Signal neutrality, scalar property, and collapsing boundaries as consequences of a learned multi-time scale strategy

Luca Manneschi[1☯], Guido Gigante[2,3☯*], Eleni Vasilaki[1], Paolo Del Giudice[2,3‡],

1 Department of Computer Science, University of Sheffield, Sheffield, UK

2 Istituto Superiore di Sanità, Rome, Italy

3 INFN, Sezione di Roma, Rome, Italy

☯These authors contributed equally to this work.

†Deceased

* guido.gigante@gmail.com

Abstract

Experiments and models in perceptual decision-making point to a key role of an integration process that accumulates sensory evidence over time. We endow a probabilistic agent comprising several such integrators with widely spread time scales and let it learn, by trial-and-error, to weight the different filtered versions of a noisy signal. The agent discovers a strategy markedly different from the literature "standard", according to which a decision made when the accumulated evidence hits a predetermined threshold. The agent instead decides during fleeting windows corresponding to the alignment of many integrators, akin to a majority vote. This strategy presents three distinguishing signatures. 1) Signal neutrality: a marked insensitivity to the signal coherence in the interval preceding the decision, as also observed in experiments. 2) Scalar property: the mean of the response times varies glaringly for different signal coherences, yet the shape of the distribution stays largely unchanged. 3) Collapsing boundaries: the agent learns to behave as if subject to a non-monotonic urgency signal, reminiscent in shape of the theoretically optimal. These three characteristics, which emerge from the interaction of a multi-scale learning agent with a highly volatile environment, are hallmarks, we argue, of an optimal decision strategy in challenging situations. As such, the present results may shed light on general information-processing principles leveraged by the brain itself.

Author summary

The rate of integration of sensory information prior to a decision-making process needs to be versatile and adaptable to different situations. While driving can require quick reactions, evaluating the authenticity of a painting can require long observations, and consequently the concept of representations over multiple timescales appears necessary from an intuitive perspective. Nevertheless, there is a lack of theoretical research that exploits multiple timescales, despite the presence of a variety of integration rates have been experimentally observed. In the following work, we developed a decision-making model based on integrators with multiple characteristic times and analysed its behaviour on a highly volatile and biologically relevant task. Through trial and error and reward maximisation, the model discovers an effective strategy that is surprisingly different and more robust in comparison to the more "classical", single time-scale approach. More importantly, the strategy learnt exhibits remarkable agreement with experimental findings, suggesting a fundamental role of multiple timescales for decision-making. Our model, despite being abstract, achieves a good degree of biological realism and perform robustly in different environments.

1 Introduction

Perceptual decision-making is one of the most fundamental interactions of a biological agent with its environment. It is not by chance that perceptual decision-making processes have been long studied in the context of operant conditioning [1], where an animal learns to associate choices and consequences by trial-and-error, and where imperfect performance is considered a consequence of imperfect learning or the reflex of the learning strategy itself [2].

The research on perceptual decision-making, on the other hand, has mainly focused on tasks where uncertainty (typically in the form of noisy signals) and time (for instance duration of the observation and response delays) play a pivotal role [3–6]. In such scenarios, the errors made by the subject at the end of a training phase, as well as the relevant performance metrics (*e.g.* accuracy or speed of response), are deemed informative of the cognitive mechanisms involved [7–10]. In fact, there have been numerous attempts to compare the behaviour of animal subjects to the performance of different algorithms and determine how optimal the displayed behaviour is [7, 11–15].

The present paper sits at the junction of these two traditions - learning theory and cognitive psychology. We present an artificial agent whose task is simply to determine whether a noisy signal has positive or negative mean value. This problem represents an idealised version of tasks often used in decision making experiments (*e.g.*, random dots [16–18]). The agent integrates the noisy signal over multiple time scales and takes a decision in a probabilistic manner. Over many task repetitions, by observing the consequences of its choices, the agent learns to maximise the expected reward for each presentation of the signal.

The accumulation of evidence over time is one of the key ideas emerged from the perceptual decision-making field [5, 7, 19–22]. A fundamental model across psychology and neuroscience, the drift-diffusion model (also know as ‘bounded evidence accumulation’ model), consists of two or more competing traces that accumulate sensory evidence for different choices; the first trace to hit a threshold makes the associated option the final decision [23]. The drift-diffusion model is a continuous time variant of the sequential probability ratio test [24, 25] and has a strong theoretical support: in the case of two-alternative forced choices it is optimal in selecting between two hypotheses. Despite its simplicity, this model can account for many psychophysical and neural observations, such as distribution of response times and performance when varying sensory coherence [26, 27].

Notwithstanding its success, alternatives have been proposed to the standard drift-diffusion model [7, 28, 29] to account for unexplained phenomena such as primacy and recency effects, asymptotic accuracy, and “fast errors” [30–32]. Of notable importance is the Ornstein–Uhlenbeck model, which modifies the standard drift-diffusion model by including a decay term in the dynamics of the accumulation. Although the Ornstein–Uhlenbeck model is capable to account for many experimental observations, including neurophysiological ones [28, 32], it introduces a characteristic time scale over which the model 'forgets' the past sensory information. This begs the question of how to determine the value of the relevant time scale. A common approach in the literature is to treat the time scale of the accumulation as a free parameter that is optimised to match experimental data [32]. However, such approach does not address how the time scale of the accumulator could be tuned in the first instance, particularly since different tasks may require processing information at different time scales, and thus such a tuning should be context dependent.

Here we tackle a different, but closely related question. We hypothesise that our artificial agent is equipped with many forgetful integrators of the incoming signal, of widely different characteristic times, and we ask how the agent would choose among these accumulators, or, more precisely, how it would assign a weight to each of them. In doing this, we are relying on a key result from the field of reservoir computing: the projection of an input signal over many different time scales allows, by means of a linear transformation, to implement a wide range of mappings between the input and an output signal [33, 34]. Beyond the computational advantage, such approach is consistent with the ample evidence of the coexistence of many time scales in brain functionality [35–39], even at the single neuron level [40–42].

Our proposed agent combines weighted sums of its accumulators to estimate, instant by instant, the optimal probability of choosing one of the two options ('move right', 'move left') or whether to wait to accumulate more evidence. Unlike the standard drift-diffusion model, there is no fixed threshold for a decision, but instead the decision time is a function of the accumulated evidence. Besides being arguably more biologically plausible, a probabilistic decision-making mechanism has clear computational advantages, for instance, it allows for better balancing of exploration and exploitation and better strategies to deal with unpredictable or antagonistic environments ([43], Chapter 13). In this context, the learnable parameters are the weights assigned to each accumulator throughout the repetition of the task. Learning takes place within the framework of reinforcement learning [43, 44]. The agent is not told which actions to take but instead must discover which actions yield the highest reward by trial-and-error. Though reinforcement learning has been applied in the context of perceptual decision-making [9, 45–47], to our knowledge this is the first time that reinforcement learning is used to optimise the evidence accumulation over multiple time scales.

There are three key characteristics in decision making processes, as suggested by psychophysical experiments and models: signal neutrality, the scalar property, and collapsing boundaries. In what follows, we will discuss how the strategy discovered by the agent during the training phase, while strikingly different to the one suggested by the drift-diffusion model, fits well the relevant psychophysical results.

With signal neutrality we denote the insensitivity, on part of the accumulation process, to the signal's mean value, in the interval preceding the decision. Such insensitivity has been repeatedly observed in the lateral intraparietal cortex during a motion-discrimination task [4, 48]. This characteristic trend is consistently found in the agent, but not in the single accumulators.

Secondly, the agent displays a distribution of response times whose shape is invariant for different mean values of the signal, notwithstanding a corresponding wide change in the average response time. This is the scalar property, as extensively reported in the temporal cognition [49], multistable perception [50], and even (with some caveats) perceptual decision making literature [51].

By looking at the internal workings of the agent after the training phase, we will conclude that signal neutrality and the scalar property have a common origin in how the agent leverages the multiple time scales at its disposal. Notably, the agent tries to achieve an integration process whose average scales like a power-law in time, but with a nearly-constant variance.

The collapsing boundaries (also referred to as urgency signal [52]) are introduced to push drift-diffusion models towards a decision even in more difficult or uncertain cases, by making the decision threshold a decreasing function of time. We will show how the agent learns to behave, thanks to the multiple time scales at its disposal, as if subject to non-monothonic boundaries (collapsing for later response times), having a shape reminiscent of the ones proposed in the literature [8, 10] for solving perceptual decision problems like the one confronted here.

These three characteristics (signal neutrality, scalar property, collapsing boundaries), we argue, are hallmarks of an optimal strategy to make decisions in a highly volatile environment. Our model proposes that these characteristics are consistent with a multi-scale agent that makes use of a simple learning rule, as such it could be implemented in the brain itself, at least in its fundamental determinants, obviously disregarding many layers of complexity. From this perspective, we believe that the present model, despite its simplicity and abstraction from a biological implementation, could shed light on general information-processing principles that can be possibly leveraged by the brain.

2 Results

2.1 Model description

Inspired by classical random dots experiments, we model a two-alternative forced-choice task as a decision over the sign of the mean value of a noisy signal $\mathrm{s}(t)$ (see Fig. 1). The signal (black line) consists of independent samples from a Gaussian distribution of mean μ and standard deviation σ, each drawn every time step $\Delta t = 10$ ms.

The agent is not required to decide at a prescribed time, it has the option to wait and then see another sample, or to perform one of two actions, 'left' and 'right', respectively associated with the decision $\mu < 0$ and $\mu > 0$, at each step. When an action is made, the episode ends, and a reward is delivered only if the agent correctly guessed the sign of μ; otherwise, the agent receives nothing. Each episode has a maximum duration $T_{\max}$. When $T_{\max}$ is reached, another 'wait' from the agent leads to the end of the episode and no reward is delivered.

Whilst σ is constant, the value of μ is instead re-sampled at the beginning of each episode from a Gaussian distribution $p(\mu)$ of zero mean and variance σ_μ. This second-order uncertainty makes the agent experience a wide range of values of μ, putting severely to the test its ability to generalise to episodes of varying signal-to-noise ratios.

The agent comprises $n_\tau = 10$ leaky integrators x_τ (dark blue to cyan lines in Fig. 1) that independently integrate the noisy signal over different time scales τ:

$$\dot{x}_\tau = -\frac{x_\tau - \mathrm{s}(t)}{\tau}, \tag{1}$$

and correspondingly n_τ leaky integrators $x_\tau^{\mathcal{T}}$ (yellow to red lines in Fig. 1) that integrate a constant input (a 'time signal', here valued 1), to account for the possible effects of an internal 'clock':

$$\dot{x}_\tau^{\mathcal{T}} = -\frac{x_\tau^{\mathcal{T}} - 1}{\tau}. \tag{2}$$

Both the x_τ and the $x_\tau^\mathcal{T}$ are reset to 0 at the beginning of each episode (note, therefore, that $x_\tau^\mathcal{T}(t) = 1 - \mathrm{e}^{-\frac{t}{\tau}} \geq 0$ for all t).

The τs are chosen on a logarithmic scale (*i.e.*, $\tau_i = \alpha\,\tau_{i-1}$, with α a suitable constant), with $\tau_1 = \tau_{\text{min}} = 100$ ms and $\tau_{n_\tau} = \tau_{\text{max}} = 10$ s, so as to allow the agent to accumulate information over a wide range of different time scales.

At each time step t, the agent computes six weighted sums, three for the signal $x_\tau(t)$ and three for the clock $x_\tau^\mathcal{T}(t)$. The first four of these weighted sums are related to the two possible actions:

$$\Sigma^\mathcal{S}_{\text{right}}(t) \equiv \sum_\tau w_{\text{right},\tau}\left[x_\tau(t) + \xi_\tau(t)\right] \tag{3}$$

$$\Sigma^\mathcal{T}_{\text{right}}(t) \equiv \sum_\tau w^\mathcal{T}_{\text{right},\tau}\left[x^\mathcal{T}_\tau(t) + \xi^\mathcal{T}_\tau(t)\right] + b_{\text{right}} \tag{4}$$

$$\Sigma^\mathcal{S}_{\text{left}}(t) \equiv \sum_\tau w_{\text{left},\tau}\left[x_\tau(t) + \xi_\tau(t)\right] \tag{5}$$

$$\Sigma^\mathcal{T}_{\text{left}}(t) \equiv \sum_\tau w^\mathcal{T}_{\text{left},\tau}\left[x^\mathcal{T}_\tau(t) + \xi^\mathcal{T}_\tau(t)\right] + b_{\text{left}} \tag{6}$$

where b_{right} and b_{left} are constants and all the $\xi_\tau(t)$s and $\xi^\mathcal{T}_\tau(t)$s are drawn independently for each t and each τ from a Gaussian distribution with zero mean and standard deviation σ_I. The $\Sigma^\mathcal{S}$s and the $\Sigma^\mathcal{T}$ carry information, respectively, on the signal and the time elapsed since the beginning of each episode. Even though the $x^\mathcal{T}_\tau$ increase with time, the $\Sigma^\mathcal{T}$s can be non-monotonic, something that will play an important in role in implementing an effective 'moving threshold' for the decision mechanism.

The other two sums are instead related to the 'wait' option:

$$\Sigma^\mathcal{S}_{\text{wait}}(t) \equiv \sum_\tau w_{\text{wait},\tau}\left|x_\tau(t) + \xi_\tau(t)\right| \tag{7}$$

$$\Sigma^\mathcal{T}_{\text{wait}}(t) \equiv \sum_\tau w^\mathcal{T}_{\text{wait},\tau}\left[x^\mathcal{T}_\tau(t) + \xi^\mathcal{T}_\tau(t)\right] + b_{\text{wait}}, \tag{8}$$

where the absolute value in Eq. 7 is taken to account for the intuition that a signal and its negative mirror should equally affect the agent's propensity to defer a decision. The $\xi_\tau(t)$s and $\xi^\mathcal{T}_\tau(t)$s are introduced to model the intrinsic noise implied in any plausible biological implementation of the integration process, such as fluctuations in the instantaneous firing rate of a network of neurons.

By setting:

$$\Sigma_x \equiv \Sigma_x^{\mathcal{S}} + \Sigma_x^{\mathcal{T}} \tag{9}$$

(with $x \in \{\text{left, right, wait}\}$), the six sums are then non-linearly combined through a softmax function (the white circles on the far right of Fig. 1) to define a probability distribution over the possible actions:

$$p_{\text{right}}(t) = \frac{\mathrm{e}^{\Sigma_{\text{right}}(t)}}{\mathrm{e}^{\Sigma_{\text{left}}(t)} + \mathrm{e}^{\Sigma_{\text{wait}}(t)} + \mathrm{e}^{\Sigma_{\text{right}}(t)}} \tag{10}$$

and analogous expressions for 'left' and 'wait'. By definition, $p_{\text{left}}(t) + p_{\text{wait}}(t) + p_{\text{right}}(t) = 1$ for every t. The agent then randomly chooses an option according to the three probabilities.

The agent is thus completely determined by the choice of the six sets of n_τ weights: $w_{\text{left},\tau}$, $w_{\text{wait},\tau}$, $w_{\text{right},\tau}$, $w^{\mathcal{T}}_{\text{left},\tau}$, $w^{\mathcal{T}}_{\text{wait},\tau}$, $w^{\mathcal{T}}_{\text{right},\tau}$, and three constant offsets b_{left}, b_{wait}, and b_{right}. These weights and offsets are learned by trial-and-error through a reinforcement learning procedure aiming to maximise the reward received on a large number of episodes (see Methods). All the results shown, if not otherwise stated, are obtained using the same set of weights, at the end of the training procedure, with $T_{\max} = 2$ s, $\sigma = 0.18\,\mathrm{s}^{-\frac{1}{2}}$, $\sigma_\mu = 0.25$, and $\sigma_I = 0.02$.

