Instability with a purpose: how the visual brain makes decisions in a volatile world

Robin Cao^{1,2}, Alexander Pastukhov¹, Stepan

Aleshin¹, Maurizio Mattia^{3*}, and Jochen Braun^{1*}

¹*Cognitive Biology, Center for Behavioral Brain Sciences, Magdeburg, Germany*

²*Gatsby Computational Neuroscience Unit, London, United Kingdom and*

³ *Istituto Superiore di Sanità, Rome, Italy*

(Dated: June 9, 2020)

Abstract

In ambiguous or conflicting sensory situations, perception is often ‘multistable’ in that it changes abruptly at irregular intervals, shifting perpetually between distinct alternatives. Intriguingly, the interval statistics of these alternations exhibits quasi-universal characteristics, suggesting a general mechanism. Here we show that the stereotypical features of multistable perception, exemplified by binocular rivalry, are reproduced in detail by a hierarchical dynamics operating out of equilibrium. Its constitutive elements are discretely stochastic and idealize the metastability of cortical networks. Independent elements accumulate visual evidence at one level, while groups of coupled elements compete for dominance at another level. As soon as one group dominates perception, feedback inhibition suppresses supporting evidence. This mechanism is corroborated compellingly by unexpected serial dependencies of perceptual alternations. Moreover, it satisfies normative constraints of continuous decision-making. We conclude that multistable perception reflects decision-making in a volatile world: integrating evidence over space and time, choosing categorically between hypotheses, while concurrently evaluating alternatives.

* These authors contributed equally to this work.

INTRODUCTION

In deducing the likely physical causes of sensations, perception goes beyond the immediate sensory evidence and draws heavily on context and prior experience [5, 44, 100, 118]. Numerous illusions in visual, auditory, and tactile perception – all subjectively compelling, but objectively false – attest to this extrapolation beyond the evidence. In natural settings, perception explores alternative plausible causes of sensory evidence by active repositioning of sensors [“active perception”, 78, 85, 129]. In general, perception is thought to actively select plausible explanatory hypotheses, to predict the sensory evidence expected for each hypothesis from prior experience, and to compare the observed sensory evidence at multiple levels of scale or abstraction [“analysis by synthesis”, “predictive coding”, “hierarchical Bayesian inference”, 86, 91, 97, 130]. Active inference engages the entire hierarchy of sensory cortical areas, including both feedforward and feedback projections [4, 40, 60, 87, 107].

The dynamics of active inference becomes experimentally observable when perceptual illusions are ‘multistable’ [65]. In numerous ambiguous or conflicting situations, phenomenal experience switches at irregular intervals between discrete alternatives, even though the sensory scene is stable [3, 83, 95, 96, 102, 105, 126]. Multistable illusions are enormously diverse, involving visibility or audibility, perceptual grouping, visual depth or motion, and many kinds of sensory scenes, from schematic to naturalistic. Average switching rates differ greatly and range over at least two orders of magnitude [19], depending on sensory scene, perceptual grouping [53, 112, 125], continuous or intermittent presentation [64, 72], attentional condition [88], individual observer [16, 32, 89], and many other factors.

In spite of this diversity, the stochastic properties of multistable phenomena appear to be quasi-universal, suggesting that the underlying mechanisms may be general. Firstly, average dominance duration depends in a characteristic and counter-intuitive manner on the strength of dominant and suppressed evidence [“Levelt’s propositions I–IV”, 14, 15, 48, 51, 66, 80]. Secondly, the statistical distribution of dominance durations shows a stereotypical shape, namely a Gamma distribution with shape parameter $r \simeq 3-4$ [“scaling property”, 8, 12, 13, 19, 28, 32, 39, 82, 88, 119]. Thirdly, the durations of successive dominance periods are correlated positively, over at least two or three periods [32, 39, 114, 119].

Here we show that these quasi-universal characteristics are comprehensively and quantitatively reproduced, indeed guaranteed, by an interacting hierarchy of birth-death processes. operating out of equilibrium. While the proposed mechanism combines key features of previous models, it far surpasses their explanatory power.

Several possible mechanisms have been proposed for perceptual dominance, the triggering of reversals, and the stochastic timing of reversals. That a single, coherent interpretation typically dominates phenomenal experience is thought to reflect competition (explicit or implicit) at the level of explanatory hypotheses [*e.g.*, 27], sensory inputs [*e.g.*, 63], or both [*e.g.*, 127]. That a dominant interpretation is periodically supplanted by a distinct alternative has been attributed to fatigue processes [*e.g.*, neural adaptation, synaptic depression, 59], spontaneous fluctuations [‘noise’, *e.g.*, 50, 128], stochastic sampling [*e.g.*, 104], or combinations of these [*e.g.*, adaptation and noise, 89, 106, 108]. The characteristic stochasticity (Gamma-like distribution) of dominance durations has been attributed to Poisson counting processes [*e.g.*, birth-death processes, 19, 41, 111] or stochastic accumulation of discrete samples [82, 104, 110, 124].

‘Dynamical’ models combining competition, adaptation, and noise capture well the characteristic dependence of dominance durations on input strength (“Levelt’s propositions”) [2, 59, 128], especially when inputs are normalized [23, 79, 80], and when the dynamics emphasize noise [89, 106, 108]. However, such models do not preserve a Gamma-like distribution over the full range of input strengths [19, 23]. On the other hand, ‘sampling’ models based on discrete random processes produce Gamma-like distributions, [19, 82, 104, 110, 111, 124], but fail to reproduce the dependence on input strength. Neither type of model accounts for the sequential dependence of dominance durations [59].

Here we reconcile ‘dynamical’ and ‘sampling’ approaches to multistable perception, extending an earlier effort [41]. Importantly, every part of the proposed mechanism is justified normatively in that it serves to optimize perceptual choices in a general behavioural situation, namely, continuous inference in uncertain and volatile environments [9, 117]. We propose that sensory inputs are represented by birth-death processes, in order to accumulate sensory information over time and in a format suited for Bayesian inference [70, 93]. Further, we suggest that explanatory hypotheses are evaluated competitively, with a hypothesis attaining

dominance (over phenomenal experience) when its support exceeds the alternatives by a certain finite amount, consistent with optimal decision making between multiple alternatives [9]. Finally, we assume that a dominant hypothesis suppresses its supporting evidence, as required by ‘predictive coding’ implementations of hierarchical Bayesian inference [46, 90, 97]. In contrast to many previous models, we do not require a mechanism of fatigue, adaptation, or decay.

Based on these assumptions, the proposed mechanism reproduces dependence on input strength, as well distribution of dominance durations and positive sequential dependence. Additionally, it predicts novel and unsuspected dynamical features confirmed by experiment.

RESULTS

Next we introduce each component of the mechanism with its normative justification and describe the reversal dynamics resulting from the interaction of all components. Further below, we compare model predictions with multistable perception of human observers, specifically, the dominance statistics of binocular rivalry at various combinations of left- and right-eye contrasts (**Fig. 1a**).

Hierarchical dynamics

Bistable assemblies: ‘local attractors’

As operative units of sensory representation, we postulate neuronal assemblies with bistable ‘attractor’ dynamics. Effectively, assembly activity moves in an energy landscape with two distinct quasi-stable states – dubbed ‘on’ and ‘off’ – separated by a ridge (**Fig. 1b**). Driven by noise, assembly activity mostly remains near one quasi-stable state (‘on’ or ‘off’), but occasionally ‘escapes’ to the other state [29, 45, 47, 55, 67].

An important feature of ‘attractor’ dynamics is that the energy of quasi-stable states depends sensitively on external input. Net positive input destabilizes (*i.e.*, raises the potential of) the ‘off’ state and stabilizes (*i.e.*, lowers the potential of) the ‘on’ state. Transition rates $\nu^\pm$ depend approximately *exponentially* on the height of the energy ridge (‘activation

Figure 1. Proposed mechanism of binocular rivalry. **(a)** When the left and right eyes see incompatible images, phenomenal appearance reverses at irregular intervals, sometimes being dominated by one image and sometimes by the other (grey and white regions). Sir Charles Wheatstone studied this multistable percept with a mirror stereoscope (not as shown!). **(b)** Spiking neural network implementation of a ‘local attractor’. An assembly of 150 neurons (schematic, dark gray circle) interacts competitively with multiple other assemblies (light gray circles). Population activity of the assembly explores an effective energy landscape (right) with two distinct steady-states (circles), separated by a ridge (diamond). Driven by noise, activity transitions occasionally between ‘on’ and ‘off’ states (bottom), with transition rates $\nu^{\pm}$ depending sensitively on external input to the assembly (not shown). Here, $\nu^{+}=\nu^{-}\approx 1$ Hz. Spike raster shows 10 representative neurons. **(c)** Nested attractor dynamics (schematic) which quantitatively reproduces the dynamics of binocular rivalry. Independently bistable variables (‘local attractors’, small circles) respond probabilistically to input, transitioning stochastically between on- and off-states (red and white, respectively). The entire system comprises four pools, with 25 variables each, linked by excitatory and inhibitory projections. Phenomenal appearance is decided by competition between decision pools R and R' forming ‘non-local attractors’ (cross-inhibition w_{comp} and self-excitation w_{coop}). Visual input c and c' accumulates, respectively, in evidence pools E and E' and propagates to decision pools (feedforward selective excitation w_{exc} and indiscriminate inhibition w_{inh}). Decision pools suppress associated evidence pools (feedback selective suppression w_{supp}). **(d)** Representative example of model dynamics. The time-course of the number of active variables (active count) is shown for decision pools (top) and evidence pools (bottom), representing the left eye (red traces) and the right eye image (blue traces). Left and right pools are shown both separately (left and middle columns, respectively) and together (right column). The state of individual variables (black horizontal traces in left and middle columns) and of perceptual dominance (grey and white regions) are also shown. In decision pools, almost all variables become active (thick trace) or inactive (thin trace) simultaneously. In evidence pools, only a small fraction of variables is active at any given time.

energy’).

