A cognitively plausible architecture for a metacognitive agent

Baptiste Pesquet

January 06, 2026

Outline

Decision-making, learning and confidence in natural cognition

Decision-making as a sequential process

- A *decision* is a deliberative process leading to a *choice.*
- Decision-makers need *time* to collect and process informative cues.
- Decision-making is often modeled as an *accumulation-to-threshold* process [1].
- The balance between response time and choice accuracy (when available) is called the *Speed/Accuracy Trade-off* [2].

Models of sequential decision-making

A popular model for binary decisions is the *Diffusion Decision Model* [3].

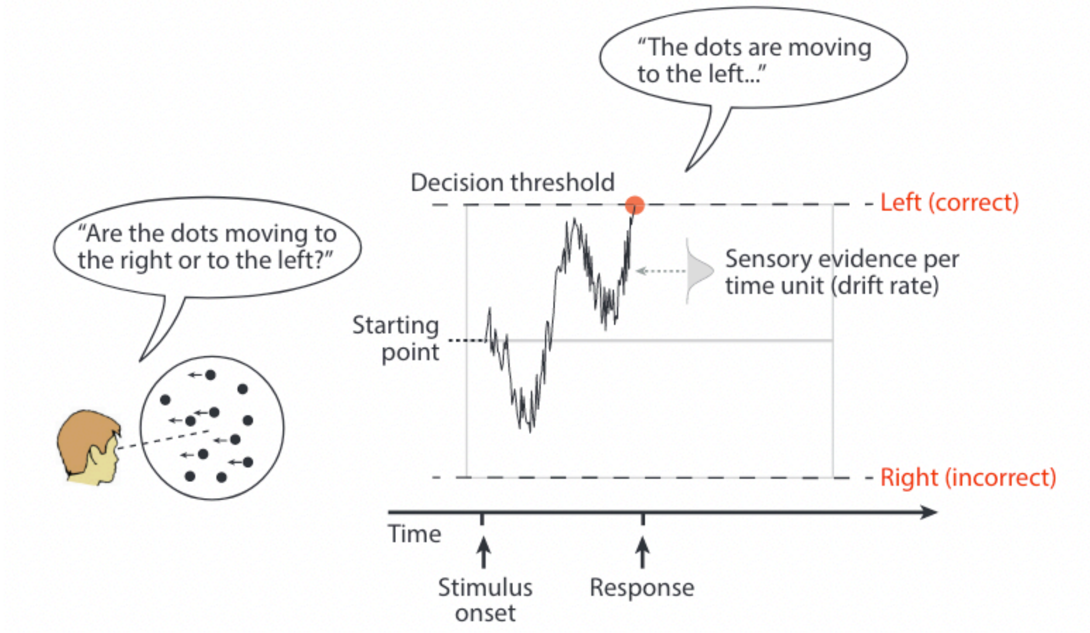

Figure 1: Illustration of the DDM model [4].

Models of sequential decision-making

In the DDM, evidence is accumulated through Equation 1.

$$\mathrm{d}x = v\,\mathrm{d}t + sW \tag{1}$$

- x: accumulated evidence.
- v: drift rate (speed of evidence accumulation).
- $\mathrm{d}t$: time unit.
- W: within-trial accumulation white noise.
- s: standard deviation of W.

Models of sequential decision-making

Multi-alternative decisions are often modeled as a *race* between accumulators, each one representing a possible choice.

Figure 2: Illustration of a race model (Source).

Learning and decision-making

- Decisions followed by *rewards* produce *learning effects*.
- Joint models of decision-making combine evidence accumulation and Reinforcement Learning to account for these effects [5].

Learning and decision-making

Q-values adjusted through delta update rule (Equation 2) are combined to drive accumulators (Equation 3) [6].

$$Q_{i,t+1} = Q_{i,t} + \alpha\left(r_t - Q_{i,t}\right) \tag{2}$$

$$\mathrm{d}x = w\left(Q_{1,t} - Q_{2,t}\right)\mathrm{d}t + sW \tag{3}$$

- $Q_{i,t}$: value representation of choice i on trial t.
- r_t: reward received on trial t.
- $\alpha \in [0, 1]$: learning rate. w: weighting factor.

Confidence in decision-making

- *Uncertainty* is inherent to all stages of neural computation [7].
- Our brain might be able to manipulate uncertainties as *probability distributions* [8].
- *Confidence* quantifies the degree of *certainty* associated with a decision.
- More formally, confidence can be defined as the *probability that a choice is correct given available evidence* [9].