In random dots experiments, usually a number of dots moves randomly on a screen, with a fraction of them moving instead coherently in one direction (either left or right in different episodes). The percentage of coherently moving dots ('coherence') is a measure of how difficult an episode is, not unlike $|\mu|$ in the model (with sign of μ corresponding to a coherent movement towards left or towards right respectively). To make the parallel between the present task and the experimental settings more evident, in the following we will show results using either $|\mu|$ or the coherence of the signal, the two measures being related by (see Methods):

$$|\mu| = 0.216\,\frac{\text{coherence}}{\sqrt{100 - \text{coherence}}}. \tag{11}$$

During learning, the model estimates at each step t the total future expected reward $V(t)$ for the current episode. Such estimate is computed by a linear summation of the integrators (Fig. 1, bottom-right) and is used to establish a moving baseline to modulate the changes in the model's weights during training. The adoption of the baseline $V(t)$ constitutes a standard procedure for actor-critic reinforcement learning algorithms (see Methods for more details).

Fig 1. Task and model schematic. The random movement of a group of dots on a screen (far left) is represented as a uni-dimensional noisy signal $s(t)$ (black line), sampled at discrete time steps $\Delta t = 10$ ms, from a Gaussian distribution of mean μ and variance σ^2. The task requires the subject to guess the sign of μ, by moving a lever to the right (positive sign) or to the left (negative sign); the subject can 'choose when to choose', within a maximum episode duration $T_{\max}$. The learning agent integrates the signal over different time scales τ ($x_\tau(t)$s, blue lines); over the same time scales the agent integrates a constant input ($x_\tau^{\mathcal{T}}(t)$s, yellow-red lines) to simulate an internal clock mechanism estimating the passage of time; in both cases, the darker the colour the longer the corresponding time scale. At each time instance, the weighted sums of the integrators (far right) are fed into a decision layer that computes the probability of choosing 'left' and 'right', thus terminating the episode, or to 'wait' to see another sample of $s(t)$. If the subject gives the correct answer (the guessed sign coincides with the actual sign of μ) within the time limit, a reward is delivered; otherwise, nothing happens. In any case, a new episode starts. The agent learns by observing the consequences (obtained rewards) of its actions, adapting the weights assigned to the $x_\tau(t)$s and $x_\tau^{\mathcal{T}}(t)$s. During learning, the model estimates at each step t the total future expected reward $V(t)$ for the current episode as a linear summation of the integrators (bottom right).

Fig 2. Learned decision strategy. Evolution of $p_{\text{right}}(t)$ (blue line) and $p_{\text{left}}(t)$ (red) during an episode where the correct action is 'right' (that is, $\mu > 0$). Decisions are made within short 'active' windows of time during which fleeting bursts of $p_{\text{left}}(t)$ or $p_{\text{right}}(t)$, corresponding to the alignment of many integrators, make an action possible. coloured circles: value of a subset of 5 of the 10 integrators (slow to fast associated time scales from top to bottom). Reds: negative values; blues: positive values. Uniformly positive (negative) values for the integrators are associated with bursts of p_{right} (p_{left}; see times denoted with 1 and 2 in the plot). The converse is not true: outside bursts (point 3) or when a burst withers (point 4), not all the integrators assume low absolute values.

2.2 Decision as a majority vote

Fig. 2 shows the evolution of $p_{\text{right}}(t)$ (blue line) and $p_{\text{left}}(t)$ (red) during an episode where the correct action is 'right' (that is, $\mu > 0$). As expected, $p_{\text{right}}(t)$ is for the most part greater then $p_{\text{left}}(t)$ (although this is unnoticeable in the plot where the probabilities are very small), signalling that the agent favours the action associated with the correct decision. Nevertheless, both probabilities are very low most of the time, implying that $p_{\text{wait}}(t)$ is often close to one (not shown). Thus, the agent appears to select a strategy in which decisions are made within short 'active' windows of time during which fleeting bursts of $p_{\text{left}}(t)$ or $p_{\text{right}}(t)$ make an action possible. Such strategy is not trivially associated with the intuitive picture of a process accumulating information over time until some threshold is met.

This is due to the availability of multiple time scale. In fact, the agent exploits the information carried by the different integrators by waiting for their consensus, akin to a majority vote. A short-lived fluctuation in the fastest integrators would not be enough for a decision. Yet, in conjunction with a longer-lived fluctuation of the slower integrators, a burst in one of the actions is triggered. Being the consensus fleeting, such probability bursts are usually quite low (they often stay below a probability of 0.1) and therefore function as 'open windows' paving the way to a decision, more than as 'funnels' forcing it. Decisions therefore happen when the different time scales stay in agreement for an extended period (roughly 100 ms).

This is illustrated in Fig. 2 with coloured circles, each row representing the evolution of one integrator (for a subset of 5 of the 10 integrators, with slow to fast time scales from top to bottom). As expected, inside a burst of $p_{\text{right}}(t)$ almost all the integrators present large positive values (dark blue, see for example temporal instance number 1 in Fig. 2). On the other hand, integrators typically assume negative values (light to dark red) in correspondence of bursts of $p_{\text{left}}(t)$, as it is shown in the temporal instance number 2. The converse is not true: in absence of probability bursts, not all the integrators assume low absolute values (see, for example, coloured circles corresponding to number 3). This is due to the fact that the integrators, though correlated, detect fluctuations in the signal over different time scales. Moreover, the non-linear nature of the probability function (Eq. 10) dampens integrators' fluctuations falling below a given range of values. When a burst fades away (see for example points between 2 and 4) not all the integrators go down together. Initially the faster integrators become neutral or even slightly change sign. Afterwards the slower integrators follow suit. But the process is not, of course, completely linear, and you can have (see instance number 4 and neighbouring points) higher values for intermediate integrators, while the slowest one is still decreasing, and the faster ones fluctuate rapidly.

2.3 Model's performance

Fig. 3A shows the fraction of correct choices as a function of the decision time, both for the agent at end of training (black line) and for the optimal fixed-t observer (blue line) that, at each time t, simply chooses according to the sign of the sum of the signal up to time t. Its performance can be derived analytically:

$$\text{Fraction Correct}(t) = \frac{1}{2} + \frac{1}{\pi} \arctan \sqrt{\frac{\sigma_\mu^2\, t}{\sigma^2}} \tag{12}$$

If the task were to decide exactly at time t, no other decision maker could outperform it; for this reason it is deemed optimal. Of course, the present task does not force a decision time; yet, the comparison with the fixed-t observer sheds lights on the agent's strategy and the underlying trade-offs.

The agent is free to"choose when to choose", thus it is not surprising that it clearly outperforms the optimal fixed-t observer for shorter decision times (the inset of Fig. 3A shows the distribution of decision times for the agent). We see that the two performances cross slightly above the average decision time for the agent; beyond this point, the fixed-t observer dominates. Indeed, the agent can make the easy decisions early on and wait to see how the signal evolves when the choice appears more uncertain; the fixed-t observer, on the other hand, is bound to decide at time t, no matter how clear or ambiguous the observed signal was up to that point. Thus, the steep rise of the agent's performance for very short decision times is mainly a reflection of its ability to tell apart the easy episodes from the hard ones. The fixed-t observer catches up for longer times, where the agent is left with only the most difficult decisions and its performance consequently declines. For the fixed-t observer, instead, larger ts always mean more information and therefore its performance monotonically increases. We notice how at the crossing point, the agent has already made the large part of its decisions, as it is apparent from the distribution of decision times.

The agent outperforms all the single-time-scale integrators, Eq. 1, when each time scale is in turn taken from the set of the τs available to the agent (Fig. 3B; agent: dashed line; single-time-scale integrators: circles). Single-time-scale integrators correspond to the Ornstein-Uhlenbeck decision process that extends the standard drift-diffusion model [28, 32]: whenever the integrator crosses a threshold or its negative mirror a decision is made ($\mu > 0$ and $\mu < 0$ respectively). For each integrator, the threshold was chosen by numerically maximizing the fraction of correct responses on a sample of signals. The performance of the single-time-scale integrator peaks for intermediate values of the associated time scale τ, though it always stays well below the performance attained by the agent. The agent, therefore, is able to leverage the information on multiple time scales to gain a clear performance advantage with respect to the drift-diffusion model on the whole spectrum of τs.

Fig. 3C and D show the accuracy and the mean response time of the agent, as the coherence of the signal varies (Eq. 11). The black line in panel Fig. 3C is computed as:

$$\text{Fraction Correct(coherence)} = 1 - \frac{1}{2}\exp\left[-\left(\frac{\text{coherence}}{7.97}\right)^{1.62}\right] \qquad (13)$$

as in Fig. 3 of [4], where the parameters of the curve were fitted to experimental data; the match between the experimental fit and the result of the agent is striking. In Fig. 3D, instead, the black line is a generic sigmoidal function plotted for illustration purposes. As found in the experiments, the agent's responses become faster as the task becomes easier (larger coherences).

Fig 3. Performance after training. **A**: Fraction of correct choices as a function of the decision time, both for the agent at end of training (black line) and for optimal fixed-t observer (blue line) that simply chooses according to the sign of the accumulated signal up to time t (see text). The agent clearly outperforms the fixed-t observer for shorter decision times, thanks to its freedom to 'choose when to choose'. The steep rise of the agent's performance for very short decision times is mainly a reflection of its ability to tell apart the easy episodes from the hard ones. Inset: response time histograms for correct (grey) and wrong (green) decisions **B**: the agent (dashed line) outperforms, considering the fraction of correct choices on a sample of episodes, all the single-time-scale integrators with optimised decision threshold (dots; the continuous line is a second-degree polynomial fit for illustration purposes). The performance of the single-time-scale integrator peaks for intermediate values of the associated time scale τ, though it always stays below the performance attained by the agent. **C** and **D**: Accuracy and mean response times for different values of coherence (dots). **C**: The accuracy curve for the agent is in very good agreement with experimental findings: the black line is the result of a fit on experimental data ([4]; see text for more details). **D**: As accuracy increases, responses become faster, as found in experiments (black line: fit with a sigmoid-like function).

2.4 Signal neutrality

A more microscopic look at the decision process surprisingly uncovers shared features between the internal dynamics of the artificial agent and the activity observed in neurons in the lateral intraparietal cortex (LIP) during a random dots task [4, 48].

We will focus our attention on the evolution of a key observable in the model, defined as (see Eqs. 9, 3, and 4):

$$\Delta\Sigma_{\text{right}}(t) \equiv \Sigma_{\text{right}}(t) - \Sigma_{\text{wait}}(t) \qquad (14)$$

and its ‘left’ counterpart: $\Delta\Sigma(t)$ provides a direct measure of the propensity of the agent, at time t, to make a ‘right’ or ‘left’ decision respectively (see Eq. 10).

Fig 4. Signal neutrality. $\Delta\Sigma_{\mathrm{right}}(t)$ (see Eq. 14) provides a direct measure of the propensity of the agent, at time t, to make a ‘right’ decision. **A** Evolution of $\Delta\Sigma_{\mathrm{right}}$, averaged over many successful episodes with the same signal coherence. On the left, the episodes are aligned to the beginning of the episode and $\Delta\Sigma_{\mathrm{right}}$ shows a marked sensitivity to the coherence of the signal. When the average is performed by aligning all the episodes to the time of the decision (right), signal neutrality clearly appears: the sensitivity to the signal strength is completely lost and all the lines collapse on the same curve for several hundreds of milliseconds. Inset: the same analysis on wrong episodes. The similarities with what is found in the discharge of LIP neurons during a motion-discrimination task are striking (see, *e.g.*, Fig. 7 in [4]). **B**: Time course of x_τ for a single-time-scale integrator with $\tau = 2s$ and optimised decision threshold (x_τ, for an integrator with threshold, plays the role that $\Delta\Sigma_{\mathrm{right}}$ has in the agent). **C**: Time course of $\Delta\Sigma_{\mathrm{right}}$ (see Eq. 14 for an agent optimised with a single timescale $\tau = 2s$. In both **B** and **C** the collapse of the curves for different signal coherences is imperfect (rightmost part of the plots). **D**: Comparison of performance and signal neutrality for the single-τ agent and the single-time-scale integrator as τ varies. The proposed model (dashed lines) shows better accuracy while exhibiting the experimentally observed collapse of the time course of neuronal activity aligned at the decision time.

Fig. 4A shows the evolution of $\Delta\Sigma_{\mathrm{right}}$, averaged over many episodes in which the agent has made the correct decision ‘right’. The traces are grouped by signal coherence. The left part of Fig. 4A shows the evolution of the average $\Delta\Sigma_{\mathrm{right}}$, with traces aligned to the beginning of the episode (so that $t = 0$ in the plot corresponds to $t = 0$ of each signal). $\Delta\Sigma_{\mathrm{right}}$ shows a marked sensitivity to the coherence of the signal. Moreover, the traces do not saturate over several hundreds of milliseconds, highlighting how the agent is making use of its slower integrators.

Ramp-like changes in the discharge of LIP neurons have been repeatedly observed, with steeper rise in spike rate for higher stimulus coherence (see, *e.g.*, Fig. 7 in [4]). Such ramps have been interpreted as a signature of the accumulation of evidence (originating in the extrastriate visual cortex, in the case of LIP neurons), for or against a specific behavioural response (‘left’ or ‘right’) [9, 19]. This interpretation is fully compatible with what is seen in the agent.

However, when the averages of the $\Delta\Sigma_{\mathrm{right}}$ traces (or of the activity of LIP neurons) are performed by aligning the episodes to the time of the decision, a clear signature of signal neutrality emerges: sensitivity to the stimulus’s coherence is lost and all the lines surprisingly collapse on the same curve (Fig. 4A, right).

For the experimental data, a reasonable explanation for such collapse is that the neuronal circuitry is engaged in stereotyped dynamics, independent from the signal, just after a decision is made and before it is manifested with a physical action, perhaps as the result of a feedback from downstream areas.

But this cannot hold for the agent, where instead signal neutrality arises precisely from the presence of multiple time scales. Figs. 4B and C show the time course of the equivalent of $\Delta\Sigma_{\text{right}}$ for a single-time-scale integrator (with optimised threshold) and for an agent trained with just one time scale available (in both cases, $\tau = 2$ s). For an integrator with threshold, x_τ plays the role that $\Delta\Sigma_{\text{right}}$ has in the agent.

In these cases, there is no clear collapse of the curves for different signal coherences (rightmost part). To make this statement more systematic, we introduce an operative measure of signal neutrality, that is basically the inverse of the maximum distance between the curves for different coherences, averaged over the interval prior to the decision (see Methods). In Fig. 4D we report this measure (upper bars) for five of the 10 time-constants τs used by the agent, both for the single-τ agent after training and for the single-time-scale integrator with optimised threshold: signal neutrality is clearly lower than that of the agent using all the 10 τs (upper dashed black line). At the same time, all the single-τ models achieve lower accuracy with respect to the multi-τ agent (lower bars *vs* lower dashed black line).

On the other hand, $\Delta\Sigma_{\text{left}}$ (*i.e.*, the propensity of the agent to make the wrong decision; in this case, to choose 'left') does not display signal neutrality. The same holds true for its experimental counterpart, that is the activity of LIP neurons when the random dot motion is away from their receptive field (see Fig. 7 in [4], dashed lines).

2.5 The scalar property

The agent's behaviour conforms to one of the hallmarks of temporal cognition: the scalar property [49]. This is illustrated in Fig. 5: the distributions of response times of the agent are shown for three different values of coherence (histograms; black lines are best fits with a Gamma distribution). As the coherence increases, as expected, the average response time of the agent decreases from 4.6 s to 370 ms.