Figure 1b illustrates ‘attractor’ dynamics for an assembly of 150 spiking neurons with activity levels of approximately 7 Hz and 21 Hz per neuron in the ‘off’ and ‘on’ states, respectively. Further details are given in the supplementary text (**Metastable population dynamics** and **Fig. 8**). Note that our model is independent of network details and relies only on an idealized description.

Binary stochastic variables

Our model reduces bistable assemblies to discretely stochastic, binary activity variables $x(t) \in \{0, 1\}$, which activate and inactivate with Poisson rates ν^+ and ν^- , respectively. These rates $\nu^\pm(s)$ vary exponentially with activation energy $\Delta u = u(s) + u_0$:

$$\nu^+ = \nu/2 \exp(\Delta u/2), \quad \nu^- = \nu/2 \exp(-\Delta u/2) \quad (1)$$

where u_0 and ν are baseline potential and baseline rate, respectively, and where input-dependent part $u(s) = ws$ varies linearly with input s , with synaptic coupling constant w (supplementary section **Metastable population dynamics** and **Fig. 8e**).

Pool of N binary variables

An extended network, containing N individually bistable assemblies with shared input s , reduces to a ‘pool’ of N binary activity variables $x_i(t) \in \{0, 1\}$ with identical rates $\nu_\pm(s)$. Although all variables are independently stochastic, they are coupled through their shared input s . The number of active variables $n(t) = \sum_i x_i(t)$ or, equivalently, the active fraction $x(t) = n(t)/N$, forms a discretely stochastic process (“birth-death” or “Ehrenfest” process [49]).

Relaxation dynamics

While activity $x(t)$ develops discretely and stochastically according to Eq. 5 (Methods), its expectation $\langle x(t) \rangle$ develops continuously and deterministically,

$$\langle \dot{x} \rangle = (1 - \langle x \rangle) \nu^+ - \langle x \rangle \nu^- \quad (2)$$

relaxing with characteristic time $\tau_x = \frac{1}{\nu^+ + \nu^-}$ towards asymptotic value $x_\infty = \frac{\nu^+}{\nu^+ + \nu^-}$. As rates $\nu^\pm$ change with input s (Eq. 1), so do $\tau_x = \Upsilon(s)$ and $x_\infty = \Phi(s)$ (see Methods). Characteristic

time τ_x is longest for small input $s \simeq 0$ and shortens for larger positive or negative input $|s| \gg 0$. Asymptotic value x_∞ ranges over the interval $[0, 1]$ and varies sigmoidally with input s , reaching half-activation for $s = -u_0/w$.

Quality of representation

Pools of bistable variables belong to a class of neural representations particularly suited for Bayesian integration of sensory information [6, 93]. In general, summation of activity is equivalent to optimal integration of information, provided that response variability is Poisson-like, and response tuning differs only multiplicatively [70, 71]. Pools of bistable variables closely approximate these properties (see supplementary text **Quality of representation: Suitability for inference**).

The representational accuracy of even a comparatively small number of bistable variables can be surprisingly high. For example, if normally distributed inputs drive the activity of initially inactive pools of bistable variables, pools as used in the present model ($N=25$, $w=2.5$) readily capture 90% of the Fisher information (supplementary text **Quality of representation: Integration of noisy samples**).

Conflicting evidence

Any model of binocular rivalry must represent the conflicting evidence from both eyes (*e.g.*, different visual orientations), which supports alternative perceptual hypotheses (*e.g.*, distinct grating patterns). We assume that conflicting evidence accumulates in two separate pools of $N=25$ bistable variables, E and E' , ('evidence pools', **Fig. 1c**). Fractional activations $e(t)$ and $e'(t)$ develop stochastically following Eq. 5 (Methods). Transition rates $\nu_e^\pm$ and $\nu_{e'}^\pm$ vary exponentially with activation energy (Eq. 1), with baseline potential u_e^0 and baseline rate ν_e . The variable components of activation energy, u_e and $u_{e'}$, are synaptically modulated by image contrasts, c and c' :

$$\begin{aligned} u_e &= w_{vis} I \\ u_{e'} &= w_{vis} I' \end{aligned} \tag{3}$$

where w_{vis} is a coupling constant and $I=f(c) \in [0, 1]$ is a monotonic function of image contrast $c \in [0, 1]$ (see Methods).

Competing hypotheses: ‘non-local attractors’

Once evidence for, and against, alternative perceptual hypotheses (*e.g.*, distinct grating patterns) has been accumulated, reaching a decision requires a sensitive and reliable mechanism for identifying the best supported hypothesis and amplifying the result into a categorical read-out. Such a winner-take-all decision [52] is readily accomplished by a dynamical version of biased competition [30, 31, 120, 121].

We assume that alternative perceptual hypotheses are represented by two further pools of $N = 25$ bistable variables, R and R' , forming two ‘non-local attractors’ (‘decision pools’, **Fig. 1c**). Similar to previous models of decision making and attentional selection [30, 31, 120, 121], we postulate recurrent excitation within pools, but recurrent inhibition between pools, to obtain a ‘winner-take-all’ dynamics. Importantly, we assume that ‘evidence pools’ project to ‘decision pools’ not only in the form of selective excitation (targeted at the corresponding decision pool), but also in the form of indiscriminate inhibition (targeting both decision pools), as suggested previously [10, 34].

Specifically, fractional activations $r(t)$ and $r'(t)$ develop stochastically according to Eq. 5 (Methods). Transition rates $\nu_s^\pm$ and $\nu_{s'}^\pm$ vary exponentially with activation energy (Eq. 1), with baseline difference u_r^0 and baseline rate ν_r . The variable components of activation energy, u_r and $u_{r'}$, are synaptically modulated by evidence and decision activities:

$$\begin{aligned} u_r &= w_{exc} e - w_{inh} (e + e') + w_{coop} r - w_{comp} r' \\ u_{r'} &= w_{exc} e' - w_{inh} (e + e') + w_{coop} r' - w_{comp} r \end{aligned} \quad (4)$$

where coupling constants w_{exc} , w_{inh} , w_{coop} , w_{comp} reflect feedforward excitation, feedforward inhibition, lateral cooperation within decision pools, and lateral competition between decision pools, respectively.

This biased competition circuit expresses a categorical decision by either raising r towards unity (and lowering r' towards zero) or *vice versa*. The choice is random when visual input is ambiguous, $I \simeq I'$, but becomes deterministic with growing input bias $|I - I'| > 0$. This probabilistic sensitivity to input bias is reliable and robust under arbitrary initial conditions of e , e' , r and r' (see supplementary section **Categorical choice** with **Fig. 10**).

Feedback suppression

Finally, we assume feedback suppression, with each decision pool selectively targeting the corresponding evidence pool. A functional motivation for this systematic bias *against* the currently dominant appearance is given momentarily. Its effects include curtailment of dominance durations and ensuring that reversals occur from time to time. Specifically, we modify Eq. 3 to

$$\begin{aligned} u_e &= w_{vis} f(c) - w_{supp} r \\ u_{e'} &= w_{vis} f(c') - w_{supp} r' \end{aligned} \quad (3')$$

where w_{supp} is a coupling constant.

Previous models of binocular rivalry [27, 46] have justified selective feedback suppression of the evidence supporting a winning hypothesis in terms of “predictive coding” and “hierarchical Bayesian inference” [61, 97]. An alternative normative justification is that, in volatile environments, where the sensory situation changes frequently (‘volatility prior’), optimal inference requires an exponentially growing bias *against* evidence for the most likely hypothesis [117]. Note that feedback suppression applies selectively to evidence for a winning hypothesis and is thus materially different from visual adaptation [123], which applies indiscriminately to all evidence present.