Computing confidence in sequential decision-making

In *decisional focus models*, confidence is directly indexed by the state of evidence at the time of choice.

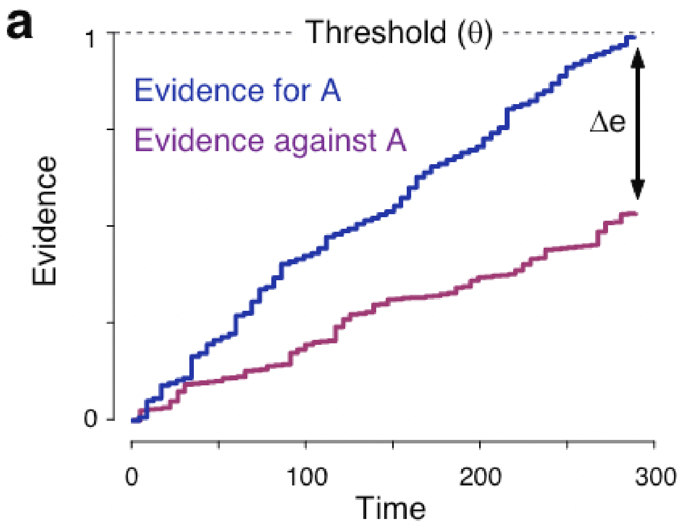

Figure 3: Computation of confidence based on the difference between accumulators at choice time [10].

Computing confidence in sequential decision-making

Post-decisional focus models posit that evidence accumulation goes on after decision time to account for confidence.

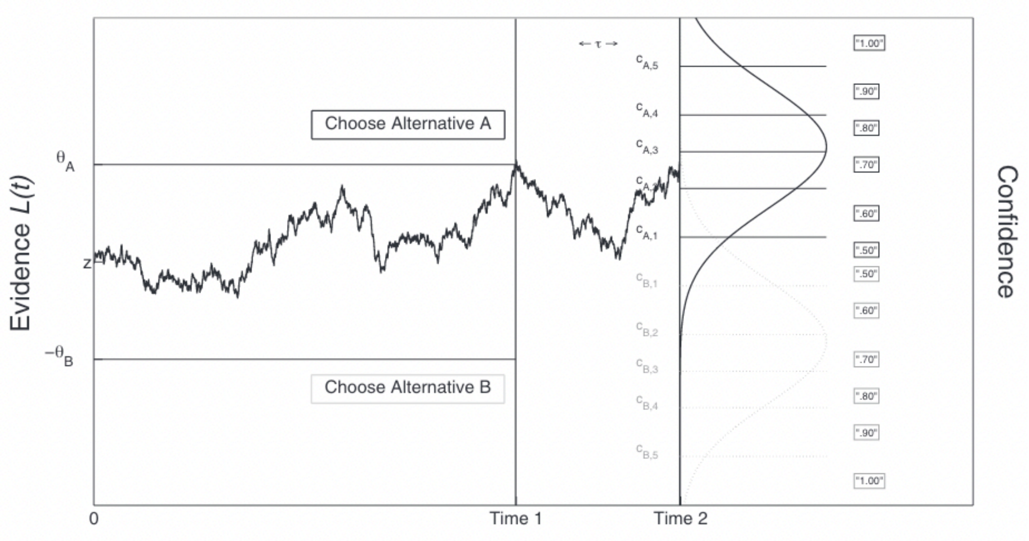

Figure 4: Computation of confidence based on post-choice accumulation [11].

Confidence as a doorway to metacognition

- *Metacognition* is the ability to monitor and regulate one's cognitive processes [12].
  - Example: should I study more or differently for an upcoming exam?
- As part of metacognitive monitoring, confidence judgments may inform the processes of *cognitive control* [13].

Building a cognitively plausible metacognitive agent

🚧

Core ideas

- Give our agent the capacity to learn a set of *behavioral rules.*
- Model decision-making by choosing between the multiple possible rules through evidence accumulation.
- Use confidence to adjust the hyperparameters of decision (e.g. accumulation thresholds) as a first step towards metacognition.