Simply stated, the scalar property — as observed for example in interval timing [49], and multistable perception [50] — implies that higher moments of the intervals' distribution scale as appropriate powers of the mean (in particular, this implies a constant coefficient of variation). Or, in other words, that the shape of the distribution does not change when its mean varies, even over very wide ranges.

And indeed, the coefficient of variation of the agent moves in a very narrow range (0.44 - 0.46, see legend; compatible with the experimental findings, see [49, 50]), whilst the mean value varies by more than one order of magnitude. The invariance of the shape of the distribution is made immediately evident in the inset of Fig. 5. Here the fitted Gamma distributions (black lines in the main plot) are rescaled so to all have mean equal to 1 (colours are consistent with the histograms): the similarity of the three curves is striking. Lastly, we note how the highest value of coherence reported in the plot is very unlikely under the distribution used during the training phase; indeed it corresponds to a value of μ five times the standard deviation σ_μ of the distribution of μ. Thus, the scalar property appears to be a very robust property of the learned decision strategy, holding well beyond the range of functioning to which the agent has been accustomed during training.

Fig 5. Scalar property. Increasing the signal coherence, the average response time of the agent decreases; still the coefficient of variation of the response times varies in a very narrow range (see legend). The black lines are the best fit of the simulation histograms with a Gamma distribution. Inset: the fitted Gamma distributions are rescaled so to have mean equal to 1, making immediately evident how the shape of the distribution stays almost unchanged as its average moves over almost one order of magnitude (colours consistent with the histograms in the main plot). Note how the highest value of coherence is very unlikely under the distribution used for training the agent (corresponding to a value of μ five times the standard deviation σ_μ of the distribution of μ): the 'invariant shape' property of the response time distribution therefore holds well beyond the typical range of functioning of the agent.

In the following, we will show semi-analytically that signal neutrality and the scalar property hold for a stochastic process, $\Delta\Sigma^{\mathcal{S}}_{\text{right}}(t)$, with a decision threshold θ, provided that it has two specific characteristics: $\Delta\Sigma^{\mathcal{S}}_{\text{right}}(t)$ must possess a power-law mean and a constant variance. That is,

$$\langle \Delta\Sigma^{\mathcal{S}}_{\text{right}} \rangle(t) \propto \mu\, t^a \tag{15}$$

$$\text{Var}[\Delta\Sigma^{\mathcal{S}}_{\text{right}}](t) \simeq \varsigma^2, \tag{16}$$

where μ, as in the agent's task, gives a measure of how difficult the task is.

Indeed, a simple argument shows how Eqs. 15 and 16 are compatible with the scalar property. Imagine that the decisions are made when $\Delta\Sigma^{\mathcal{S}}_{\text{right}}$ reaches a threshold θ.

Consider as the mean of the response time $t = \mathrm{RT}$, in a first approximation, the time at which $\langle \Delta\Sigma^{\mathcal{S}}_{\mathrm{right}} \rangle(t) = \theta$; that is:

$$\mathrm{RT}(\mu) = \left(\frac{\theta}{\mu}\right)^{\frac{1}{a}}. \tag{17}$$

A similar calculation then provides the times $\mathrm{RT}^{\pm}$ at which the mean ± the (constant) standard deviation ς cross the threshold:

$$\mathrm{RT}^{\pm} = \left(\frac{\theta \mp \varsigma}{\mu}\right)^{\frac{1}{a}} \tag{18}$$

Taking $\mathrm{RT}^{-} - \mathrm{RT}^{+}$ as a rough measure of the standard deviation of the distribution of response times, one has that the coefficient of variation:

$$\mathrm{CV} \equiv \frac{\mathrm{RT}^{-} - \mathrm{RT}^{+}}{\mathrm{RT}} = \frac{(\theta + \varsigma)^{\frac{1}{a}} - (\theta - \varsigma)^{\frac{1}{a}}}{\theta^{\frac{1}{a}}} \tag{19}$$

does not depend on μ, that is on the difficulty of the task — this is our operative definition of the scalar property.

The most straightforward explanation for signal neutrality, on the other hand, is that, when crossing the threshold, the behavior of the stochastic process is dominated by fluctuations, that are naturally independent of the coherence of the signal. Yet, such fluctuations need to have a similar structure at different times, otherwise the behaviour close to the decision threshold would still depend on the coherence: a strongly coherent signal will lead to a short decision time and *viceversa*. In other words, a minimal requirement would be a nearly-constant variance of the integration variable, and Eq. 16 is a prerequisite for signal neutrality.

We now will show how our model complies with these two conditions (Eqs. 15 and 16). Now, we rewrite Eq. 14 as (see Eqs. 3-9):

$$\Delta\Sigma_{\mathrm{right}} = \Delta\Sigma^{\mathcal{S}}_{\mathrm{right}} - \Delta\Sigma^{\mathcal{T}} \tag{20}$$

where:

$$\Delta\Sigma^{\mathcal{S}}_{\mathrm{right}} \equiv \Sigma^{\mathcal{S}}_{\mathrm{right}} - \Sigma^{\mathcal{S}}_{\mathrm{wait}} \tag{21}$$

is a term that provides information on the signal only. And:

$$\Delta\Sigma^{\mathcal{T}} \equiv \Sigma^{\mathcal{T}}_{\text{wait}} - \Sigma^{\mathcal{T}}_{\text{right}} \tag{22}$$

carries information on the passage of time only. We note that on the r.h.s. of Eq. 22 we could insert $\Sigma^{\mathcal{T}}_{\text{left}}$ in place of $\Sigma^{\mathcal{T}}_{\text{right}}$ with no notable numerical difference in the result. This is because the right and left choices are *a priori* equivalent in the present task, and therefore the inferred $w^{\mathcal{T}}_{\text{right},\tau}$ and $w^{\mathcal{T}}_{\text{left},\tau}$ are in fact very similar. For this reason $\Delta\Sigma^{\mathcal{T}}$ does not carry a 'right' label.

Focusing on Eq. 21, we note that, in fact, $\Delta\Sigma^{\mathcal{S}}_{\text{right}}$ can be understood as a stochastic process for which we can compute, under some approximations, how the mean and the standard deviation evolve in time (see Methods). Defining:

$$\Delta w_\tau \equiv w_{\text{right},\tau} - w_{\text{wait},\tau}, \tag{23}$$

we have:

$$\langle \Delta\Sigma^{\mathcal{S}}_{\text{right}} \rangle (t|\mu) \simeq \mu \sum_\tau \Delta w_\tau \left(1 - \exp(-\frac{t}{\tau})\right) \tag{24}$$

and:

$$\begin{aligned}\text{Var}[\Delta\Sigma^{\mathcal{S}}_{\text{right}}](t|\sigma) \simeq \sigma^2 \sum_{\tau_1} \sum_{\tau_2} \frac{\Delta w_{\tau_1} \Delta w_{\tau_2}}{\tau_1 + \tau_2} \\ \Big[1 - \exp\big(-(\frac{1}{\tau_1} + \frac{1}{\tau_2}) t\big)\Big].\end{aligned} \tag{25}$$

Fig. 6 shows the result of Eqs. 24 (left) and 25 (right). What one finds is that both $\langle \Delta\Sigma^{\mathcal{S}}_{\text{right}} \rangle$ and $\text{Var}[\Delta\Sigma^{\mathcal{S}}_{\text{right}}]$ (normalized as to have maximum value of 1) can be well fitted by a saturating power law:

$$y(t) = \left(\frac{t}{t + t_0}\right)^a. \tag{26}$$

In particular, the fitted parameters for $\langle \Delta\Sigma^{\mathcal{S}}_{\text{right}} \rangle$ ($t_0 = 2.07$ s, $a = 0.76$) make it close to a power law, as in Eq. 15, on a broad range of t (roughly speaking $t < t_0$, where $t_0 > T_{\text{max}} = 2$ s):

$$\langle \Delta\Sigma^{\mathcal{S}}_{\text{right}} \rangle (t|\mu) \propto \mu\, t^a. \tag{27}$$

Fig 6. Mean and standard deviation of $\Delta\Sigma^{\mathcal{S}}_{\text{right}}(t)$. A: the mean (see Eq. 24) is close to a power-law ($\propto \mu\, t^{a}$) in the typical range of reaction times ($0.2 - 1.6$ s). The red line is a saturating power-law fit (see text for details). B: the standard deviation (see Eq. 25) moves, instead, in a quite limited range, approximating a constant time-course; the dashed black line shows the evolution of the standard deviation for the drift-diffusion model for comparison. The red line is a saturating power-law fit (see text for details). The curves in panels A and B are a combination of contributions on multiple time scales (Eqs. 24 and 25), and provide a possible explanation for signal neutrality and the scalar property displayed by the agent.

We note how a power-law trend is consistent with the seemingly non-saturating behavior observed on the left part of Fig. 4A.

Moreover, $\text{Var}[\Delta\Sigma^{\mathcal{S}}_{\text{right}}]$ ($t_0 = 0.117$ s, $a = 1.71$) moves in quite a narrow range of values, *i.e.*:

$$\text{Var}[\Delta\Sigma^{\mathcal{S}}_{\text{right}}](t) \simeq \text{const}, \tag{28}$$

an approximation to Eq. 16. Thus, the agent roughly satisfies the conditions in which signal neutrality and the scalar property hold for the very simple model introduced above.

Such result is a direct consequence of having multiple time scales. Indeed, referring back to Eqs. 24 and 25, the power-law form of the mean and the nearly-constant variance of $\Delta\Sigma^{\mathcal{S}}_{\text{right}}$ are achieved by aptly combining exponential contributions with different saturation times.

On the contrary, the drift-diffusion model (that can be seen as having a single infinite time scale) satisfies Eq. 15 (with $a = 1$) but not Eq. 16; indeed $\text{Var}[\Delta\Sigma](t|\sigma) = \sigma^2\, t \neq \text{const}$ (see the dashed black line in Fig. 6B for a graphical comparison). In this case, the non-constant variance leads to a coefficient of variation CV that strongly depends on μ:

$$\text{CV} = \frac{\sigma}{\sqrt{\mu\,\theta}}. \tag{29}$$

We note that such result is not in contrast with the linear relationship between mean and standard deviation of the response times found in [51]: such linear relationship is a necessary, but not sufficient condition for a constant CV (the scalar property). In the range of average response times explored in the reference, the CV varies on a quite broad range, from approximately 0.5 (consistent with our findings and the evidence in other experimental settings [49, 50]) to almost 1 (corresponding to an exponential distribution of response times).

In the above argument, we assumed a constant decision threshold θ; but, we will see in the next section, the agent behaves as if subject to a non-fixed threshold. However, if for a single signal coherence μ the distribution of reaction times is not too wide (as suggested by a coefficient of variation 0.45, see Fig. 5), θ can be assumed to vary in a range narrow enough not to invalidate the argument.

Fig 7. Signal neutrality and scalar property during training. Evolution of signal neutrality (black line), scalar property (blue line), and accuracy (dashed red line, scale on the right) as the training progresses. Signal neutrality attains a broad maximum where the performance has almost plateaued. Thus signal neutrality can be interpreted as the signature of a 'satisficing' strategy, rather than of an optimal one. The scalar property, on the other hand, keeps growing even for very long training. Yet, the evolution of signal neutrality and the scalar property are highly correlated, suggesting a common origin for the two (see Text for discussion).

In view of the above considerations, then, signal neutrality and the scalar property share a same origin. Further evidence of this can be found in the evolution of the two measures during the training phase.

Fig. 7 shows the average evolution, during training, of signal neutrality (black line; the same measure reported in Fig. 4D), scalar property (blue line; see Methods for the definition of the metric), and accuracy (dashed red line, scale on the right y-axis). Accuracy is computed on a sample taken from the Gaussian $p(\mu)$ used for training. All the lines are computed by averaging the results of 100 different realizations of the training.

The evolution of signal neutrality and the scalar property are highly correlated for much of the training phase, with an initial fast increase that continues up to about $10^4 - 10^5$ episodes, where the accuracy has almost plateaued — this is the region of all that has been shown above (Figs. 4A, 5, and 6). Such correlated progress naturally hints to a common origin for the two measures, and makes us advance the hypothesis that a behavioural policy displaying these two properties could represent an 'optimal' information-extraction strategy for dealing with a decision task in a volatile environment. It wouldn't be by chance then that the agent robustly finds such a strategy by tuning its parameters in a ecologically plausible way.

Yet, after about 10^5 training episodes — and therefore probably far beyond the experimental training durations, the behaviour of the two curves in Fig. 7 starts to diverge: whilst the scalar property keeps improving, signal neutrality attains a broad peak, after which it gradually breaks down in the face of very modest performance gains.

Therefore, the scalar property seems to be more fundamental then signal neutrality, at least for what concerns the strategy asymptotically discovered by the learning agent.

In this sense, signal neutrality *per se* cannot be viewed as signature of an optimal strategy for the agent, but rather of a ‘satisficing’ one [53]. Faced with a wide distribution of coherences, the agent pretty quickly finds a robust strategy that, at around decision time, disregards coherence by relying on fluctuations to make decisions, and still ensures a very good performance. Nevertheless, the agent can do slightly better, given enough training time, by giving more weight to the ‘drift’ component (Eq. 24) and less to the ‘diffusion’ component (Eq. 25): this is what happens on the far right of the plot.

2.6 Collapsing boundaries

The hypothesised optimality of the agent’s strategy finds indirect support in the behaviour displayed by the component $\Delta\Sigma^{\mathcal{T}}$ of $\Delta\Sigma$ (Eq. 22) that depends only on the passage of time and not on the signal.

In a sense, $\Delta\Sigma^{\mathcal{T}}(t)$ measures the propensity of the agent at time t to wait for another input instead of making a (either right or left) decision, independently from the signal. Indeed, looking back at Eq. 20, $\Delta\Sigma^{\mathcal{T}}$ effectively acts as a time-dependent bias term that, in the context of a drift-diffusion model, could be straightforwardly interpreted as a time-dependent threshold. Lacking a threshold mechanism, such interpretation cannot be directly extended to the learning agent; yet it is reasonable to expect that the range of values attained by $\Sigma^{\mathcal{S}}_{\text{right}}$ at decision time shifts in accordance with the time-dependent bias. And this is indeed the case.

Fig. 8 shows (black thick line) the evolution of $\Delta\Sigma^{\mathcal{T}}(t)$ from 0 to $T_{\max} = 2$ s. In addition, three sample trajectories of $\Delta\Sigma^{\mathcal{S}}(t)$ (coloured lines) are shown from $t = 0$ to decision time (marked by circles). The shaded grey area marks the region where 70% of the (correct) decisions are made; as expected, the region’s boundaries mostly run parallel to $\Delta\Sigma^{\mathcal{T}}(t)$.

Therefore, in this sense, $\Delta\Sigma^{\mathcal{T}}$ does work a soft threshold for the decision. Conversely, looking at Eq. 20, one can view $-\Delta\Sigma^{\mathcal{T}}$ as an ‘urgency’ signal that pushes for a decision as the episode time elapses, not unlike what has been observed experimentally in the lateral intraparietal area [52].

In this respect we want to point out how the soft threshold $\Delta\Sigma^{\mathcal{T}}$ does not behave purely as an urgency signal. In fact the decision is made more and more likely as the time passes only after about 200 ms (when $\Delta\Sigma^{\mathcal{T}}$ reaches a peak); but initially earlier decisions are favoured by a rise of the threshold. Such behaviour could be beneficial to effectively exploit the tails of the distribution of μ: sometimes the signal received is very clear, and the agent learns in those cases that an extremely quick decision is the best option.

