "I" the expected reward is highest due to the weighted mean of both signals preferring this output. If both measurements are either higher or lower than the actual action (Arrows), the weighted mean will favour a different action.