Perceptual task: Random Dot Motion discrimination

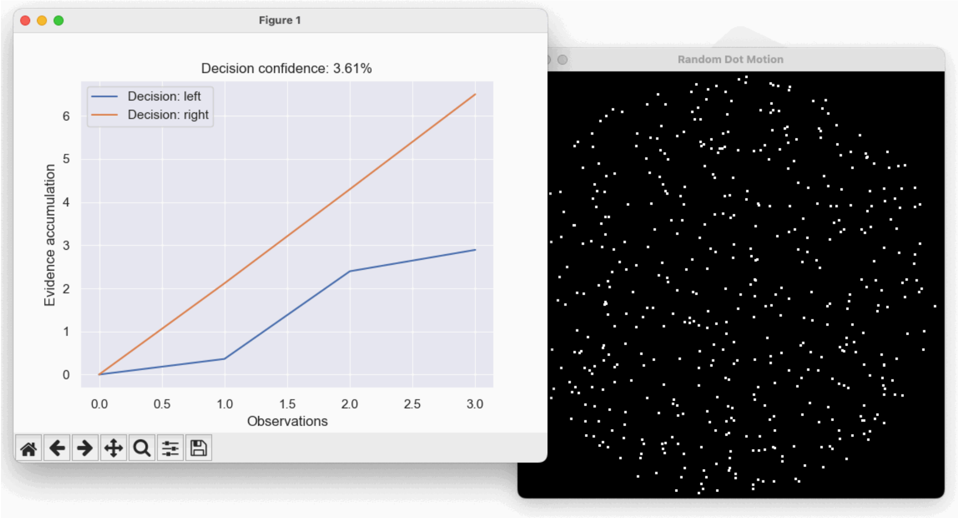

Figure 5: Evidence integration with high motion coherence [14].

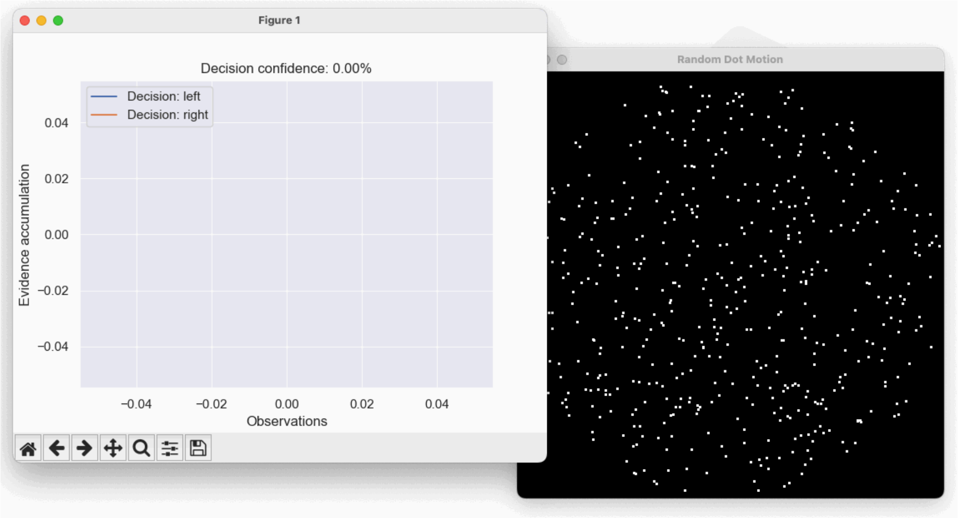

Figure 6: Evidence integration with low motion coherence [14].