Yet, beyond such interpretations, it is interesting to note - coming back to our point about the presumed optimality of the agent - how this shape of the moving threshold qualitatively matches the one demonstrated to be optimal in [8] (see Fig. 2B therein; see also [10]). Even if the models in the references and in the present paper are not structurally equivalent, it is nonetheless striking that the agent, by trial-and-error, seems able to approximate the optimal behaviour at least in this respect. And, on the other hand, this piece of evidence gives support to the hypothesis that the agent could somehow learn to leverage, through the multiple time scales at its disposal, some deeper information processing strategy possibly exploited by the brain.

Fig 8. Collapsing boundaries. $\Delta\Sigma(t)_{\mathrm{right}}$ (see Eqs. 14 and 20) can be decomposed in a signal-dependent part ($\Delta\Sigma^{\mathcal{S}}_{\mathrm{right}}$) and a time-dependent part ($\Delta\Sigma^{\mathcal{T}}$; see Eq. 22), that measures the propensity of the agent at each time to wait for another input instead of making a decision. In the plot, $\Delta\Sigma^{\mathcal{S}}_{\mathrm{right}}$ for three sample episodes (coloured lines) is depicted, alongside $\Delta\Sigma^{\mathcal{T}}$ (thick black line). $\Delta\Sigma^{\mathcal{T}}$ acts as a time-dependent threshold: most of the decisions (dots mark the decision times) indeed fall inside a strip running parallel to it (the grey area is where 70% of the decisions are made). The resulting boundaries do collapse, but only for longer response times: until about 200 ms, a rise of the effective threshold favours early decisions.

2.7 Robustness

The utilisation of a wide range of time scales makes the performance of the agent robust to variations of the task and to the intrinsic noise. This is shown in Fig. 9A and B. We varied T_{max} (the maximum duration of an episode) and σ_I (the standard deviation of the intrinsic noise, ξ_τs and $\xi^{\mathcal{T}}_\tau$s in Eqs. 3-8) systematically and, for each value, run the learning process from scratch. The results of the agent are then compared to those of the single integrators, each with the decision threshold optimised for each individual condition.

In Fig. 9A, as T_{max} increases (and σ_I stays at its reference point of 0.02), the fraction of correct responses rises monotonically both for the agent (black line) and for three single integrators (colored lines), with the performance of the agent staying superior on the whole range of T_{max} explored. Two features are noteworthy: first the lines for the fastest and slowest integrators ($\tau = 0.1$ s and $\tau = 10$ s respectively) cross at intermediate values of T_{max}, with the longer τ surpassing the shorter ones for higher episode durations; and, second, the advantage of the learning agent shrinks in comparison to the longer τ for longer T_{max}. These features have a common origin. Indeed, from Eq. 1, a signal s(t) of mean μ will lead all the integrators, given enough time, to the same (statistically) stationary value of μ, but with different levels of noise: integrators with longer τs will have a smaller variance and thus will be more reliable in detecting whether $\mu > 0$ or $\mu < 0$. On the other hand the time needed to reach the stationary state will be longer for longer τs. Longer integrators will still be integrating the signal for shorter T_{max} and, as a consequence, their value will carry less information on the μ. Hence, the smaller τs will dominate for shorter T_{max}, the larger τ for longer T_{max}. The intermediate $\tau = 2.1$ s, on the other hand, shows a steadier, intermediate, trend.

In Fig. 9B, the level σ_I of intrinsic noise is varied, with T_{max} kept constant at 2 s. The performance of the agent (black line) is always substantially higher than that of the single integrators (coloured lines). As expected, performance deteriorates as σ_I increases from 0 to 0.2; yet the decrease is only surprisingly slight, considering that the maximum value attained by σ_I is comparable with the typical dynamical range of the integrators x_τ. Such range is determined by the distribution $p(\mu)$ (here, a Gaussian of standard deviation $\sigma_\mu = 0.25$). It is then clear that the highest levels of intrinsic noise really affect the typical value of the integrators. This is even more true taking into account that the slowest integrators operate far from the asymptotic value, given the limited integration time. This consideration is clearly reflected in the behaviour of the single integrators. The fast integrators ($\tau = 0.1$ s and $\tau = 2.1$ s) indeed are scarcely affected by the increase in noise. On the other hand, the slowest integrator ($\tau = 10$ s) shows good accuracy for very low levels of noise, but then becomes rapidly ineffective for higher values of σ_I.

Fig. 9C shows the evolution of the 'moving threshold' $\Delta\Sigma^{\mathcal{T}}$ (Eq. 22) for three values of $T_{\max}$. For very low $T_{\max}$ (black line) the threshold only decays, always pushing for a decision. For higher values of $T_{\max}$, instead, as we have already seen in Fig. 8, the moving threshold initially rises; it reaches a peak and then decays afterwards, making a decision ever more likely. Such peak shifts with $T_{\max}$ and so does, even more clearly, the time at which the threshold reaches back its initial value (around 1 s for $T_{\max} = 2.0$ s, and around 5 seconds for $T_{\max} = 10$ s).

Fig. 9D shows w_{right} (Eq. 3; $w_{\text{left}} \simeq w_{\text{right}}$) for different values of intrinsic noise σ_I (continuous lines are fourth degree polynomial fits for illustrative purposes). Coherently with what we have seen in Fig. 9B, the peak of the lines, corresponding to the most exploited time scale, shifts towards lower τ values as σ_I increases.

Fig 9. The wide range of time scales makes the agent's performance robust to variations of the task and to the intrinsic noise. **A**: as $T_{\max}$ increases, the fraction of correct responses rises monotonically both for the agent (dashed black line) and for all the single integrators, with the performance of the agent staying superior on the whole range of $T_{\max}$ explored. **B**: varying the level σ_I of intrinsic noise, the performance of the agent (dashed black line) stays always substantially higher than that of the single integrators, notably for stronger noise. As expected, the performance does deteriorate, but the decrease is surprisingly slight, considering that the maximum value attained by σ_I is comparable with the typical dynamical range of the integrators x_τ. **C**: evolution of the 'moving threshold' $\Delta\Sigma^{\mathcal{T}}$ (Eq. 22) for three values of $T_{\max}$. For higher values of $T_{\max}$ (see also Fig. 8), the moving threshold presents a peak whose position shifts with $T_{\max}$. **D**: w_{right} (Eq. 3) for different values of intrinsic noise σ_I (continuous lines are fourth degree polynomial fits for illustrative purposes). The peak of the lines, corresponding to the most exploited time scale, shifts towards lower τ values as σ_I increases.

3 Evolution during training

Fig 10. Learning is characterised by a non-monotonic adaptation of the average response time that is consequent to the necessity of finding a fine balance between integrating information and the cost of waiting to make decisions. **A**: Accuracy of the model for signals with different coherences across learning. **B**: Average response times and probability of not making a decision before the end of the episode, i.e. after $T_{\max}$. Trials with increasing level of coherences correspond to greater response times and greater probabilities of 'late' responses. The initial descending trend (around 100 episodes) of the response times common to all coherences is due to the initial ignorance of the agent about the nature of the task, on the tendency to avoid late decisions and to prefer immediate rewards.

Fig. 10 illustrates how the behaviour of the agent evolves as it encounters new episodes during learning. Fig. 10A shows the performance attained on average for four different values of signal coherence at different times during the training phase. The performance is of course always higher for higher values of coherence ('easier' episodes), and tends to increase monotonically for all the values of coherence during training. This monotonic trend is not preserved, instead, looking at the average response time (Fig. 10B). The response time drops at the beginning of training with values that are very close for every value of coherence. The reason for such behaviour is related to how the agent is initialised. At the beginning, the agent is ignorant about the rules of the task and pre-programmed to make a random choice after having waited for a finite random length of time. Without such random initialisation, the learning would not proceed, since the agent needs to perform actions to learn the relative consequences. While the agent is unable to tell apart signals with different coherences, the response time then decreases. In fact, longer average response times are detrimental due to late responses (no decision before the maximum time allowed $T_{\max}$) that are not rewarded. This is made clear in the inset, that shows how the fraction of late responses quickly drops to almost zero, and it stays there. Afterwards, the model starts to statistically differentiate between signals with different coherences (the four lines diverge) and the response time begins to rise. In this regime, waiting means accumulating more information and helps to improve the performance.

4 Discussion

Decision making and reinforcement learning are fields with overlapping contributions: both attempt to answer the question of how decisions are taken. In this work, we unified these two views by having a reinforcement learning agent solving a classical perceptual decision making task; in this we are not alone [9, 45–47]. This work is novel for studying multiple time scales, that arguably exist in the brain [35–42], and their effect on the decision making process, leading to surprising conclusions.

Here we have shown how the agent is able to learn an effective policy to solve the task in a relatively small number of episodes. Effective at least in the sense that the agent performs better than any single-time-scale drift-diffusion integrator. And that the performance curve of the model fits remarkably well with the psychophysical results. Also the experimental relationship between signal coherence and reaction time is semi-quantitatively reproduced, with differences that can be likely traced back to the fact that our agent does not include any non-decision delays. The agent's performance, moreover, is much more robust than the single accumulators' to variations of the task.

It comes therefore not unexpected that, at a more microscopic level, the policy devised by the agent after the training is markedly different from the one suggested by the drift-diffusion model, where the decision is taken when one of the integrating processes reach the decision threshold. Instead, the proposed agent makes decisions within short 'active' windows of time during which fleeting bursts in the probability of choosing an action make that action possible. We interpret such behaviour as arising from a 'surge' of probability in those short windows resulting from the broad agreement on the decision of many integrators with different time scales, akin to the concept of majority voting. This feature of the model could be in principle tested experimentally; in this respect, we note how it is compatible with the analysis performed in [54] on single-neuron single-trial spike trains in LIP area to uncover sudden activity jumps and their informativeness about choice.

We have moreover shown how the time course of the key variable in our model, *i.e.* the weighted combination of the integrators, reproduces many qualitative characteristics observed in the activity of neurons in LIP area during a motion-discrimination task [4, 48], notably what we have termed 'signal neutrality': the collapse of neuronal activities to a single trajectory for different values of signal to noise ratio. The model strongly suggests that such a collapse is due to the availability of multiple time scales: in fact, at the level of the single integrator and when the model is trained with only one of the integrators, the collapse is not observed.

Yet fluctuations too play a key role in signal neutrality. And indeed, as far as we can discern, the observation of the phenomenon in [19], where no multiple time scales are present, is rooted in the presence of large fluctuations in the activity traces being averaged. Such fluctuations are smoothed out in the model, by a first order filter and by the introduction of a random post-decision time, to form the decision; but nonetheless they give a major contribution to the observed collapse, as testified by peak values well above the decision threshold.

We have also shown how the distribution of response times exhibits a well defined scalar property [49, 50]: as the mean response time spans more than one order of magnitude, the shape of the distribution stays quite unchanged. The drift-diffusion model is capable of exhibiting a linear relationship between the mean response time and its standard deviation [51], a prerequisite for the scalar property. Yet, it seems difficult for the drift-diffusion model to show a clear scalar property (see Eq. 29), namely to produce response times with a constant coefficient of variation. Nonetheless, the experimental results reported in [51] seem to support to a degree the linear relationship (though the range of values of the coefficient of variation does not match the one from the drift-diffusion model), and not a proper scalar property. Yet the very low coefficients of variation for fast responses could result from ignoring the effect of non-decision times, that can be assumed to have low variability [22].

We have suggested that signal neutrality and the scalar property share a common origin in how the agent leverages the multiple time scales at its disposal to make its integration variable ($\Delta\Sigma^{\mathcal{T}}$) a stochastic process whose mean scales like a power-law in time, but with a nearly-constant variance. Such claim is corroborated by the observation that signal neutrality and the scalar property build up together during training; this represents a genuine prediction of the model, amenable to experimental testing.

One of the major distinctive points of this work is that the agent autonomously learns how to behave pursuing the maximization of reward, with no strategy *a priori* prescribed, not unlike a biological agent during a perceptual decision making experiment. Our model is abstract of any detailed biological elements and provide little information about the corresponding mechanisms at circuit level. Yet, similar to others, such as the drift-diffusion model, it provides useful insights to complex processes. In fact we argue that it might provide the right trade off between complexity and simplicity [55]: it does not model directly the decision making process but rather learns when to make actions, incorporating at the same time the concept of multiple time scales.

With this in mind, and given the evidence of reinforcement learning-compatible signaling in the brain [56, 57], we advance the hypothesis that the similarities between the model workings and the experimental results originate from a common high-level 'optimal' strategy for dealing with a volatile environment, discovered through the interaction with the rules of the task. Though the implementation substrates are very different and present very different degrees of complexity, this shared strategy can help shed light on general information-processing principles leveraged by the brain itself.

In this line of reasoning, we have shown how the agent, by means of the integrators 'accumulating' the passage of time, implements an effective decision threshold that changes non-monotonically over the course of each episode. Such moving threshold matches qualitatively the optimal strategy for fixed-duration tasks [8, 10]. Although the existence itself of 'collapsing boundaries' has been disputed [58], this results lends support to the presumed optimality of the strategy discovered by the agent.

Timescales have been implicit in the reinforcement learning framework, in the context of propagating information about the success (or failure) of the task in cases where reward is not immediate, see for instance eligibility traces [43, 59]. This is not entirely the same as the concept of time scales in this model, where the emphasis is on acquiring and retaining sensory information from the environment, not unlike what happens in the field of Reservoir Computing [60].

Admittedly abstract, the proposed framework allows for more detailed and biologically plausible implementations. In this respect, we note how the building blocks of the present model, *i.e.* the signal accumulators, have likely biological counterparts [61, 62], and at the same time have been subject to deep theoretical and modeling analysis [50, 54, 63]. One notable idea emerging from such analysis is that integrators with wildly different time scales, as required by a truly multi-scale system, can be effectively implemented by pools of noisy attractors. Attractor dynamics has been long one of the main staples of theoretical neuroscience, and as such it suffers no lack of detailed spiking implementations [64]. On the other hand, several winner-take-all spiking networks capable of implementing a probabilistic classification of noisy signal have been described in the literature [65, 66]. Therefore we see no conceptual barriers to a more detailed, spiking model mimicking the workings of the agent.

Beyond the specific interpretations provided by the model presented here, we would like to advocate the consideration of multiple time scales in models handling non-stationary and noisy information: there is indeed increasing evidence that performance is improved or becomes more robust to changes to the environment as a consequence. This is achieved thanks to the degeneracy offered by the wide range of time scales, that allows the agent to adapt effectively to different conditions. Such 'adaptive' degeneracy can be seen as an instance of a general strategy seemingly implemented in many biological systems: deploying a large number of elements performing similar functions in order to build 'sloppy' neutral regions in the space of conformations, where equivalent or nearly equivalent behaviours originate [67–70]. Just those neutral regions allow for rapid and effective adaptation. In our opinion, considering the complexity and unexpected shifts of the environments in which biological agents live and operate, modelling efforts — even when very abstract — should see robustness not just as an interesting characteristic, but as a paramount requisite.

