**Abstract**

Confidence estimation is a crucial component of decision making. Both decision confidence and the uncertainty in sensory and cognitive variables that give rise to such confidence estimates have been studied extensively in the literature (Pouget et al., 2016). These studies mostly examined confidence and uncertainty in the context of perceptual decision making, and model-free value-based decision making. Less attention has been paid to confidence judgements in model-based learning and decision making in which animal forms a mental model of the environment to guide planning and adaptive behavior. Little is known about how animals estimate the uncertainty in the environment model (e.g., state transition rules) and how model uncertainty contributes to decision confidence in model-based learning. In this work, we explore such relationships in artificial neural network agents. We extended the post-decision wagering paradigm that is used to study decision confidence (Persaud et al., 2007) to study model confidence in model-based agents using a modified version of the two-step task (Kim et al., 2019). Specifically, before the model-based agents make a final decision, we ask them to predict and report an upcoming state transition. They will receive a positive reward for a correct model judgement, a negative reward for an incorrect model judgement, and crucially, they also have a third safe-bet option that gives them a small fraction of positive reward regardless of the correctness of their model judgements. We expect our trained agent to choose the safe-bet option more when they are less confident about their learnt model. By measuring model confidence and decision confidence within the same framework, our approach allows us to study the computational mechanisms underlying how model confidence contributes to decision confidence.

**References**

Kim, D., Park, G. Y., JP, O. D., & Lee, S. W. (2019). Task complexity interacts with state-space uncertainty in the arbitration between model-based and model-free learning. *Nat Commun*, *10*(1), 5738. <https://doi.org/10.1038/s41467-019-13632-1>

Persaud, N., McLeod, P., & Cowey, A. (2007). Post-decision wagering objectively measures awareness. *Nat Neurosci*, *10*(2), 257-261. <https://doi.org/10.1038/nn1840>

Pouget, A., Drugowitsch, J., & Kepecs, A. (2016). Confidence and certainty: distinct probabilistic quantities for different goals. *Nat Neurosci*, *19*(3), 366-374. <https://doi.org/10.1038/nn.4240>