Reversal dynamics

A representative example of the joint dynamics of evidence and decision pools is illustrated in **Fig. 1cd**, both at the level of pool activities $e(t)$, $e'(t)$, $r(t)$, $r'(t)$, and at the level of individual bistable variables $x(t)$. The top row shows decision pools R and R' , with instantaneous active counts, $Nr(t)$ and $Nr'(t)$ and active / inactive states of individual variables $x(t)$. The bottom row shows evidence pools E and E' , with instantaneous active counts, $Ne(t)$ and $Ne'(t)$ and active / inactive states of individual variables $x(t)$. Only a small fraction of evidence variables is active at any one time.

Phenomenal appearance reverses when the differential activity $\Delta e = e - e'$ of evidence pools, E and E' , *contradicts* sufficiently strongly the differential activity $\Delta r = r - r'$ of decision pools, R and R' , such that the steady-state of decision pools is destabilized (see further below

and **Fig. 5**). As soon as the reversal has been effected at the decision level, feedback suppression *lifts from* the newly non-dominant evidence, and *descends upon* the newly dominant evidence. Due to this asymmetric suppression, the newly non-dominant evidence recovers, whereas the newly dominant evidence habituates. This opponent dynamics progresses, past the point of equality $s \simeq s'$, until differential evidence activity $\Delta e = e - e'$ once again *contradicts* differential decision activity $\Delta r = r - r'$. Accordingly, the respective times at which evidence and decision activities reverse are phase-shifted by approximately 90° relative to each other (**Fig. 1d**), consistent with previous models [1, 41, 124].

Figure 2. Dependence of mean dominance duration on dominant and suppressed image contrast (“Levelt’s propositions”). **(a)** Mean dominance duration $\langle T \rangle$ (color scale), as a function of dominant contrast c_{dom} and suppressed contrast c_{sup} , in model (left) and experiment (right). **(b)** Model prediction (solid traces) and experimental observation (dashed traces and symbols) compared. Levelt I & II: weak increase of $\langle T \rangle$ with c_{dom} when $c_{sup}=1.0$ (red traces and symbols), and strong decrease with c_{sup} when $c_{dom}=1.0$ (brown traces and symbols). Levelt III: Symmetric increase of $\langle T \rangle$ with c_{dom} (red traces and symbols) and decrease with c_{sup} (brown traces and symbols), when $c_{dom}+c_{sup}=100\%$. Alternation rate (green traces and symbols) peaks at equidominance and decreases symmetrically to either side. **(c)** Levelt IV: Decrease of $\langle T \rangle$ with image contrast, when $c_{sup}=c_{sup}$. **(d)** Predicted dependence of sequential correlation CC_1 (color scale) on c_{dom} and c_{sup} . **(e)** Model prediction (black trace, $N=25$) and experimental observation (blue trace and symbols, mean $\pm$ SEM, Spearman’s rank correlation ρ), when $c_{sup}=c_{sup}$. Also shown is a second model prediction (red trace, $N=40$).

Binocular rivalry

To compare predictions of the model described above to experimental observations, we measured spontaneous reversals of binocular rivalry for different combinations of image contrast. Binocular rivalry is a particularly well studied instance of multistable perception [15, 33, 65, 66, 126]. When conflicting images are presented to each eye (*e.g.*, by means of a mirror stereoscope or of colored glasses, see Methods), the phenomenal appearance reverses from time to time between the two images (**Fig. 1a**). Importantly, the perceptual conflict involves also representations of coherent (binocular) patterns and is not restricted to eye-specific (monocular) representations [7, 11, 54, 68].

Specifically, our experimental observations established reversal sequences for 5×5 combinations of image contrast, $c_{dom}, c_{sup} \in \{1/16, 1/8, 1/4, 1/2, 1\}$, where c_{dom} pertains to the phenomenally dominant image and c_{sup} to the phenomenally suppressed image (see Methods). We analysed these observations in terms of mean dominance durations $\langle T \rangle$, higher moments c_v and γ_1/c_v of the distribution of dominance durations, and sequential correlation cc_1 of successive dominance durations. Additional aspects of serial dependence are discussed further below.

As described in Methods, we fitted twelve (12) model parameters to reproduce observations with more than 50 degrees of freedom: 5×5 mean dominance durations $\langle T \rangle$, coefficients of variation c_v , and values of skewness $\gamma_1/c_v=2$, as well as a uniform correlation coefficient $cc_1=0.06$. Model predictions and experimental observations are juxtaposed in **Fig. 2** and **Fig. 3**.

The complex and asymmetric dependence of mean dominance durations on image contrast — aptly summarized by Levelt’s ‘propositions’ I to IV [15, 66] — is fully reproduced by the model (**Fig. 2a**). Increasing the contrast of one image, increases the fraction of time during which this image dominates appearance (‘predominance’, Levelt I). Counter-intuitively, this is due more to shortening dominance of the *unchanged* image than to lengthening dominance of the *changed* image (Levelt II, **Fig. 2b**, left panel). Mean dominance durations grow (and alternation rates decline) symmetrically around equal predominance as contrast difference $c_{dom} - c_{sup}$ increases (Levelt III, **Fig. 2b**, right panel). Mean dominance durations

Figure 3. Shape of dominance distribution depends only weakly on image contrast (‘scaling property’). Distribution shape is parametrized by coefficient of variation c_v and relative skewness γ_1/c_v . **(a)** Coefficient of variation c_v (color scale), as a function of dominant contrast c_{dom} and suppressed contrast c_{sup} , in model (left) and experiment (right). **(b)** Model prediction (solid traces) and experimental observation (dashed traces and symbols) compared. Left: increase of c_v with c_{dom} (red traces and symbols), and symmetric decrease with c_{sup} (brown traces and symbols), when $c_{sup} + c_{sup} = 1.0$. Right: weak dependence when $c_{dom} = c_{sup}$ (black traces and symbols). **(c)** Predicted dependence of relative skewness γ_1/c_v (gray scale) on c_{dom} and c_{sup} . **(d)** Model prediction (solid traces) and experimental observation (blue dashed trace and symbols, mean $\pm$ SEM), when $c_{dom} = 1 - c_{sup}$.

shorten when both image contrasts $c_{dom} = c_{sup}$ increase (Levelt IV, **Fig. 2c**).

Successive dominance durations are typically correlated positively [39, 89, 119]. Averaging over all contrast combinations, observed and predicted correlation coefficients were comparable with $cc_1 = 0.06 \pm 0.06$ (mean and standard deviation). Unexpectedly, both observed and predicted correlations coefficients increased systematically with image contrast ($\rho = 0.9$, $p < .01$), growing from $cc_1 = 0.02 \pm 0.05$ at $c_{dom} = c_{sup} = 1/16$ to 0.21 ± 0.06 at $c_{dom} = c_{sup} = 1$ (**Fig. 2e**, blue symbols). To our knowledge, this contrast-sensitivity of sequential correlations has not been observed previously.

The distribution of dominance durations typically takes a characteristic shape [8, 12, 13, 19, 28, 32, 39, 82, 88, 119], approximating a Gamma distribution with shape parameter $r \simeq 3-4$, or coefficient of variation $c_v = 1/\sqrt{r} \simeq 0.5-0.6$. The fitted model fully reproduces this ‘scaling property’ (**Fig. 3a**). The observed coefficient of variation remained in the range $c_v \simeq 0.5-0.6$ for nearly all contrast combinations (**Fig. 3b**). Unexpectedly, both observed and predicted values increased above, or decreased below, this range at extreme contrast combinations (**Fig. 3b**, left panel). At intermediate contrast, where observed values of skewness were most reliable, both observed and predicted values of skewness were $\gamma_1/c_v \simeq 2$ and conformed to a Gamma distribution (**Fig. 3d**, blue symbols).

Reasons for success

What are the reasons for the surprising success of the model described above? Which features of its dynamics are responsible for reproducing different aspects of binocular rivalry? Here we explain why no component of the model may reasonably be omitted and how highly non-trivial interactions between these components guarantee the universal characteristics of multistable phenomena.

Levelt’s propositions I, II and III

The characteristic input-dependence of average dominance durations emerges in two steps. First, inputs and feedback suppression shape the birth-death dynamics of evidence pools E and E' (by setting disparate transition rates $\nu^\pm$, following Eq. 3’ and Eq. 1). Second, this sets in motion two opponent developments (habituation of dominant evidence activity and recovery of non-dominant evidence activity, both following Eq. 2) that jointly determine dominance duration.

To elucidate this mechanism, it is helpful to consider the limit of large pools ($N \rightarrow \infty$) and its deterministic dynamics (**Fig. 4a**), which corresponds to the *average* stochastic dynamics. In this limit, periods of dominant evidence E or E' start and end at the same levels ($e_{start}=e'_{start}$ and $e_{end}=e'_{end}$), because reversal thresholds Δ_{rev} are the same for evidence difference $e-e'$ and $e'-e$ (see section *Levelt IV* below).