5 Methods

5.1 Coherence

In [4], every three frames on the screen, a fraction c ('coherence') of dots are moved coherently in the chosen direction by $\mathrm{d}x$, while the other $1 - c$ dots are randomly displaced. We assume that each of the randomly moving dots is subjected to a change Δx in their position following a probability distribution, with $\langle \Delta x \rangle = 0$ and $\mathrm{Var}[\Delta x] = \sigma_x^2$. Imagining that neurons with different receptive fields help to estimate the average movement of the dots at each time step, we end up with a signal s of mean:

$$\mu \equiv \langle s \rangle = c\,\mathrm{d}x \tag{30}$$

and variance:

$$\sigma^2 \equiv \mathrm{Var}[s] = (1 - c)\,\sigma_x^2 \tag{31}$$

Then, we have the relationship:

$$\frac{\mu}{\sigma} = \frac{c}{\sqrt{1-c}}\frac{\mathrm{d}x}{\sigma_x} \tag{32}$$

or:

$$\mu \propto \frac{\text{coherence}}{\sqrt{100 - \text{coherence}}}, \tag{33}$$

where we have expressed the coherence as a percentage. Eq. 11 is a special case of this one, with a proportionality constant chosen to match experimental ranges.

5.2 Learning

The learning algorithm adopted is a reinforcement learning actor-critic model with eligibility traces [43]. The goal of reinforcement learning is to maximise the cumulative reward:

$$G_t = R_{t+1} + \gamma R_{t+2} + ... + \gamma^{t+N-1} R_{t+N} \tag{34}$$

where $0 \leq \gamma \leq 1$ is a discount factor that describes the tendency of the agent to invest in future rewards (here $\gamma \simeq 1 - 10^{-7}$). In our specific case, an episode ends when the agent chooses 'left' or 'right', or the maximum allowed time $T_{\max}$ is reached without a decision. Rewards are given at the end of the episode only. The reward is 1 for the case of correct decision, and 0 otherwise. The policy (the actor) is embodied by the probabilities p_a ($a =$ 'right', 'left' or 'wait') defined in Eq. 10. Instead, the parametrisation defining the critic, *i.e.* the value function $V(t) = \mathbf{E}_p\{G_t|s_t\}$, is:

$$V(t) = \sum_{\tau} \tilde{w}_\tau |x_\tau(t)| + \tilde{w}^{\mathcal{T}}_\tau x^{\mathcal{T}}_\tau(t) + b_v \tag{35}$$

where the weights $\tilde{w}$s are learned alongside the ws, and we used the absolute value of the integrators because, similarly to the definition of Σ_{wait}, positive and negative fluctuations of the signal should contribute in the same way to the expected reward. For each episode, the algorithm defines two sets of eligibility traces, $\mathbf{e}^{\tilde{w}}_t$ and $\mathbf{e}^{w}_t$, for the critic and the actor respectively:

$$\begin{aligned}
\mathbf{e}^{\tilde{w}}_t &= \gamma \lambda^{\tilde{w}} \mathbf{e}^{\tilde{w}}_{t-\Delta t} + \gamma^t \nabla_{\tilde{w}} V(t) \\
\mathbf{e}^{w}_t &= \gamma \lambda^{w} \mathbf{e}^{w}_{t-\Delta t} + \gamma^t \nabla_{w}\, p_a(t)
\end{aligned}$$

where t is the time inside an episode, and $0 \leq \lambda^w \leq 1$ and $0 \leq \lambda^{\tilde{w}} \leq 1$ are the traces decay parameters (here $\lambda^{\tilde{w}} = \lambda^w = 1 - 10^{-5}$). The parameters are then updated according to:

$$\begin{aligned}
&\delta \equiv R_{t+\Delta t} + \gamma V(t + \Delta t) - V(t) \\
&\boldsymbol{w} \leftarrow \boldsymbol{w} + \eta^w\, \delta\, \mathbf{e}^w \\
&\tilde{\boldsymbol{w}} \leftarrow \tilde{\boldsymbol{w}} + \eta^{\tilde{w}}\, \delta\, \mathbf{e}^{\tilde{w}}.
\end{aligned}$$

5.3 Rescaling σ_I for single integrators

When comparing the agent with a single integrator, we rescaled the amount of noise σ_I affecting the single integrator by a factor α_I, defined as

$$\alpha_I = \frac{1}{\sqrt{\sum_\tau w^2_{\text{right},\tau} / \max_\tau (w^2_{\text{right},\tau})}} \leq 1 \tag{36}$$

(we could equivalently use $w_{\text{left},\tau}$, since $w_{\text{left},\tau} \simeq -w_{\text{right},\tau}$).

Eq. 36 takes into account the fact that the agent effectively lowers the total noise by summing up n_τ integrators x_τ affected by independent sources of noise ξ_τ. Thus, $\alpha_I = 1$ when just one of the $w_{\text{right},\tau}$ is different from 0, *i.e.* when the agent utilises just one integrator. On the other hand, the maximum $\alpha_I = \frac{1}{\sqrt{n_\tau}}$ is attained when the agent weights equally all the integrators.

5.4 Signal neutrality and scalar property measures

To measure signal neutrality, we take the average $\Delta\Sigma_{\text{right}}(t)$, aligned to decision time, for six different coherences (0%, 3.2%, 6.4%, 12.8%, 25.6%, 51.2%); each curve is considered for an interval between 0 and 600 ms before the decision is taken; if the number of points to average for a given coherence drops below 100 before the 600 ms, the interval of definition of that curve is shrunk accordingly. We then rescale all the curves to fit inside the range 0-1, so that the minimum of the minimum values attained by each curve is 0; and the maximum of the maxima is 1. Then we compute, for each time, the maximum distance between any pairs of rescaled curves (this distance is of course always ≤ 1 thanks to the rescaling). Finally we take the average of such maximum distance, and take the inverse: this is the operative measure of signal neutrality used throughout the paper.

To give a measure of scalar property, we compute the coefficient of variation CV for the distribution of response times corresponding to six values of coherence (0%, 3.2%, 6.4%, 12.8%, 25.6%, 51.2%). We then take the inverse of the difference between the maximum and the minimum value of CV: this is the reported measure of the scalar proprety (see Fig. 7).

5.5 Mean and variance of $\Delta\Sigma^{\mathcal{S}}_{\text{right}}$

Being the signal $s(t)$ a constant μ plus a white noise W_t, Eq. 1 is a stochastic differential equation:

$$dx_\tau(t) = -\frac{x_\tau(t)-\mu}{\tau}\,\mathrm{d}t + \frac{\sigma}{\tau}\,dW_t \tag{37}$$

whose solution is (since by definition $x(0)_\tau = 0$):

$$x_\tau(t) = \mu\left(1-\exp\big(-\frac{t}{\tau}\big)\right) + \frac{\sigma}{\tau}\int_0^t \exp\big(\frac{s-t}{\tau}\big)\,dW_s. \tag{38}$$

It is to find how the mean value of $x_\tau(t)$ depends on t:

$$\langle x_\tau(t)\rangle = \mu\left(1-\exp\big(-\frac{t}{\tau}\big)\right). \tag{39}$$

In the following we will make use of the covariance $\langle (x_{\tau_1}(t)-\langle x_{\tau_1}(t)\rangle)\,(x_{\tau_2}(t)-\langle x_{\tau_2}(t)\rangle)\rangle$ for two generic τ_1 and τ_2; defining:

$$\tau^* \equiv \frac{\tau_1\,\tau_2}{\tau_1+\tau_2}, \tag{40}$$

and being $\delta(t)$ the Dirac delta function, we have:

$$\begin{aligned}\langle (x_{\tau_1}(t)-\langle x_{\tau_1}(t)\rangle)\,(x_{\tau_2}(t)-\langle x_{\tau_2}(t)\rangle)\rangle &= \frac{\sigma^2}{\tau_1\,\tau_2}\int_0^t\int_0^t \mathrm{e}^{\frac{u-t}{\tau_1}}\,\mathrm{e}^{\frac{v-t}{\tau_2}}\,\langle dW_u dW_v\rangle \\ &= \frac{\sigma^2}{\tau_1\,\tau_2}\int_0^t\int_0^t \mathrm{e}^{\frac{u-t}{\tau_1}}\,\mathrm{e}^{\frac{v-t}{\tau_2}}\,\delta(u-v)\,\mathrm{d}u\mathrm{d}v \\ &= \frac{\sigma^2}{\tau_1\,\tau_2}\int_0^t \mathrm{e}^{\frac{u-t}{\tau^*}}\,\mathrm{d}u = \sigma^2\,\frac{\tau^*}{\tau_1\,\tau_2}\left[1-\exp\big(-\frac{t}{\tau^*}\big)\right] \\ &= \frac{\sigma^2}{\tau_1+\tau_2}\left[1-\exp\big(-(\frac{1}{\tau_1}+\frac{1}{\tau_2})\,t\big)\right].\end{aligned} \tag{41}$$

Neglecting the intrinsic noises ξ_τ and $\xi^{\mathcal{T}}_\tau$ and assuming that the x_τs do not change sign during an episode, we can omit the absolute value in Eq. 7 and write, using Eq. 23:

$$\Delta\Sigma^{\mathcal{S}}_{\text{right}}(t|\mu,\,\sigma) = \sum_\tau \Delta w_\tau\,x_\tau(t). \tag{42}$$

The assumption about the sign of the x_τs holds on average, but not at the beginning of each episode when the x_τs are, by construction, close to 0. Disregarding this initial phase, though, the approximation is probably good most of the time. We can now easily recover, from Eqs. 39 and 41, the mean and the variance of $\Delta\Sigma^{\mathcal{S}}_{\text{right}}$, as given by Eqs. 24 and 25.

References

1. Skinner BF. Operant behavior. American psychologist. 1963;18(8):503.

2. Pisupati S, Chartarifsky-Lynn L, Khanal A, Churchland AK. Lapses in perceptual decisions reflect exploration. Elife. 2021;10:e55490.

3. Link S, Heath R. A sequential theory of psychological discrimination. Psychometrika. 1975;40(1):77–105.

4. Roitman JD, Shadlen MN. Response of neurons in the lateral intraparietal area during a combined visual discrimination reaction time task. Journal of neuroscience. 2002;22(21):9475–9489.

5. Gold JI, Shadlen MN. The neural basis of decision making. Annual review of neuroscience. 2007;30.

6. Hanks TD, Mazurek ME, Kiani R, Hopp E, Shadlen MN. Elapsed decision time affects the weighting of prior probability in a perceptual decision task. Journal of Neuroscience. 2011;31(17):6339–6352.

7. Bogacz R, Brown E, Moehlis J, Holmes P, Cohen JD. The physics of optimal decision making: a formal analysis of models of performance in two-alternative forced-choice tasks. Psychological review. 2006;113(4):700.

8. Drugowitsch J, Moreno-Bote R, Churchland AK, Shadlen MN, Pouget A. The cost of accumulating evidence in perceptual decision making. Journal of Neuroscience. 2012;32(11):3612–3628.

9. Rao RP. Decision making under uncertainty: a neural model based on partially observable markov decision processes. Frontiers in computational neuroscience. 2010;4:146.

10. Balsdon T, Wyart V, Mamassian P. Confidence controls perceptual evidence accumulation. Nature communications. 2020;11(1):1–11.

11. Gold JI, Shadlen MN. Banburismus and the brain: decoding the relationship between sensory stimuli, decisions, and reward. Neuron. 2002;36(2):299–308.

12. Liu Y, Blostein SD. Optimality of the sequential probability ratio test for nonstationary observations. IEEE Transactions on Information Theory. 1992;38(1):177–182.

13. Bogacz R. Optimal decision-making theories: linking neurobiology with behaviour. Trends in cognitive sciences. 2007;11(3):118–125.

14. Moran R. Optimal decision making in heterogeneous and biased environments. Psychonomic bulletin & review. 2015;22(1):38–53.

15. Rahnev D, Denison RN. Suboptimality in perceptual decision making. Behavioral and Brain Sciences. 2018;41.

16. Gold JI, Shadlen MN. Neural computations that underlie decisions about sensory stimuli. Trends in cognitive sciences. 2001;5(1):10–16.

17. Marcos E, Cos I, Girard B, Verschure PF. Motor cost influences perceptual decisions. PLoS One. 2015;10(12):e0144841.

18. Lamichhane B, Adhikari BM, Dhamala M. The activity in the anterior insulae is modulated by perceptual decision-making difficulty. Neuroscience. 2016;327:79–94.

19. Mazurek ME, Roitman JD, Ditterich J, Shadlen MN. A role for neural integrators in perceptual decision making. Cerebral cortex. 2003;13(11):1257–1269.

20. Brown SD, Heathcote A. The simplest complete model of choice response time: Linear ballistic accumulation. Cognitive psychology. 2008;57(3):153–178.

21. Pedersen ML, Endestad T, Biele G. Evidence accumulation and choice maintenance are dissociated in human perceptual decision making. PloS one. 2015;10(10):e0140361.

22. Stine GM, Zylberberg A, Ditterich J, Shadlen MN. Differentiating between integration and non-integration strategies in perceptual decision making. Elife. 2020;9:e55365.

23. Ratcliff R. A theory of memory retrieval. Psychological review. 1978;85(2):59.

24. Wald A, Wolfowitz J. Optimum character of the sequential probability ratio test. The Annals of Mathematical Statistics. 1948; p. 326–339.

25. Roxin A. Drift–diffusion models for multiple-alternative forced-choice decision making. The Journal of Mathematical Neuroscience. 2019;9(1):5.

26. Ratcliff R, Cherian A, Segraves M. A comparison of macaque behavior and superior colliculus neuronal activity to predictions from models of two-choice decisions. Journal of neurophysiology. 2003;90(3):1392–1407.

27. Ratcliff R, McKoon G. The diffusion decision model: theory and data for two-choice decision tasks. Neural computation. 2008;20(4):873–922.

28. Ratcliff R, Smith PL. A comparison of sequential sampling models for two-choice reaction time. Psychological review. 2004;111(2):333.

29. Smith PL, Ratcliff R. Psychology and neurobiology of simple decisions. Trends in neurosciences. 2004;27(3):161–168.

30. Busemeyer JR, Townsend JT. Decision field theory: a dynamic-cognitive approach to decision making in an uncertain environment. Psychological review. 1993;100(3):432.

31. Usher M, McClelland JL. The time course of perceptual choice: the leaky, competing accumulator model. Psychological review. 2001;108(3):550.

32. Ratcliff R, Hasegawa YT, Hasegawa RP, Smith PL, Segraves MA. Dual diffusion model for single-cell recording data from the superior colliculus in a brightness-discrimination task. Journal of neurophysiology. 2007;97(2):1756–1774.

33. Lukoševičius M, Jaeger H. Reservoir computing approaches to recurrent neural network training. Computer Science Review. 2009;3(3):127–149.

34. Schrauwen B, Verstraeten D, Van Campenhout J. An overview of reservoir computing: theory, applications and implementations. In: Proceedings of the 15th european symposium on artificial neural networks. p. 471-482 2007; 2007. p. 471–482.

35. Pritchard WS. The brain in fractal time: 1/f-like power spectrum scaling of the human electroencephalogram. International Journal of Neuroscience. 1992;66(1-2):119–129.

36. Kiebel SJ, Daunizeau J, Friston KJ. A hierarchy of time-scales and the brain. PLoS computational biology. 2008;4(11).

37. Wacongne C, Labyt E, van Wassenhove V, Bekinschtein T, Naccache L, Dehaene S. Evidence for a hierarchy of predictions and prediction errors in human cortex. Proceedings of the National Academy of Sciences. 2011;108(51):20754–20759.

38. Linkenkaer-Hansen K, Nikouline VV, Palva JM, Ilmoniemi RJ. Long-range temporal correlations and scaling behavior in human brain oscillations. Journal of Neuroscience. 2001;21(4):1370–1377.

39. Honey CJ, Kötter R, Breakspear M, Sporns O. Network structure of cerebral cortex shapes functional connectivity on multiple time scales. Proceedings of the National Academy of Sciences. 2007;104(24):10240–10245.

40. La Camera G, Rauch A, Thurbon D, Luscher HR, Senn W, Fusi S. Multiple time scales of temporal response in pyramidal and fast spiking cortical neurons. Journal of neurophysiology. 2006;96(6):3448–3464.