Value-based task: collaborative sorting

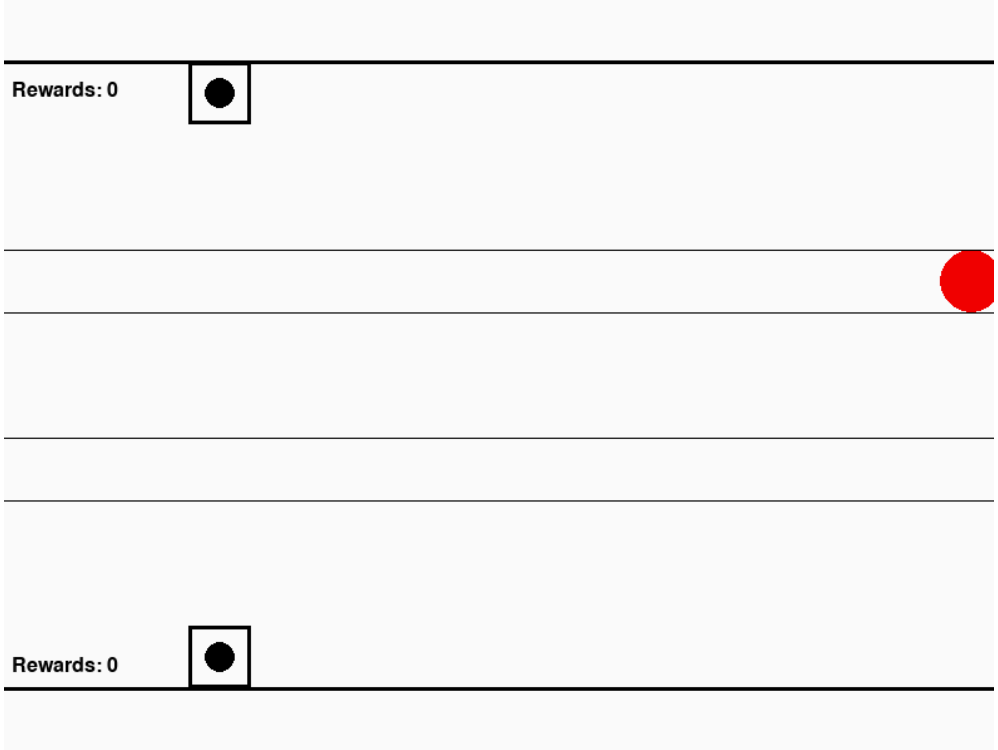

Figure 7: Two agents involved in a cooperative pick-and-place task (Source).

References

[1] J. I. Gold and M. N. Shadlen, “The Neural Basis of Decision Making,” *Annual Review of Neuroscience*, vol. 30, no. 1, pp. 535–574, July 2007, doi: 10.1146/annurev.neuro.29.051605.113038.

[2] R. P. Heitz, “The Speed-Accuracy Tradeoff: History, Physiology, Methodology, and Behavior,” *Frontiers in Neuroscience*, vol. 8, June 2014, doi: 10.3389/fnins.2014.00150.

[3] R. Ratcliff and G. McKoon, “The Diffusion Decision Model: Theory and Data for Two-Choice Decision Tasks,” *Neural Computation*, vol. 20, no. 4, pp. 873–922, Apr. 2008, doi: 10.1162/neco.2008.12-06-420.

[4] B. Forstmann, R. Ratcliff, and E.-J. Wagenmakers, “Sequential Sampling Models in Cognitive Neuroscience: Advantages, Applications, and Extensions,” *Annual Review of Psychology*, vol. 67, no. 1, pp. 641–666, Jan. 2016, doi: 10.1146/annurev-psych-122414-033645.

[5] S. Miletić, R. J. Boag, A. C. Trutti, N. Stevenson, B. U. Forstmann, and A. Heathcote, “A New Model of Decision Processing in Instrumental Learning Tasks,” *eLife*, vol. 10, p. e63055, Jan. 2021, doi: 10.7554/eLife.63055.

[6] L. Fontanesi, S. Gluth, M. S. Spektor, and J. Rieskamp, “A Reinforcement Learning Diffusion Decision Model for Value-Based Decisions,” *Psychonomic Bulletin & Review*, vol. 26, no. 4, pp. 1099–1121, Aug. 2019, doi: 10.3758/s13423-018-1554-2.

[7] S. M. Fleming, “Metacognition and Confidence: A Review and Synthesis.”

[8] F. Meyniel, M. Sigman, and Z. Mainen, "Confidence as Bayesian Probability: From Neural Origins to Behavior," *Neuron*, vol. 88, no. 1, pp. 78–92, Oct. 2015, doi: 10.1016/j.neuron.2015.09.039.

[9] A. Pouget, J. Drugowitsch, and A. Kepecs, "Confidence and Certainty: Distinct Probabilistic Quantities for Different Goals," *Nature Neuroscience*, vol. 19, no. 3, pp. 366–374, Mar. 2016, doi: 10.1038/nn.4240.

[10] A. Kepecs, N. Uchida, H. A. Zariwala, and Z. F. Mainen, "Neural Correlates, Computation and Behavioural Impact of Decision Confidence," *Nature*, vol. 455, no. 7210, pp. 227–231, Sept. 2008, doi: 10.1038/nature07200.

[11] T. J. Pleskac and J. R. Busemeyer, "Two-Stage Dynamic Signal Detection: A Theory of Choice, Decision Time, and Confidence.," *Psychological Review*, vol. 117, no. 3, pp. 864–901, 2010, doi: 10.1037/a0019737.

[12] J. Flavell, "Metacognition and Cognitive Monitoring." Oct. 1979.

[13] S. M. Fleming and H. C. Lau, "How to Measure Metacognition," *Frontiers in Human Neuroscience*, vol. 8, July 2014, doi: 10.3389/fnhum.2014.00443.

[14] B. Pesquet and F. Alexandre, "Towards Metacognitive Agents: Integrating Confidence in Sequential Decision-Making."

Thank you for your attention!

Any questions? 😀