Figure 4. Joint dynamics of evidence habituation and recovery. Exponential development of evidence activities is governed by input-dependent asymptotic values and characteristic times. **(a)** Fractional activities e (blue traces) and e' (red traces) of evidence pools E and E' , respectively, over several dominance periods for unequal stimulus contrast ($c=7/8$, $c'=1/8$). Stochastic reversals of finite system ($N = 25$ units per pool, left) and deterministic reversals of infinite system ($N \rightarrow \infty$, right). Perceptual dominance (decision activity) is indicated along the upper margin (red or blue stripe). Dominance evidence habituates (*dom*), and non-dominant evidence recovers (*sup*), until evidence contradicts perception sufficiently (black vertical lines) to trigger a reversal (grey and white regions). **(b)** Development of stronger-input evidence e (blue) and weaker-input evidence e' (red), over two successive dominance periods ($c=15/16$, $c'=1/16$). Activities recover, or habituate, exponentially until reversal threshold Δ_{rev} is reached. Thin curves extrapolate to the respective asymptotic values, e_∞ and e'_∞ . Dominance durations depend on distance Δ_∞ and on characteristic times τ_e and $\tau_{e'}$. Left: incrementing non-dominant evidence e (dashed curve), raises upper asymptotic value e_∞ and shortens dominance T' by $\Delta T'$. Right: incrementing dominant evidence e (dashed curve), raises lower asymptotic value e_∞ and shortens dominance T by ΔT . **(c)** Increasing asymptotic activity difference Δ_∞ accelerates the development of differential activity and curtails dominance periods T , T' (and *vice versa*). As the dependence is hyperbolic, any change to Δ_∞ disproportionately affects longer dominance periods. If $T > T'$, then $\Delta T > \Delta T'$ (and *vice versa*).

The rates at which evidence habituates or recovers depend, in the first instance, on asymptotic levels e_∞ and e'_∞ (Eq. 1 and 2, **Fig. 4b** and **Fig. 11**). In general, dominance durations depend on distance Δ_∞ between asymptotic levels: the further apart these are,

the faster the development and the shorter the duration. As feedback suppression *inverts* the sign of the opponent developments, dominant evidence decreases (habituates) while non-dominant evidence increases (recovers). Due to this inversion, Δ_∞ is roughly proportional to $e_\infty^{non-dom} - e_\infty^{dom} + w_{supp}$. It follows that distance Δ_∞ is *smaller* and the reversal dynamics *slower* when dominant input is *stronger*, and *vice versa*. It further follows that incrementing one input (and raising the corresponding asymptotic level), speeds up recovery or slows down habituation, shortening or lengthening periods of non-dominance and dominance, respectively (Levelt I).

In the second instance, rates of habituation or recovery depend on characteristic times τ_e and $\tau_{e'}$ (Eq. 1 and 2). When these rates are unequal, dominance durations depend more sensitively on the *slower* process. This is why dominance durations depend more sensitively on non-dominant input (Levelt II): recovery of non-dominant evidence is generally slower than habituation of dominant evidence, independently of which input is weaker or stronger. The reason is that the respective effects of characteristic times τ_e and $\tau_{e'}$ and asymptotic levels e_∞ and e'_∞ are synergistic for weaker-input evidence (in both directions), whereas they are antagonistic for stronger-input evidence (supplementary text **Deterministic dynamics: Evidence pools** and **Fig. 11**).

In general, dominance durations depend hyperbolically on Δ_∞ (**Fig. 4d** and Eq. 7 in supplementary text). Dominance durations become infinite (and reversals cease) when Δ_∞ falls below the reversal threshold Δ_{rev} . This hyperbolic dependence is also why alternation rate peaks at equidominance (Levelt III): increasing the difference between inputs always lengthens longer durations more than it shortens shorter durations, thus lowering alternation rate.

Distribution of dominance durations

For all combinations of image contrast, the mechanism accurately predicts the experimentally observed distributions of dominance durations. This is owed to the stochastic activity of pools of bistable variables.

Firstly, dominance distributions retain nearly the same shape, even though average durations vary more than three-fold with image contrast (supplementary **Fig. 12ab**).

This ‘scaling property’ is due to the Poisson-like variability of birth-death processes (see supplementary text **Stochastic dynamics**). Generally, when a stochastic accumulation approaches threshold, the rates of both accumulation and dispersion of activity affect the distribution of first-passage-times [18, 19]. In the special case of Poisson-like variability, the two rates vary proportionally and preserve distribution shape (supplementary **Fig. 12cd**).

Secondly, predicted distributions correspond to Gamma distributions with scale factor $r \simeq 3-4$. As shown previously [18, 19], this is due to birth-death processes accumulating activity within a narrow range (*i.e.*, evidence difference $\Delta e \lesssim 0.2$). In this low-threshold regime, the first-passage-times of birth-death processes are both highly variable and Gamma distributed, consistent with experimental observations .

Thirdly, the predicted variability (coefficients of variation) of dominance periods varies along the $c+c'=1$ axis, being larger for longer than for shorter dominance durations (**Fig. 3ab**). The reason is that stochastic development becomes noise-dominated. For longer durations, stronger-input evidence habituates rapidly into a regime where random fluctuations gain importance (see also supplementary **Fig. 11ab**).

Levelt’s proposition IV

The model accurately predicts how dominance durations shorten with higher image contrast $c=c'$ (Levelt IV). Surprisingly, this reflects the dynamics of decision pools R and R' (**Fig. 5**).

Here again it is helpful to consider the deterministic limit of large pools ($N \rightarrow \infty$). In this limit, a dominant decision state $r' \simeq 1$ is destabilized when a *contradictory* evidence difference $\Delta e = e - e'$ exceeds a certain threshold value Δ_{rev} (**Fig. 5b** and supplementary text **Deterministic dynamics: Decision pools**). Due to the combined effect of excitatory and inhibitory feedforward projections, w_{exc} and w_{inh} (Eq. 4 and **Fig. 5a**), this average reversal threshold *decreases* with average evidence activity $(e+e')/2$. Simulations of the fully stochastic model ($N = 25$) confirm this analysis (**Fig. 5c**). As average evidence activity $(e+e')/2$ increases with image contrast, the average evidence bias Δe at the time of reversals decreases, resulting in shorter dominance periods (**Fig. 5d**).

Figure 5. Competitive dynamics of decision pools ensures Level IV. **(a)** The joint stable state of decision pools (here $r' \simeq 1$ and $r \simeq 0$) can be destabilized by sufficiently contradictory evidence, $e > e'$. **(b)** Effective potential $U(e, e', r, r')$ (colored curves) and steady-states r_{ss} (colored dots) for different levels of contradictory input, $\Delta e = e - e'$. Increasing Δe destabilizes the steady-state and shifts r_{ss} rightwards (curved arrow). The critical value r_{crit} (dotted vertical line), at which the steady-state turns unstable, is reached when Δe reaches the reversal threshold Δ_{rev} . At this point, a reversal ensues with $r \rightarrow 1$ and $r' \rightarrow 0$. **(c)** The reversal threshold Δ_{rev} diminishes with combined evidence $e + e'$. In the deterministic limit, Δ_{rev} decreases linearly with $\bar{s} = (s + s')/2$ (dashed red line). In the stochastic system, the average evidence bias $\langle \Delta e \rangle$ at the time of reversals decreases similarly with the average evidence mean $\bar{e}$ (black dots). Actual values of Δe at the time of reversals are distributed around these average values (grey shading). **(d)** Average evidence mean $\langle \bar{s} \rangle$ (left) and average evidence bias $\langle \Delta s \rangle$ (middle) at the time of reversals, as a function of image contrast c or c' . Decrease of average evidence bias $\langle \Delta s \rangle$ with contrast shortens dominance durations (Level IV). At low contrast (blue dot), higher reversal thresholds Δ_{rev} result in less frequent reversals (bottom right, grey and white regions) whereas, at high contrast (red dot), lower reversal thresholds lead to more frequent reversals (top right).

Serial dependence

The proposed mechanism accurately predicts positive correlations between successive dominance durations, a well-known characteristic of multistable phenomena [32, 39, 114, 119]. In addition, it predicts further aspects of serial dependence not described previously.

In both model and experimental observations, a long dominance periods tends to be followed by another periods, and a short dominance periods by another short period (**Fig. 6**). In the model, this is due to mean evidence activity $\bar{e}=(e + e')/2$ fluctuating stochastically above and below its long-term average. The autocorrelation time of these fluctuations increases monotonically with image contrast and, for high contrast, spans multiple dominance periods (supplementary text **Characteristic times** and **Fig. 13**). Note that fluctuations of $\bar{e}$ diminish as the number of bistable variables increases and vanish in the deterministic limit ($N \rightarrow \infty$).