41. Wasmuht DF, Spaak E, Buschman TJ, Miller EK, Stokes MG. Intrinsic neuronal dynamics predict distinct functional roles during working memory. Nature communications. 2018;9(1):1–13.

42. Cavanagh SE, Towers JP, Wallis JD, Hunt LT, Kennerley SW. Reconciling persistent and dynamic hypotheses of working memory coding in prefrontal cortex. Nature communications. 2018;9(1):1–16.

43. Sutton RS, Barto AG. Reinforcement learning: An introduction. MIT press; 2018.

44. Lee D, Seo H, Jung MW. Neural basis of reinforcement learning and decision making. Annual review of neuroscience. 2012;35:287–308.

45. Law CT, Gold JI. Reinforcement learning can account for associative and perceptual learning on a visual-decision task. Nature neuroscience. 2009;12(5):655.

46. Lepora NF. Threshold learning for optimal decision making. In: Advances in Neural Information Processing Systems; 2016. p. 3763–3771.

47. Pedersen ML, Frank MJ, Biele G. The drift diffusion model as the choice rule in reinforcement learning. Psychonomic bulletin & review. 2017;24(4):1234–1251.

48. Shadlen MN, Newsome WT. Neural basis of a perceptual decision in the parietal cortex (area LIP) of the rhesus monkey. Journal of neurophysiology. 2001;86(4):1916–1936.

49. Gibbon J, Malapani C, Dale CL, Gallistel CR. Toward a neurobiology of temporal cognition: advances and challenges. Current opinion in neurobiology. 1997;7(2):170–184.

50. Cao R, Pastukhov A, Mattia M, Braun J. Collective activity of many bistable assemblies reproduces characteristic dynamics of multistable perception. Journal of Neuroscience. 2016;36(26):6957–6972.

51. Wagenmakers EJ, Brown S. On the linear relation between the mean and the standard deviation of a response time distribution. Psychological review. 2007;114(3):830.

52. Churchland AK, Kiani R, Shadlen MN. Decision-making with multiple alternatives. Nature neuroscience. 2008;11(6):693–702.

53. Simon HA. Rational choice and the structure of the environment. Psychological review. 1956;63(2):129.

54. Latimer KW, Yates JL, Meister ML, Huk AC, Pillow JW. Single-trial spike trains in parietal cortex reveal discrete steps during decision-making. Science. 2015;349(6244):184–187.

55. Wilson RC, Collins AG. Ten simple rules for the computational modeling of behavioral data. Elife. 2019;8:e49547.

56. Schultz W, Dayan P, Montague PR. A neural substrate of prediction and reward. Science. 1997;275(5306):1593–1599.

57. Lin Z, Nie C, Zhang Y, Chen Y, Yang T. Evidence accumulation for value computation in the prefrontal cortex during decision making. Proceedings of the National Academy of Sciences. 2020;117(48):30728–30737.

58. Hawkins GE, Forstmann BU, Wagenmakers EJ, Ratcliff R, Brown SD. Revisiting the evidence for collapsing boundaries and urgency signals in perceptual decision-making. Journal of Neuroscience. 2015;35(6):2476–2484.

59. Vasilaki E, Frémaux N, Urbanczik R, Senn W, Gerstner W. Spike-based reinforcement learning in continuous state and action space: when policy gradient methods fail. PLoS Comput Biol. 2009;5(12):e1000586.

60. Manneschi L, Ellis MO, Gigante G, Lin AC, Del Giudice P, Vasilaki E. Exploiting multiple timescales in hierarchical echo state networks. arXiv preprint arXiv:210104223. 2021;.

61. Kiani R, Hanks TD, Shadlen MN. Bounded integration in parietal cortex underlies decisions even when viewing duration is dictated by the environment. Journal of Neuroscience. 2008;28(12):3017–3029.

62. Meister ML, Hennig JA, Huk AC. Signal multiplexing and single-neuron computations in lateral intraparietal area during decision-making. Journal of Neuroscience. 2013;33(6):2254–2267.

63. Braun J, Mattia M. Attractors and noise: twin drivers of decisions and multistability. Neuroimage. 2010;52(3):740–751.

64. Amit DJ, Brunel N. Model of global spontaneous activity and local structured activity during delay periods in the cerebral cortex. Cerebral cortex (New York, NY: 1991). 1997;7(3):237–252.

65. Wang XJ. Probabilistic decision making by slow reverberation in cortical circuits. Neuron. 2002;36(5):955–968.

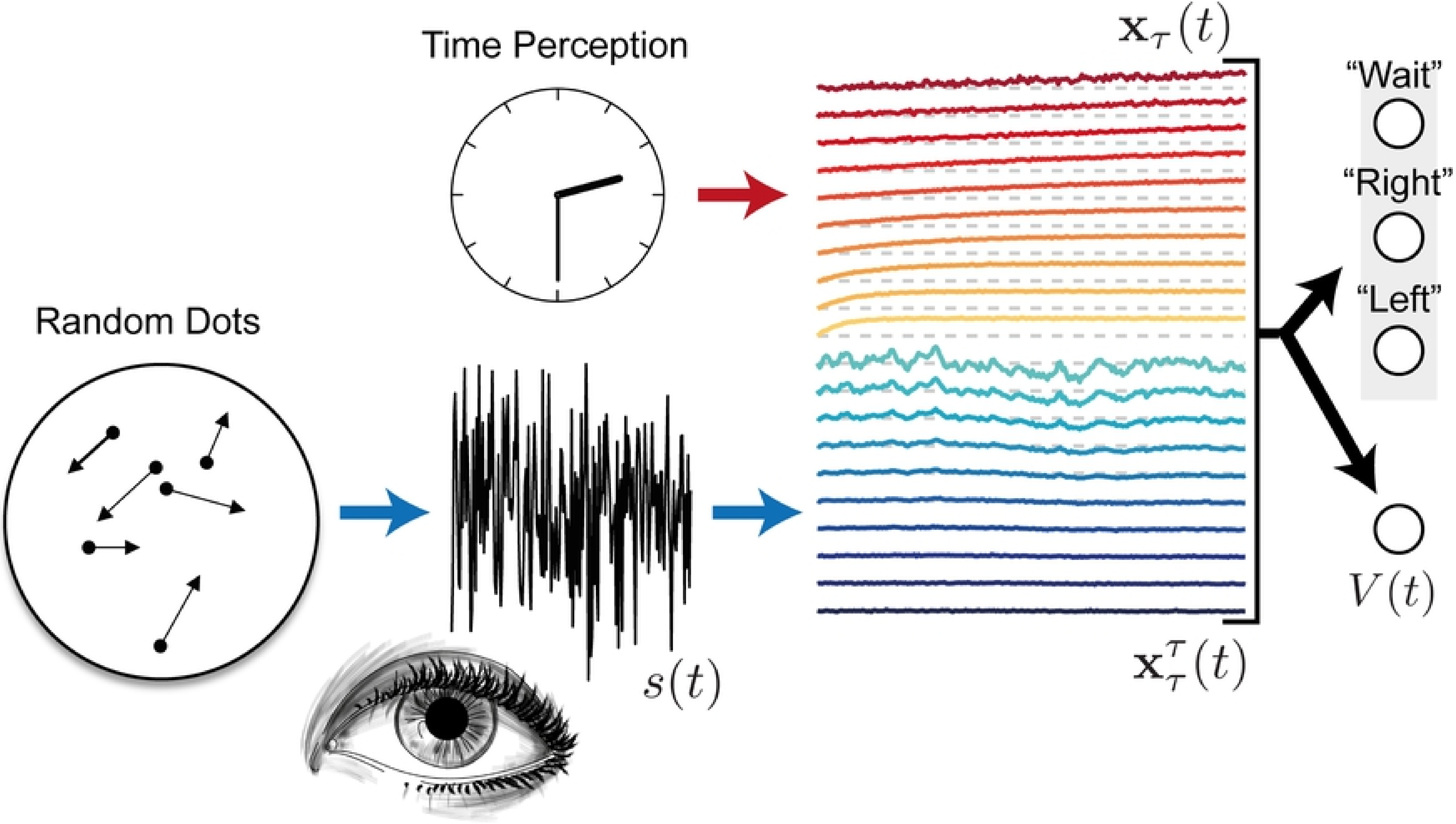
66. Wang XJ. Decision making in recurrent neuronal circuits. Neuron. 2008;60(2):215–234.

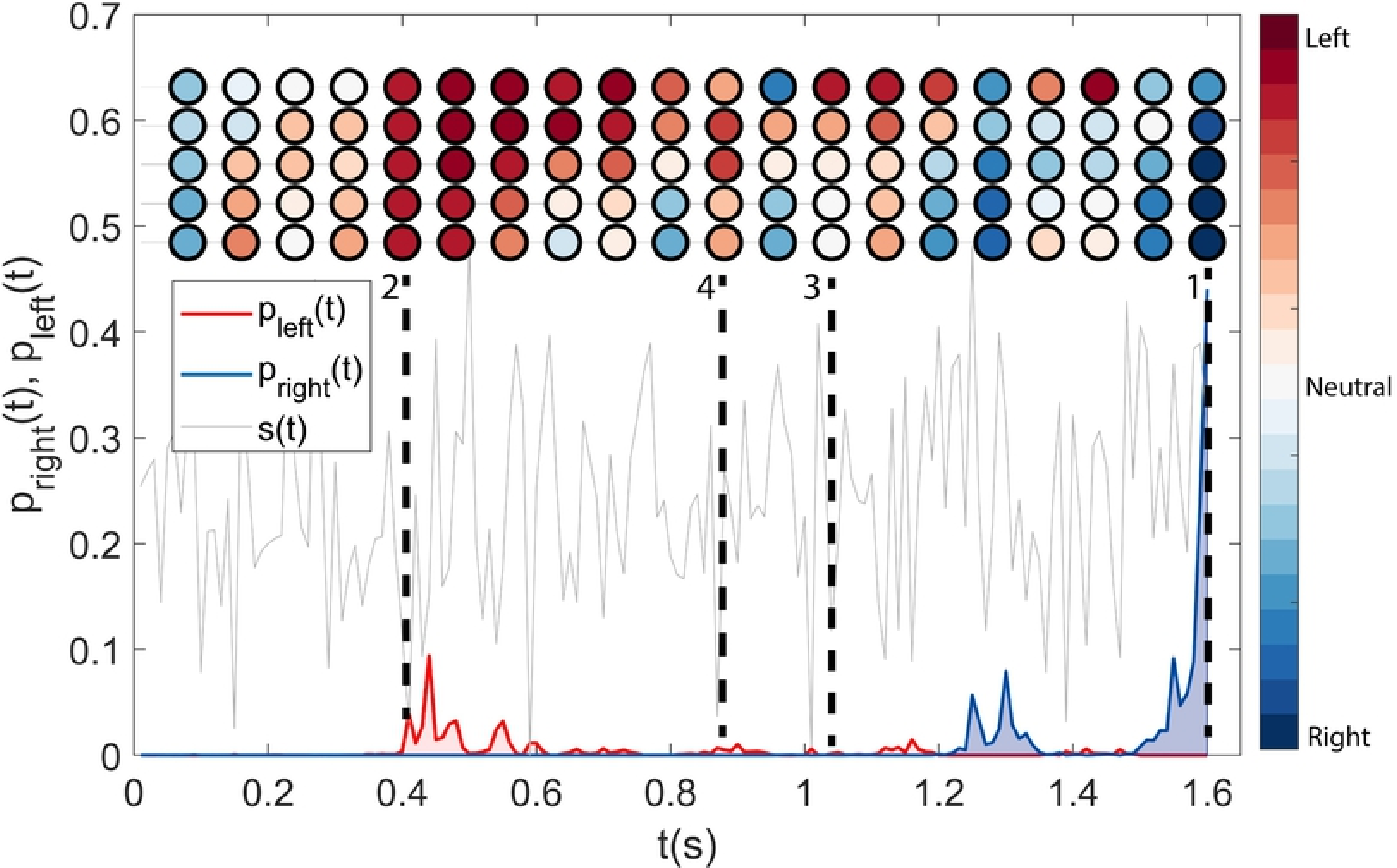
67. Edelman GM, Gally JA. Degeneracy and complexity in biological systems. Proceedings of the National Academy of Sciences. 2001;98(24):13763–13768.

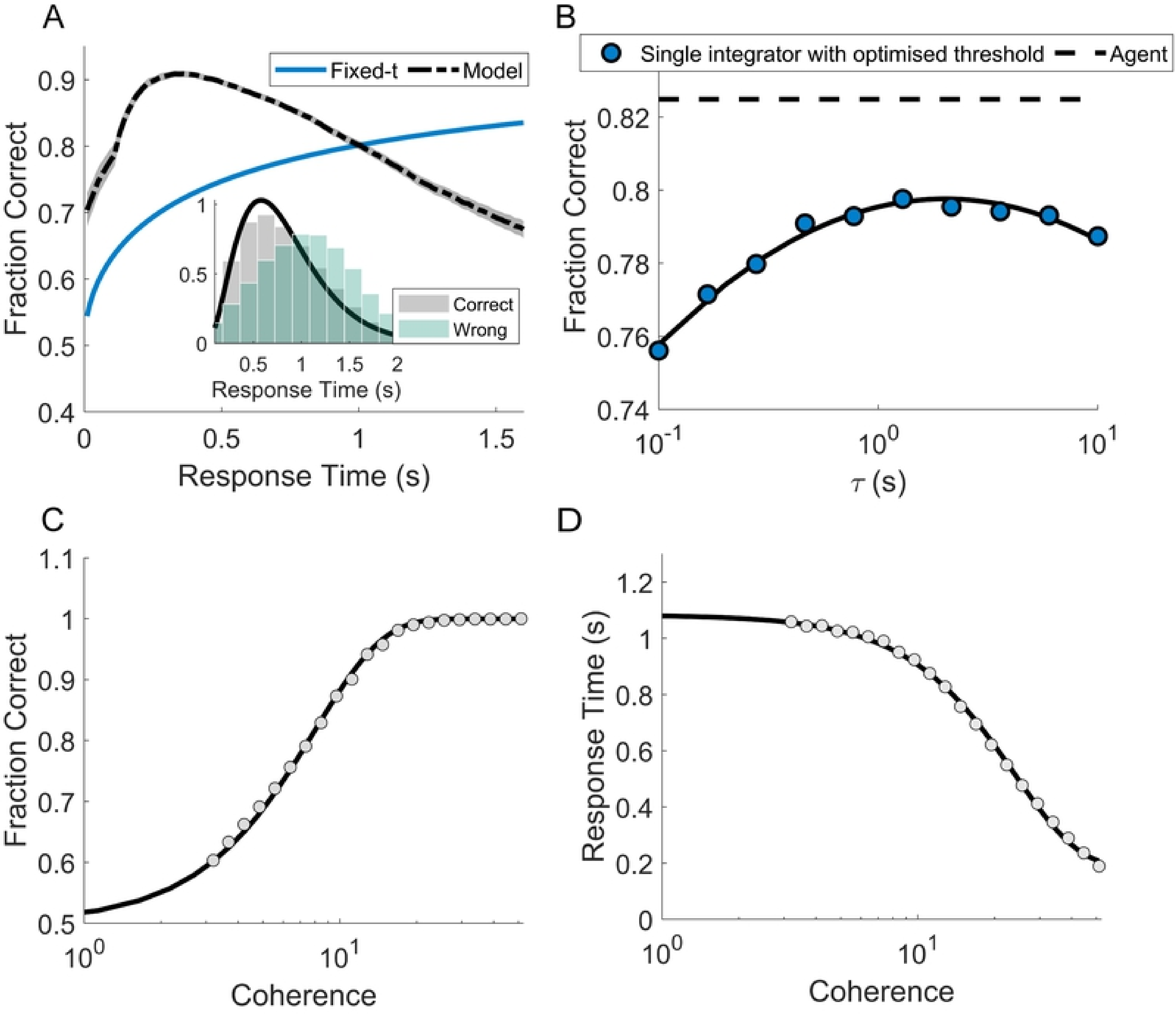
68. Gutenkunst RN, Waterfall JJ, Casey FP, Brown KS, Myers CR, Sethna JP. Universally sloppy parameter sensitivities in systems biology models. PLoS Comput Biol. 2007;3(10):e189.