Crucially, fluctuations of $\bar{e}$ modulate both reversal threshold Δ_{rev} and dominance durations T , as illustrated in **Fig. 6ab**. To obtain **Fig. 6a**, dominance durations were grouped into quantiles and the average duration $\langle T_0 \rangle$ of each quantile was compared to the conditional expectation of preceding and following durations $\langle T_{\pm n} \rangle$ (upper graph). For the same quantiles (compare color coding), average evidence activity $\bar{e}_0$ was compared to the conditional expectation $\langle \bar{e}_{\pm n} \rangle$ at the end of preceding and following periods (lower graph). Both the *inverse* relation between $\langle T_{\pm n} \rangle$ and $\langle \bar{e}_{\pm n} \rangle$ and the autocorrelation over multiple dominance periods are evident.

This source of serial dependency – comparatively slow fluctuations of $\bar{e}$ – predicts several qualitative characteristics not described previously and now confirmed by experimental observations. First, sequential correlations are predicted (and observed) to be strictly positive at all lags (next period, one-after-next period, and so on)(**Fig. 6d**). In other words, it predicts that several successive dominance periods are shorter (or longer) than average.

Second, due to the contrast-dependence of autocorrelation time, sequential correlations are predicted (and observed) to increase with image contrast (**Fig. 6d**). The experimentally observed degree of contrast-dependence is broadly consistent with pool sizes between $N=25$ and $N=40$ (black and red curves in **Fig. 2e**). Larger pools with hundreds

of bistable variables do not express the observed dependence on contrast (not shown).

Third, for high image contrast, reversal sequences are predicted (and observed) to contain extended episodes with dominance periods that are short, or extended episodes with periods that are long (**Fig. 6c**). When quantified in terms of a ‘burstiness index’, the degree of inhomogeneity in predicted and observed reversal sequences is comparable (supplementary text **Burstiness** and **Fig. 14**).

Many previous models of binocular rivalry [*e.g.*, 59] postulated selective adaptation of competing representations to account for serial dependency. However, selective adaptation is an *opponent* process which ensures *positive* correlations between *different* dominance periods, but *negative* correlations between *same* dominance periods. To demonstrate this point, we fitted such a model to reproduce our experimental observations (T , c_v , γ_1 , and cc_1) for five image contrasts $c = c'$. As expected, the alternative model predicts *negative* correlations cc_2 for *same* dominance periods (**Fig. 6d**, right panel), contrary to what is observed.

Figure 6. Serial dependency predicted by model and confirmed by experimental observations. **(a)** Conditional expectation of dominance duration $\langle T_{\pm n} \rangle$ (top) and of mean evidence activity, $\langle \bar{s}_{\pm n} \rangle$ (bottom), in model simulations with maximal stimulus contrast ($c=c'=1.0$). Dominance periods T_0 were grouped into octiles, from longest (yellow) to shortest (black). For each octile, the average duration $\langle T_{\pm n} \rangle$ of preceding and following dominance periods, as well as the average mean evidence activity $\langle \bar{s}_{\pm n} \rangle$ at the end of each period, are shown. All times in multiples of the overall average duration, $\langle T \rangle$. **(b)** Example reversal sequence from model. Bottom: Stochastic development of evidence activities e and e' (red and blue traces), with large, joint fluctuations raising or lowering mean activity $\bar{e}=(e+e')/2$ above or below long-term average (dashed line). Top left: episode with $\bar{s}$ above average, lower Δ_{rev} and shorter dominance periods. Top right: episode with $\bar{s}$ below average, higher Δ_{rev} and longer dominance durations. **(c)** Examples of reversal sequences from human observers ($c=c'=1.0$ and $c=c'=0.5$). **(d)** Positive lagged correlations predicted by model (mean, middle) and confirmed by experimental observations (mean $\pm$ std, top). Alternative model [59] with adaptation & noise (mean, bottom), fitted separately at each contrast to reproduce $\langle T \rangle$, c_v , γ_1 , and c_{C_1} correlations of the present model (blue stars).

DISCUSSION

We have shown that the universal statistical features of multistable phenomena in perception, as exemplified by binocular rivalry, are reproduced and indeed guaranteed by a particular biophysical mechanism. The explanatory power of this mechanism is extraordinarily high, in that the phenomena explained are considerably more complex (have far more degrees of freedom) than the mechanism itself. It has long been suspected that multistable phenomena in perception may share a similar mechanistic origin. Here we present for first time a possible common mechanism that is biophysically detailed and neurally plausible.

Continuous inference

The notion that multistable phenomena such as binocular rivalry reflect active exploration of explanatory hypotheses for sensory evidence has a venerable history [5, 44, 65, 118]. The mechanism proposed here is entirely in keeping with that notion: higher-level ‘explanations’ compete for control (‘dominance’) of phenomenal appearance in terms of their correspondence to lower-level ‘evidence’. An ‘explanation’ takes control if its correspondence is sufficiently superior to that of rival ‘explanations’. The greater the superiority, the longer control is retained. Eventually, alternative ‘explanations’ seize control, if only briefly. This manner of operation is also consistent with computational theories of ‘analysis by synthesis’ or ‘hierarchical inference’, although there are many differences in detail [86, 91, 97].

Interacting with an uncertain and volatile world necessitates continuous and concurrent evaluation of sensory evidence and selection of motor action [22, 43]. Multistable phenomena exemplify continuous decision-making without external prompting [17]. Sensory decision-making has been studied extensively, mostly in episodic choice-task, and the neural circuits and activity dynamics underlying episodic decision-making – including representations of potential choices, sensory evidence, and behavioural goals – have been traced in detail [22, 42, 57, 122]. Interestingly, there seems to be substantial overlap between choice representations in decision-making and in multistable situations [17].

Dynamical mechanism

Two principal *alternatives* have been considered for the dynamical mechanism of

perceptual decision-making: drift-diffusion models [69, 98] and recurrent network models [121, 122]. The mechanism proposed here *combines* both alternatives: at its evidence level, sensory information is integrated, over both space and time, by ‘local attractors’ in a discrete version of a drift-diffusion process. At its decision level, the population dynamics of a recurrent network implements a winner-take-all competition between ‘non-local attractors’. Together, the two levels form a ‘nested attractor’ system [17] operating perpetually out of equilibrium.

A recurrent network with strong competition typically ‘normalizes’ individual responses relative to the total response [77]. Divisive normalization is considered a canonical cortical computation [20], for which multiple rationales can be found. Here divisive normalization is augmented by indiscriminate feedforward inhibition. This combination ensures that decision activity rapidly and reliably categorizes *differential* input strength, largely independently of *total* input strength.

Another key feature of the proposed mechanism is that a ‘dominant’ decision pool applies feedback suppression to the associated evidence pool. Selective suppression of evidence for a winning hypothesis features in computational theories of ‘hierarchical inference’ [61, 86, 91, 97], as well as in accounts of multistable perception inspired by such theories [27, 46, 124]. A normative reason for feedback suppression arises during continuous inference in uncertain and volatile environments, where the accumulation of sensory information is ongoing and cannot be restricted to appropriate intervals [117]. Here optimal change detection requires an exponentially rising bias *against* evidence for the most likely state, ensuring that even weak changes are detected, albeit with some delay.

The pivotal feature of the proposed mechanism are pools of bistable variables or ‘local attractors’. Encoding sensory inputs in terms of persistent ‘activations’ of local attractors assemblies (rather than in terms of transient neuronal spikes) creates an intrinsically retentive representation: sites that respond are also sites that retain information (for a limited time). Our results are consistent with a few tens of bistable variables in each pool. In the proposed mechanism, *differential* activity of two pools accumulates evidence *against* the dominant appearance until a threshold is reached and a reversal ensues. This discrete non-equilibrium dynamics may well instantiate a variational principle of inference such as ‘maximum caliber’

[35, 94].

Emergent features

The components of the proposed mechanism interact to guarantee the statistical features that characterize binocular rivalry and other multistable phenomena. Discretely stochastic accumulation of differential evidence *against* the dominant appearance ensures sensitivity of dominance durations to non-dominant input. It also ensures the constant variability (‘scaling property’) and Gamma-like distribution shape of dominance durations. Due to a highly non-trivial interaction with the competitive decision, discretely stochastic fluctuations of evidence level activity express themselves in a serial dependency of dominance durations. These unexpected dependencies are experimentally confirmed and strongly corroborate the proposed mechanism.

Previous models

How does the proposed mechanism compare to previous ‘dynamical’ models of multistable phenomena? It is of similar complexity in that it assumes two dynamical levels, one slow (accumulation) and one fast (winner-take-all competition). It differs in reversing their ordering: visual input impinges first on the slow level, which then drives the fast level. It also differs in that stochasticity dominates the slow dynamics [as suggested by 115], not the fast dynamics. However, the most fundamental difference is discreteness (pools of bistable variables), which shapes all key dynamical properties.

Neural substrate

What might be the neural basis of the bistable variables / ‘local attractors’ proposed here? Ongoing activity in sensory cortex appears to be low-dimensional, in the sense that the activity of neurons with similar response properties varies concomitantly [“shared variability”, “noise correlations”, 37, 75, 76, 92, 99]. This shared variability reflects the spatial clustering of intracortical connectivity [25, 62, 81, 84, 101] and unfolds over moderately slow time-scales, with characteristic times in the range of 100 *ms* to 500 *ms*) reported in primate and rodent cortex [26, 37, 75, 76, 92, 99].

Possible dynamical origins of shared and moderately slow variability have been

studied extensively in theory and simulation [for reviews, see 47, 58, 77]. Networks with weakly clustered connectivity (*e.g.*, 3% rewiring) can express a metastable attractor dynamics with moderately long time-scales [36, 67, 101, 103]. In a metastable dynamics, individual (connectivity-defined) clusters transition spontaneously between distinct and quasi-stationary activity levels (‘attractor states’) [109, 113].

Evidence for metastable attractor dynamics in cortical activity is accumulating steadily [37, 73–76, 99]. Distinct activity states with exponentially distributed durations have been reported in sensory cortex [37, 75], consistent with noise-driven escape transitions [36, 47]. And several reports are consistent with external input modulating cortical activity mostly indirectly, *via* the rate of state transitions [21, 37, 38, 75, 76].

The proposed mechanism assumes bistable variables with noise-driven escape transitions, with transition rates modulated exponentially by external synaptic drive. Following previous work [19, 36, 47], we show this to be an accurate reduction of the population dynamics of metastable networks of spiking neurons.

As multistable phenomena and their characteristics are ubiquitous in visual, auditory and tactile perception, the mechanism we propose may form a general part of sensory processing. It bridges neural, perceptual and normative levels of description and potentially offers a ‘comprehensive task-performing model’ [56] for sensory decision-making.

METHODS

Psychophysics

Six practised observers participated in the experiment (4 male, 2 female). Stimuli were displayed on an LCD screen (EIZO ColorEdge CG303W, resolution 2560×1600 pixels, viewing distance was 104 cm, single pixel subtended 0.014° , refresh rate 60Hz) and were viewed through a mirror stereoscope, with viewing position being stabilized by chin and head rests. Display luminance was gamma-corrected and average luminance was 50 cd/m^2 .

Two grayscale circular orthogonally-oriented gratings ($+45^\circ$ and -45°) were presented foveally to each eye. Gratings had diameter of 1.6° , spatial period 2 cyc/deg . To avoid a sharp edge, grating contrast was modulated with Gaussian envelope (starting inner radius 0.6° , $\sigma = 0.2^\circ$). Tilt and phase of gratings was randomized for each block. Five contrast levels were used: 6.25%, 12.5%, 25%, 50%, and 100%. Contrast of each grating was systematically manipulated, so that each contrast pair was presented in two blocks (50 blocks in total). Blocks were 120 s long and separated by a compulsory one-minute break. Observers reported on the tilt of the visible grating by continuously pressing one of two arrow keys. They were instructed to press only during exclusive visibility of one of the gratings, so that mixed percepts were indicated by neither key being pressed (25% of total presentation time). To facilitate binocular fusion, gratings were surrounded by a dichoptically presented square frame (outer size 9.8° , inner size 2.8°).

Dominance periods of ‘clear visibility’ were extracted in sequence from the final 90 s of each block and the mean linear trend was subtracted from all values. Values from the initial 30 s were discarded. To make comparable the dominance periods of different observers, values were rescaled by the ratio of the all-condition-all-observer average (2.5 s) and the all-condition average of each observer ($2.5 \pm 1.3 \text{ s}$). Finally, dominance periods from symmetric conditions ($c_{\text{left}}, c_{\text{right}}$) were combined into a single category ($c_{\text{dom}}, c_{\text{sup}}$), where c_{dom} (c_{sup}) was the contrast viewed by the dominant (suppressed) eye.

For the dominance periods T observed in each condition, first, second, and third central moments were computed, as well coefficient of variation c_v and skewness γ_1 relative

to coefficient of variation:

$$\mu_1 = \langle T \rangle, \quad \mu_2 = \langle T^2 \rangle - \langle T \rangle^2, \quad \mu_3 = \langle T^3 \rangle - 3\langle T \rangle \langle T^2 \rangle + 2\langle T \rangle^3$$

$$c_v = \frac{\sqrt{\mu_2}}{\mu_1}, \quad \frac{\gamma_1}{c_v} = \frac{\mu_3 \mu_1}{\mu_2^2}$$

Coefficients of sequential correlations were computed from pairs of periods (T_i, T_j) with opposite dominance (first and next: ‘lag’ $j - i = 1$), pairs of periods with same dominance (first and next but one: ‘lag’ $j - i = 2$), and so on,

$$c_k = \frac{\langle T_i - \langle T_i \rangle \rangle \langle T_j - \langle T_j \rangle \rangle}{\sqrt{(\langle T_i^2 \rangle - \langle T_i \rangle^2) (\langle T_j^2 \rangle - \langle T_j \rangle^2)}}$$

where $\langle T \rangle$ and $\langle T^2 \rangle$ are mean duration and mean square duration, respectively.

To analyze ‘burstiness’, we adapted a statistical measure used in neurophysiology [24]. First, sequences of dominance periods were divided into all possible subsets of $k \in \{2, 3, \dots, 16\}$ successive periods and mean durations computed for each subset. Second, heterogeneity was assessed by computing, for each size k , the coefficient of variation c_v over mean durations, compared to the mean and variance of the corresponding coefficient of variation for randomly shuffled sequences of dominance periods. Specifically, a ‘burstiness index’ was defined for each subset size k as

$$BI(k) = \frac{c_v - \langle c_v \rangle_{shuffled}}{\sqrt{\langle c_v^2 \rangle_{shuffled} - \langle c_v \rangle_{shuffled}^2}}$$

where c_v was the coefficient of variation over subsets of size k and where $\langle c_v \rangle_{shuffled}$ and $\langle c_v^2 \rangle_{shuffled}$ were, respectively, mean and mean square of the coefficients of variation from shuffled sequences.

Model

The proposed mechanism for binocular rivalry dynamics relies on discretely stochastic processes (‘birth-death’ or generalized Ehrenfest processes). Bistable variables $x \in \{0, 1\}$ transition between active and inactive states with time-varying Poisson rates $\nu^+(t)$ (activation) and $\nu^-(t)$ (inactivation). Two ‘evidence pools’ of N such variables, E and E' ,

represent two kinds visual evidence (*e.g.*, for two visual orientations), whereas two ‘decision pools’, R and R' , represent alternative perceptual hypotheses (*e.g.*, two grating patterns) (**Fig. 10**). Thus, instantaneous dynamical state is represented by four active counts $n_e, n_{e'}, n_r, n_{r'} \in [0, N]$ or, equivalently, by four active fractions $e, e', r, r' \in [0, 1]$.

The development of pool activity over time is described by a master equation for probability $P_n(t)$ of the number $n(t) \in [0, N]$ active variables.

$$\begin{aligned} \partial_t P_n(t) = & (N - n + 1) \nu^+ P_{n-1}(t) + (n + 1) \nu^- P_{n+1}(t) \\ & - [(N - n) \nu^+ + n \nu^-] P_n(t) \end{aligned} \quad (5)$$

For constant $\nu^\pm$, the distribution $P_n(t)$ is binomial at all times [49, 116]. The time-development of the number of active units $n_X(t)$ in pool X is an inhomogeneous Ehrenfest process and corresponds to the count of activations, minus the count of deactivations,

$$\Delta n_X(t) = \underbrace{\mathcal{B}(N - n_X, \nu^+ \Delta t)}_{\text{activations}} - \underbrace{\mathcal{B}(n_X, \nu^- \Delta t)}_{\text{inactivations}}$$

where $\mathcal{B}(n, \nu \Delta t)$ is a discrete random variable drawn from a binomial distribution with trial number n and success probability $\nu \Delta t$.

All variables of a pool have identical transition rates, which depend exponentially on the ‘potential difference’ $\Delta u = u + u^0$ between states, with a input-dependent component u and a baseline component u_0 :

$$\begin{aligned} \nu_s^\pm &= \frac{\nu_s}{2} e^{\pm(u_e + u_0^e)/2}, & \nu_{s'}^\pm &= \frac{\nu_s}{2} e^{\pm(u_{e'} + u_0^e)/2} \\ \nu_r^\pm &= \frac{\nu_r}{2} e^{\pm(u_r + u_0^r)/2}, & \nu_{r'}^\pm &= \frac{\nu_r}{2} e^{\pm(u_{r'} + u_0^r)/2} \end{aligned}$$

where ν_e and ν_r are baseline rates and u_0^e and u_0^r baseline components. The input-dependent components of effective potentials are modulated linearly by synaptic couplings

$$\begin{aligned} u_s &= w_{vis} f(c) - w_{supp} r \\ u_{s'} &= w_{vis} f(c') - w_{supp} r' \\ u_r &= w_{exc} e - w_{inh} (e + e') + w_{coop} r - w_{comp} r' \\ u_{r'} &= w_{exc} e - w_{inh} (e + e') + w_{coop} r' - w_{comp} r \end{aligned}$$

Visual inputs are $I = f(c)$ and $I' = f(c')$, respectively, where $f(c) = \frac{\ln c + \gamma/\gamma}{\ln 1 + \gamma/\gamma} \in \{0, 1\}$ is a monotonically increasing, logarithmic function of image contrast $c \in [0, 1]$, with parameter γ .