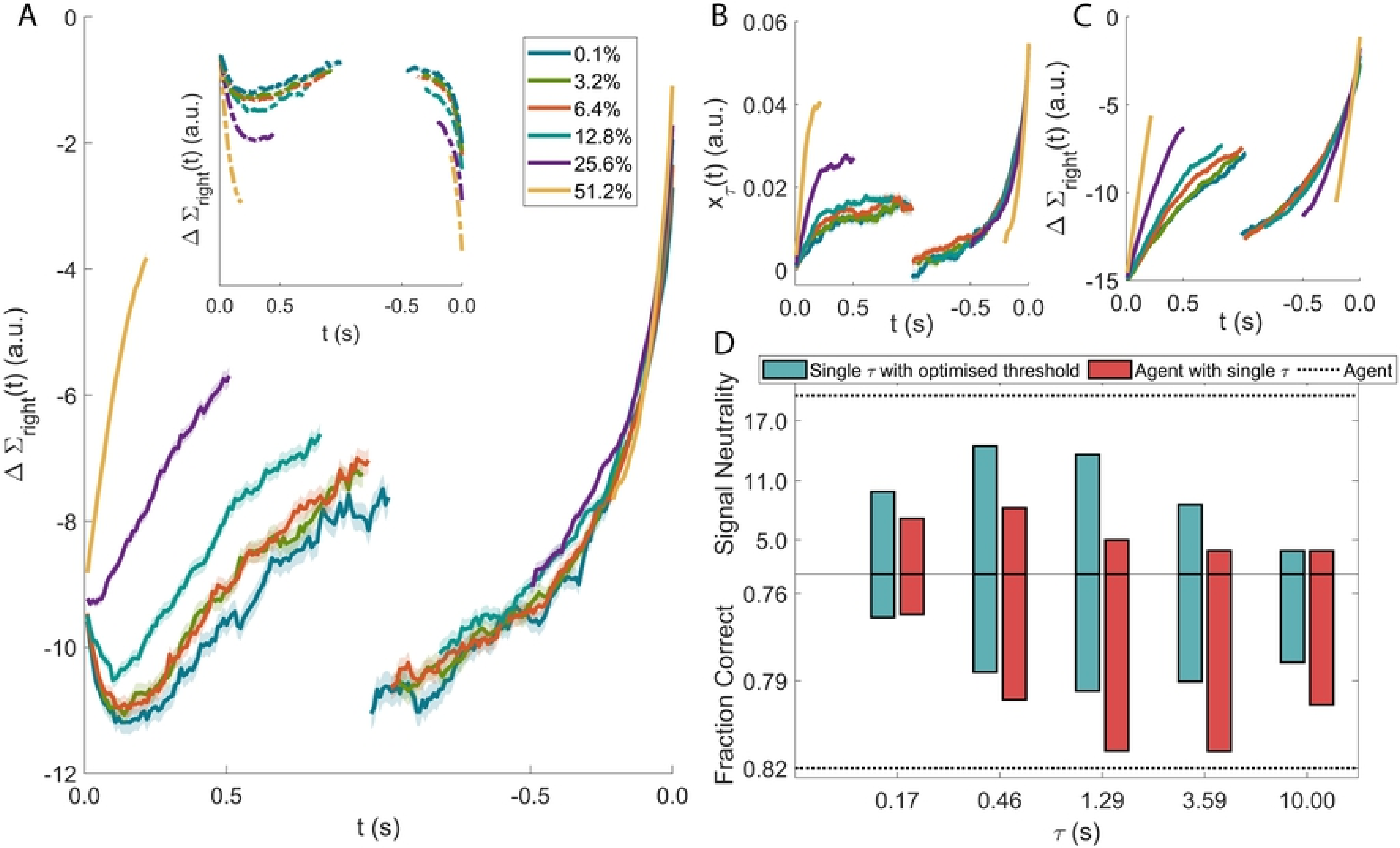
69. Daniels BC, Chen YJ, Sethna JP, Gutenkunst RN, Myers CR. Sloppiness, robustness, and evolvability in systems biology. Current opinion in biotechnology. 2008;19(4):389–395.

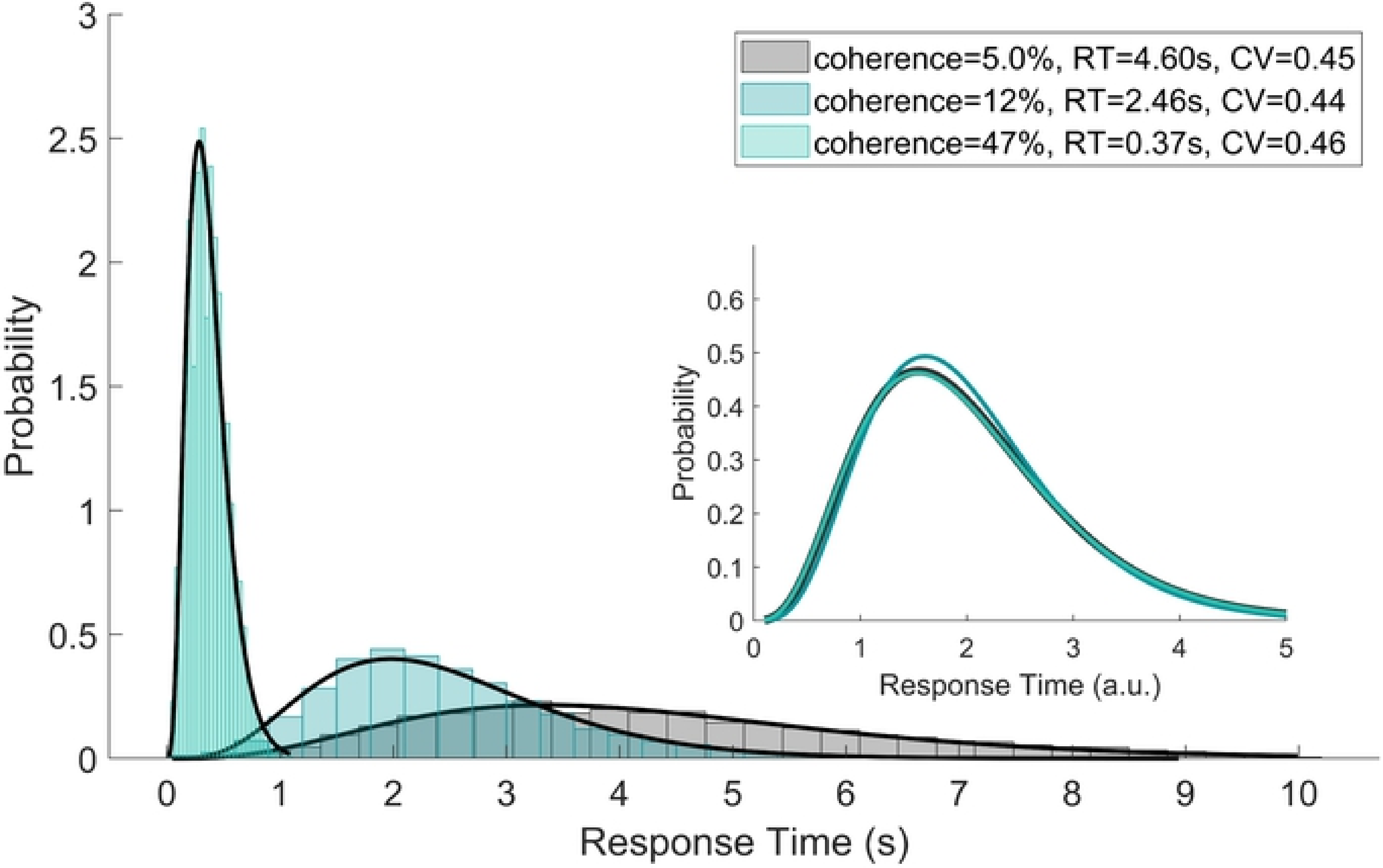
70. Panas D, Amin H, Maccione A, Muthmann O, van Rossum M, Berdondini L, et al. Sloppiness in spontaneously active neuronal networks. Journal of Neuroscience. 2015;35(22):8480–8492.

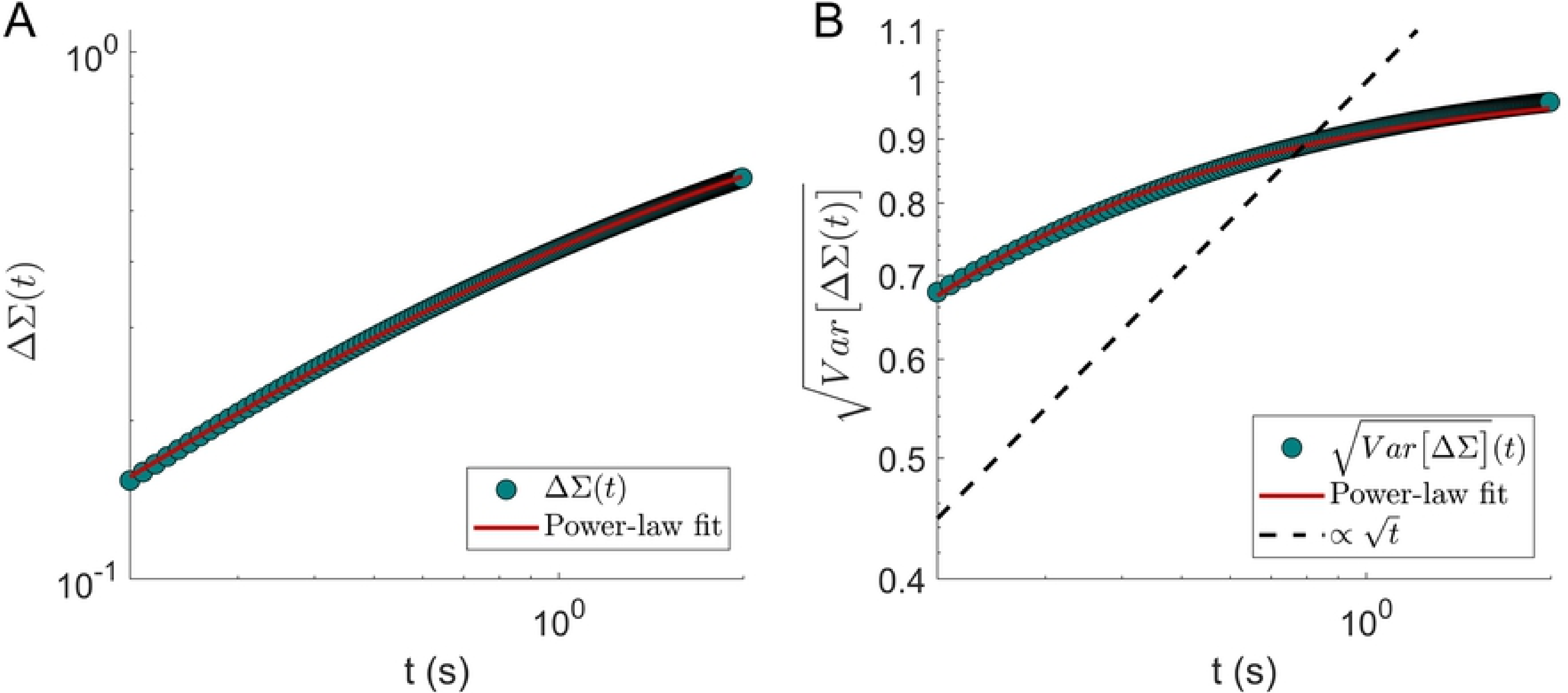

Time Perception
$\mathbf{x}_\tau(t)$
"Wait"
"Right"
"Left"
Random Dots
$s(t)$
$V(t)$
$\mathbf{x}_\tau^\tau(t)$

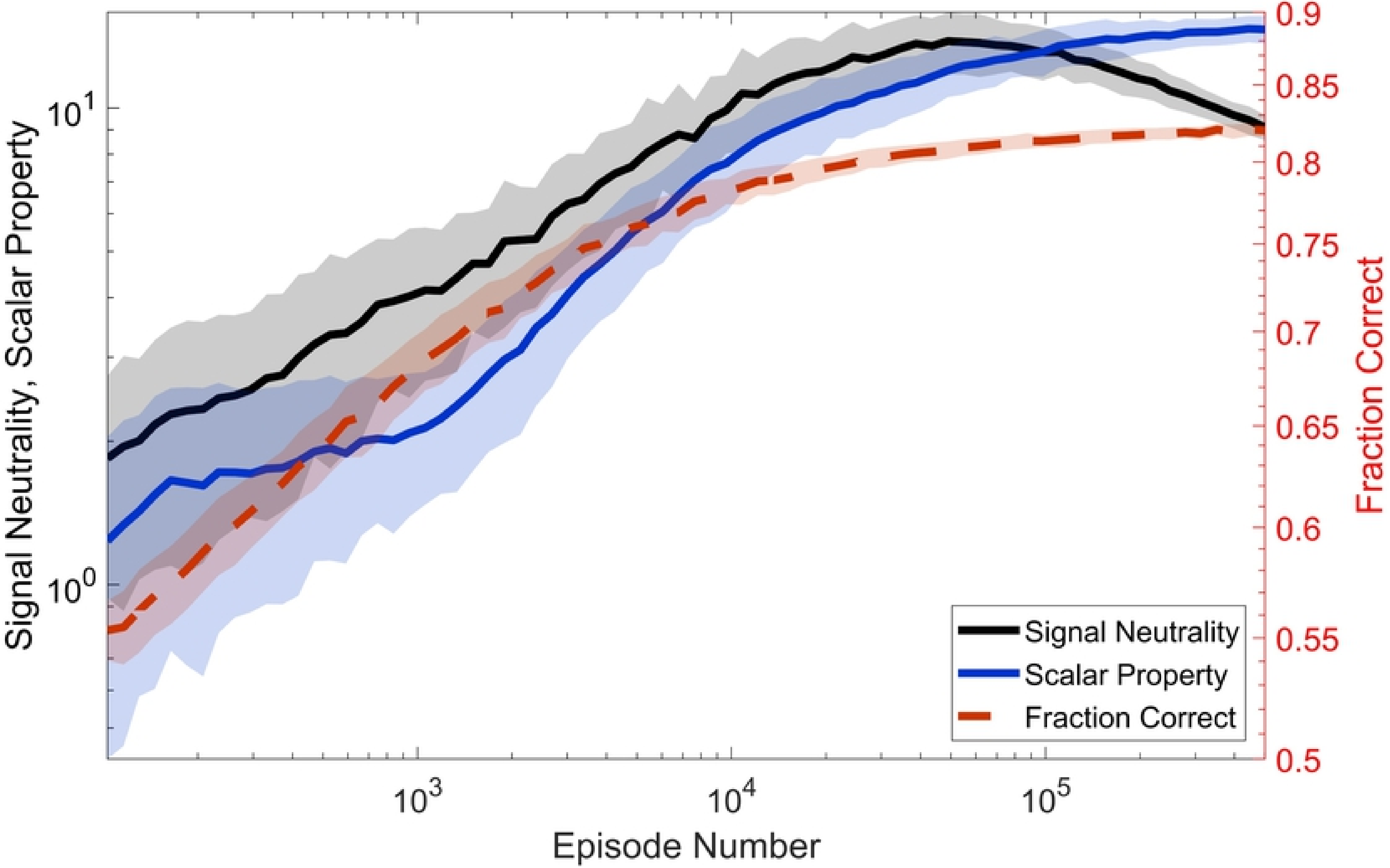

0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
$p_{right}(t)$, $p_{left}(t)$
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
t(s)
2
4
3
1
$p_{left}(t)$
$p_{right}(t)$
s(t)
Left
Neutral
Right