Degrees of freedom

The proposed mechanism has twelve (12) independent parameters: pool size, 6 synaptic couplings, 2 baseline rates, 2 baseline potentials, 1 contrast non-linearity.

Symbol	description	value
N	pool size	25
ν_e	baseline rate, evidence	1.949
ν_r	baseline rate, decision	0.01767
u_0^e	baseline potential, evidence	-1.6529
u_0^r	baseline potential, decision	-4.928
w_{vis}	visual input coupling	1.7798
w_{exc}	feedforward excitation	152.2
w_{inh}	feedforward inhibition	32.10
w_{comp}	lateral competition	33.38
w_{coop}	lateral cooperation	15.21
w_{supp}	feedback suppression	2.340
γ	contrast nonlinearity	0.07088

Fitting procedure

The experimental dataset consisted of three 5×5 arrays X_i^{exp} for mean $\langle T \rangle$, coefficient of variation c_v , skewness γ_1 , plus a scalar correlation coefficient $cc_1=0.06$, which was the average cc_1 observed over the 5×5 array. In other words, the fitting procedure prescribed contrast-dependencies for distribution moments, but not for correlation coefficients.

The fit error E_{fit} was computed as a weighted sum of relative errors

$$E_{fit} = \sum_{i=1}^4 w_i \delta_i \bigg/ \sum_{i=1}^4 w_i, \quad \delta_i = \left| \frac{X_i^{mod} - X_i^{exp}}{\bar{X}_i^{exp}} \right|$$

with weighting $w = [1, 1, 1, 1/4]$ emphasizing distribution moments.

Approximately 400 minimization runs were performed, starting from random initial configurations of model parameters. For the optimal parameter set, the resulting fit error for the mean observer dataset was approximately 15%.

To confirm that resulting fit was indeed optimal, and could not be further improved, we studied the behaviour of the fit error in the vicinity of the optimal parameter set. For each parameter α_i , 10 values $\alpha_i^{(j)}$ were picked in the direct vicinity of the optimal parameter α_i^{opt} . The resulting scatter plot of value pairs $\alpha_i^{(j)}$ and fit error $E_{fit}^{(j)}$ is approximated by a quadratic function. For most parameters the estimated quadratic function is convex, so that the corresponding coefficient of the Hessian matrix associated with the fit error is positive. Also, the estimated extremum of each parabola was close to the corresponding optimal parameter, indicating that the parameter set was indeed optimal.

Alternative model

As an alternative model [59], a combination of competition, adaptation, and noise was fitted to reproduce $\langle T \rangle$, c_v , γ_1 , and c_1 for a given image contrast $c = c'$. The model comprises four state variables and independent coloured noise,

$$\begin{aligned} \tau_r \dot{r}_{1,2} &= -r_{1,2} + F(-\beta r_{2,1} - \phi_a a_{1,2} + I_{1,2} + n_{1,2}), \\ \tau_a \dot{a}_{1,2} &= -a_{1,2} + r_{1,2} \\ \tau_n \dot{n}_{1,2} &= -n_{1,2} + \sigma_n \sqrt{2\tau_n} \xi(t) \end{aligned}$$

where $F(x) = [1 + \exp(-x/k)]^{-1}$ is a non-linear activation function and $\xi(t)$ is white noise. Parameters $\tau_r = 10 \text{ ms}$, $\tau_n = 100 \text{ ms}$, and $k = 0.1$ were fixed. Parameters for competition β , adaptation strength ϕ_a , adaptation time-constant τ_a , and noise amplitude σ_n were fitted separately for each image contrast.

Spiking network simulation

To illustrate a possible neural realization of ‘local attractors’, we simulated a competitive network with eight (8) identical assemblies of excitatory and inhibitory neurons, which collectively expresses a spontaneous and metastable dynamics [74]. One assembly (denoted as ‘foreground’) comprised 125 excitatory leaky-integrate-and-fire neurons, which were weakly coupled to the 875 excitatory neurons of the other assemblies (denoted as ‘background’), as well as 250 inhibitory neurons. Note that background assemblies are not strictly necessary and included only for the sake of verisimilitude. The connection probability between any two neurons was $c = 80\%$. Excitatory synaptic efficacy between ‘foreground’ neurons, ‘background’ neurons’, and between the two was $J_{\text{fore}} = 0.618mV$, $J_{\text{back}} = 0.438mV$, and $J_{\text{inter}} = 0.402mV$, respectively. Inhibitory synaptic efficacy was $J_I = -1.50mV$ and the efficacy of excitatory synapses onto inhibitory neurons was $J_{IE} = 0.560mV$. Finally, ‘foreground’ neurons, ‘background neurons’, and ‘inhibitory neurons’ each received independent Poisson spike trains of $2340Hz$, $2280Hz$, and $2280Hz$, respectively. Other settings were as in [74]. As a result of these settings, ‘foreground’ activity transitioned spontaneously between an ‘off’ state of approximately $3Hz$ and an ‘on’ state of approximately $40Hz$.

ACKNOWLEDGEMENTS

Funding from EU FP7-269459 Coronet, DFG BR 987/3-1 and DFG 987/4-1. The authors are obliged to Andrew Parker for helpful comments.

SUPPLEMENTARY MATERIAL

Figure 7. Proposed mechanism of binocular rivalry dynamics (schematic). Bistable variables are represented by white (inactive) or red (active) circles. Four pools, each with $N = 25$ variables, are shown: two evidence pools E and E' , with active counts $n_e(t)$ and $n_{e'}(t)$, and two decision pools, R and R' , with active counts $n_r(t)$ and $n_{r'}(t)$. Excitatory and inhibitory synaptic couplings include selective feedforward excitation w_{exc} , indiscriminate feedforward inhibition w_{inh} , recurrent excitation w_{coop} and mutual inhibition w_{comp} of decision pools, as well as selective feedback suppression w_{supp} of evidence pools. Visual input to evidence pools $f(c)$ and $f(c')$ is a function of image contrast c and c' .

Metastable attractor dynamics

We postulate assemblies or clusters of neurons with recurrent random connectivity as operative units of sensory representations. In our model, such assemblies are reduced to binary variables with Poisson transitions. Our key assumption is that the rates $\nu^\pm$ of activation and inactivation events are modulated exponentially by synaptic input (Eq. 1):

$$\nu^\pm = \nu e^{\pm(ws+u_0)}$$

Here we show that these assumptions are a plausible reduction of recurrently connected assemblies of spiking neurons.

Following earlier work, we simulated a competitive network with eight (8) identical assemblies of excitatory and inhibitory neurons (**Fig. 8a**), configured to collectively express a metastable activity dynamics [74]. Here we are interested particularly in the activity dynamics of one excitatory assembly (dubbed ‘foreground’), which expresses two quasi-stable ‘attractor’ states: an ‘on’ state with high activity. In the context of the metastable network, the ‘foreground’ assembly is bistable in that it transitions spontaneously between ‘on’ and ‘off’ states. Such state transitions are noise-driven escape events from an energy well and therefore occur with Poisson-like rates ν^+ (activation) and ν^- (inactivation). **Fig. 1b** and **Fig. 8b** illustrate this energy landscape for the ‘diffusion limit’ a very large assemblies, where quasi-stable activity levels are $\nu_{fore} \simeq 45 \text{ Hz}$ for the ‘on’ state state and $\nu_{fore} \simeq 4 \text{ Hz}$ for the ‘off’ state. For small assemblies with fewer neurons, the difference between ‘on’ and ‘off’ states is less pronounced.