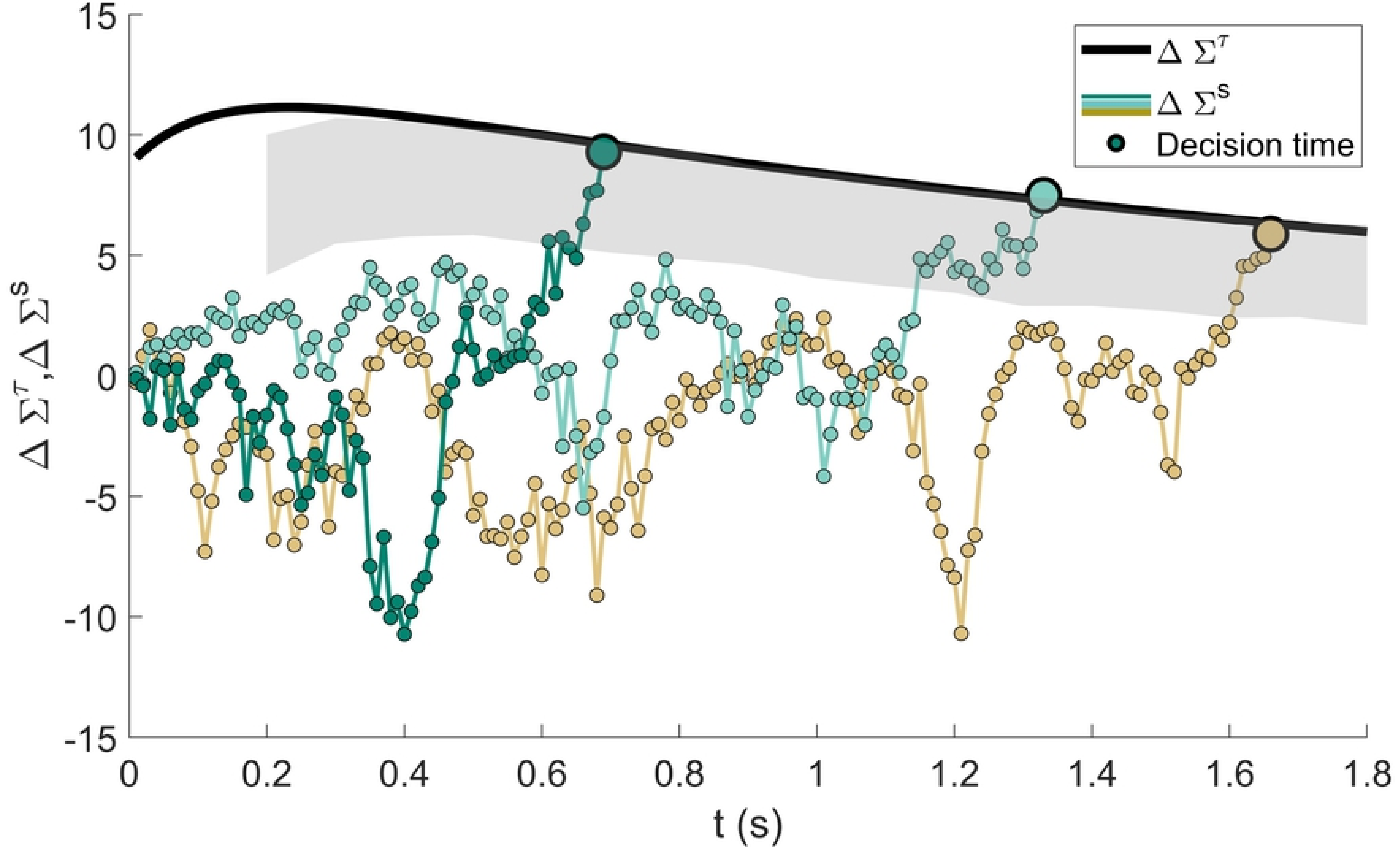

A
Fixed-t
Model
Fraction Correct
0.4
0.5
0.6
0.7
0.8
0.9
0
0.5
1
1.5
Response Time (s)
0
0.5
1
0.5
1
1.5
2
Response Time (s)
Correct
Wrong
B
Single integrator with optimised threshold
Agent
Fraction Correct
0.74
0.76
0.78
0.8
0.82
10^{-1}
10^{0}
10^{1}
τ (s)
C
Fraction Correct
0.5
0.6
0.7
0.8
0.9
1
1.1
10^{0}
10^{1}
Coherence
D
Response Time (s)
0
0.2
0.4
0.6
0.8
1
1.2
10^{0}
10^{1}
Coherence


A
0
-2
-4
-6
-8
-10
-12
$\Delta\ \Sigma_{right}(t)$ (a.u.)
0.0
0.5
-0.5
0.0
t (s)
0.1%
3.2%
6.4%
12.8%
25.6%
51.2%
$\Delta\ \Sigma_{right}(t)$ (a.u.)
0.0
0.5
-0.5
0.0
t (s)
B
0.06
0.04
0.02
0
$x_\tau(t)$ (a.u.)
0.0
0.5
-0.5
0.0
t (s)
C
0
-5
-10
-15
$\Delta\ \Sigma_{right}(t)$ (a.u.)
0.0
0.5
-0.5
0.0
t (s)
D
Single τ with optimised threshold
Agent with single τ
Agent
Signal Neutrality
17.0
11.0
5.0
Fraction Correct
0.76
0.79
0.82
0.17
0.46
1.29
3.59
10.00
τ (s)

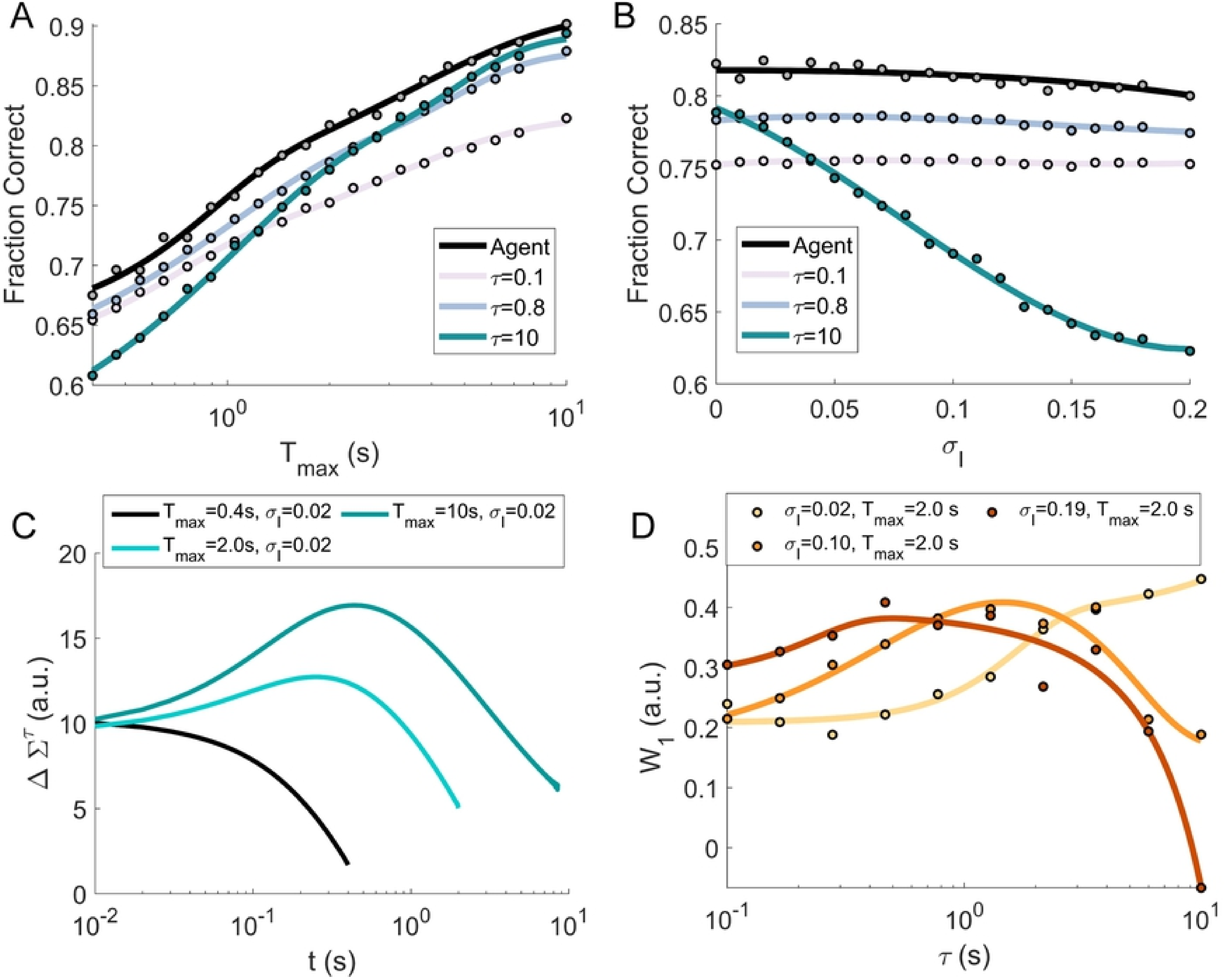

coherence=5.0%, RT=4.60s, CV=0.45
coherence=12%, RT=2.46s, CV=0.44
coherence=47%, RT=0.37s, CV=0.46
Probability
0
0.5
1
1.5
2
2.5
3
Response Time (s)
0
1
2
3
4
5
6
7
8
9
10
Probability
0
0.1
0.2
0.3
0.4
0.5
0.6
Response Time (a.u.)
0
1
2
3
4
5

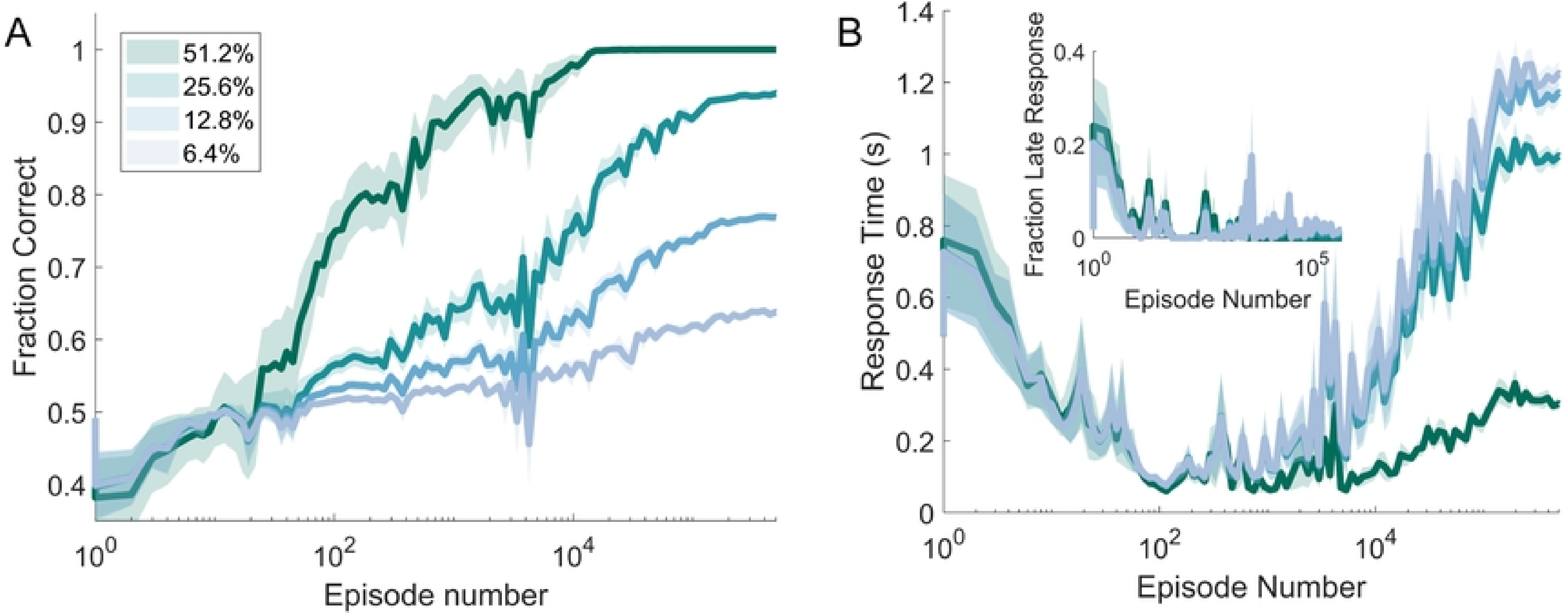

A
10^0
10^{-1}
$\Delta\Sigma(t)$
$\Delta\Sigma(t)$
Power-law fit
10^0
t (s)
B
1.1
1
0.9
0.8
0.7
0.6
0.5
0.4
$\sqrt{Var[\Delta\Sigma(t)]}$
$\sqrt{Var[\Delta\Sigma](t)}$
Power-law fit
$\propto \sqrt{t}$
10^0
t (s)


Signal Neutrality, Scalar Property
10^1
10^0
Fraction Correct
0.9
0.85
0.8
0.75
0.7
0.65
0.6
0.55
0.5
10^3
10^4
10^5
Episode Number
Signal Neutrality
Scalar Property
Fraction Correct



$\Delta\ \Sigma^{\tau}$
$\Delta\ \Sigma^{s}$
Decision time
$\Delta\ \Sigma^{\tau}, \Delta\ \Sigma^{s}$
15
10
5
0
-5
-10
-15
t (s)
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8


A
Fraction Correct
0.9
0.85
0.8
0.75
0.7
0.65
0.6
10^0
10^1
T_{max} (s)
Agent
τ=0.1
τ=0.8
τ=10
B
Fraction Correct
0.85
0.8
0.75
0.7
0.65
0.6
0
0.05
0.1
0.15
0.2
σ_I
Agent
τ=0.1
τ=0.8
τ=10
C
T_{max}=0.4s, σ_I=0.02
T_{max}=10s, σ_I=0.02
T_{max}=2.0s, σ_I=0.02
$\Delta\,\Sigma^{\tau}$ (a.u.)
20
15
10
5
0
10^{-2}
10^{-1}
10^0
10^1
t (s)
D
σ_I=0.02, T_{max}=2.0 s
σ_I=0.19, T_{max}=2.0 s
σ_I=0.10, T_{max}=2.0 s
W_1 (a.u.)
0.5
0.4
0.3
0.2
0.1
0
10^{-1}
10^0
10^1
τ (s)


A
51.2%
25.6%
12.8%
6.4%
Fraction Correct
1
0.9
0.8
0.7
0.6
0.5
0.4
10^0
10^2
10^4
Episode number
B
Response Time (s)
1.4
1.2
1
0.8
0.6
0.4
0.2
0
10^0
10^2
10^4
Episode Number
Fraction Late Response
0.4
0.2
0
10^0
10^5
Episode Number