To establish the dependence of transition rates on external input to the ‘foreground’ assembly, we stepped external input rate $\Delta\nu_{ext}$ between two values selected from a range $\Delta\nu_{ext} \in [-120 \text{ Hz}, 50 \text{ Hz}]$ and monitored the resulting spiking activity in individual neurons, as well as activity ν_{fore} of the entire population (**Fig. 8c**, upper and middle panels). Comparing population activity to a suitable threshold, we identified ‘on’ and ‘off’ states of the ‘foreground’ assembly (**Fig. 8c**, lower panel), as well as the probability of ‘on’ or ‘off’ states at different points in time following a step in $\Delta\nu_{ext}$ (**Fig. 8d**). From the hazard rate (temporal derivative of probability) we then estimated the rates $\nu^\pm$ of state transitions shown in **Fig. 8d**. Transition rates $\nu^\pm$ vary anti-symmetrically and exponentially

Figure 8. Metastable dynamics of spiking neural network. **(a)** Eight assemblies of excitatory neurons (schematic, light and dark gray disks) and one pool of inhibitory neurons (white disc) interact competitively with recurrent random connectivity. We focus on one ‘foreground’ assembly (dark gray), with firing rate ν_{fore} and selective external input $\Delta\nu_{ext}$. **(b)** ‘Foreground’ activity explores an effective energy landscape with two distinct steady-states (circles), separated by ridge points (diamonds). As this landscape changes with external input $\Delta\nu_{ext}$, transition rates $\nu_{\pm}$ between ‘on’ and ‘off’ states also change with external input. **(c)** Simulation to establish transition rates $\nu_{\pm}$ of foreground assembly. External input $\Delta\nu_{ext}$ is stepped periodically between -44 Hz and -4 Hz. Spiking activity of 10 representative excitatory neurons in a single trial, population activity over 25 trials, thresholded population activity over 25 trials, and activation probability (fraction of ‘on’ states). **(d)** Relaxation dynamics in response to step change of $\Delta\nu_{ext}$, with ‘on’ transitions (left) and ‘off’ transitions (right). **(e)** Average state transition rates $\nu^{\pm}$ vary exponentially and symmetrically with external input. The dependence is well approximated by $\nu^{\pm} \simeq 1\text{Hz} \exp(\pm 0.04s\Delta\nu_{ext} + 1)$ (red and blue lines).

with external input $\Delta\nu_{ext}$. In the present example, the dependence is well approximated by $\nu^{\pm} \simeq 1Hz \exp(\pm 0.04s\Delta\nu_{ext} + 1)$ (**Fig. 8e**, red and blue lines). This Arrhenius-Van't-Hoff-like dependence of escape rates is a consequence of the approximately linear dependence of relative potential on external input. Escape kinetics is typical for attractor systems and motivates Eq. 1.

Quality of representation

Accumulation of information

A birth-death process – defined as N bistable variables with transition rates $\nu_{\pm} = \nu e^{\pm ws}$, where ν is a baseline rate and w a coupling constant – accumulates and retains information about input s , performing as a ‘leaky integrator’ with a characteristic time scale [17]. Specifically, the value of s may be inferred from fractional activity $x(t)$ at time t , if coupling w and baseline rate ν are known. The inverse variance of the maximum likelihood estimate is given by the Fisher information

$$J_x(s, t) = \frac{N [\partial_s \langle x \rangle]^2}{\langle x \rangle (1 - \langle x \rangle)} \quad (6)$$

Its value grows with time, approaching $J_x = Nw^2 / \cosh^2(ws/2)$ for $t \rightarrow \infty$. For small inputs $s \simeq 0$, the Fisher information increases monotonically as $J_x(t) \approx Nw^2/4 \tanh(\nu t/2)$. Surprisingly, the upper bound by $J_x \leq Nw^2/4$ depends linearly on pool size N , but quadratically on coupling w . Thus, stronger coupling substantially improves encoding accuracy (of input s).

The rate at which Fisher information is accumulated by a pool is set by the baseline transition rate ν . An initially inactive pool, with $n_0=0$, accumulates Fisher information at an initial rate of $\partial_t J_x|_{t=0} = \nu Nw^2/4 e^{ws/2}$. Thus, any desired rate of gaining Fisher information may be obtained by choosing an appropriate value for ν . However, unavoidably, after an input s has ceased (and was replaced by another), information about s is lost at the same rate.

Integration of noisy samples

Figure 9. Supplementary. Information retained by stochastic pool activity from normally distributed inputs. Inputs $s \in \mathcal{N}(\mu, \sigma)$ provide Fisher information $J_s = 1/\sigma^2$ about mean μ . Stochastic activity $n(t)$ of a birth-death process ($N \in \{10, 20, 40, 80\}$ and $w = 2.5$) driven by such inputs accumulates Fisher information $J_n(t)$ about mean μ . **(a)** Accumulation over input interval $t = [0, 1]$ of fractional information $J_{rel}(t) = J_n(t) \sigma^2$ by an initially inactive pool of size N . **(b)** Information about μ retained by *summed activity* $\hat{n} = n_1 + \dots + n_4$ of four independent pools (all initially inactive and of size N) receiving concurrently four independent inputs ($s \in \mathcal{N}(\mu, \sigma)$) over an interval $t = [0, 1]$. Retained fraction $J_{rel} = J_{\hat{n}}(1) \sigma^2/4$ depends on pool size N and input variance σ^2 . **(c)** Information about μ retained by activity n of one pool (initially inactive and of size N) receiving successively four independent inputs ($s \in \mathcal{N}(\mu, \sigma)$) over an interval $t = [0, 43]$. Retained fraction $J_{rel} = J_n(4) \sigma^2/4$ depends on pool size N and input variance σ^2 .

Birth-death processes are able to encode also noisy sensory inputs, capturing much of the information provided. When an initially inactive pool receives an input s over time t , stochastic activity $n(t)$ gradually accumulates information about the value of s . Normally distributed inputs $s \in \mathcal{N}(\mu, \sigma)$ provide Fisher information $J_s = 1/\sigma^2$ about mean μ . Pool activity $n(t)$ accumulates Fisher information $J_n(t)$ about input mean μ , which may be compared to J_s . Comparatively small pools with strong coupling (*e.g.*, $N=25$, $w=2.5$) readily capture 90% of the information provided (**Fig. 9a**).

Moreover, pools readily permit information from multiple independent inputs to be combined over space and/or time. For example, the combined activity of four pools ($N=25$, $w=2.5$), which receive concurrently four independent samples, captures approximately 80% of the information provided, and a single pool receiving four samples in succession still retains approximately 60% of the information provided (**Fig. 9bc**). In the latter case, retention is compromised by the ‘leaky’ nature of stochastic integration. Whether signals are being integrated over space or time, the retained fraction of information is highest for inputs of

moderate and larger variance σ^2 (**Fig. 9bc**). This is because inputs with smaller variance are degraded more severely by the internal noise of a birth-death process (*i.e.*, stochastic activations and inactivations).

Suitability for inference

Summation of heterogeneous neural responses can be equivalent to Bayesian integration of sensory information [6, 93]. In general, this is the case when response variability is “Poisson-like” and response tuning differs only multiplicatively [70, 71]. We now show that bistable stochastic variables $x_i(t)$, with heterogeneous transition rates $\nu_i^\pm(s)$, satisfy these conditions, as long as synaptic coupling w is uniform.

Assuming initially inactive variables, $x_i(0) = 0$, incremental responses $x_i(\Delta t)$ after a short interval Δt are binomially distributed about mean $\langle x_i(\Delta t) \rangle$, which is approximately

$$\langle x_i(\Delta t) \rangle \approx \left. \frac{d\langle x_i \rangle}{dt} \right|_{t=0} \Delta t = \underbrace{\nu_i \Delta t/2}_{\phi_i} \underbrace{e^{ws/2}}_{f(s)}$$

where $\phi_i = \nu_i \Delta t/2$ reflects (possibly heterogeneous) response tuning and $f(s) = e^{ws/2}$ represents a common response function which depends only on synaptic coupling w . The Fisher information, about s , of individual responses is

$$J_i(s) = \frac{[\partial_s \langle x_i \rangle]^2}{\langle x_i \rangle [1 - \langle x_i \rangle]} \approx \phi_i \frac{f'^2(s)}{f(s)}, \quad \langle x_i \rangle \ll 1$$

as long as expected activation $\langle x_i \rangle$ is small. The Fisher information of summed responses $\sum_i x_i$ is

$$J_{sum}(s) \approx \frac{[f'(s) \sum_i \phi_i]^2}{f(s) \sum_i \phi_i} = \frac{f'^2(s)}{f(s)} \sum_i \phi_i = \sum_i J_i(s)$$

and equals the combined Fisher information of individual responses. Accordingly, the summation of bistable activities with heterogeneous transition rates ν_i optimally integrates information, provided expected activations remain small, $\langle x_i \rangle \ll 1$, and synaptic coupling w is uniform.

Figure 10. Decision response to fixed input I, I' , for random initial conditions of e, e', r, r' .
(a) Expected differential steady-state activation $\langle |r - r'| \rangle$ of decision level. Steady-state activity $r + r' \simeq 1$ implies a categorical decision with activity $\simeq 1$ of one pool and activity $\simeq 0$ of another.
(b) Probability that decision correctly reflects input bias ($r > r'$ if $I > I'$, and *vice versa*).

Categorical choice

The ‘biased competition’ circuit proposed here expresses a categorical decision by either raising r towards unity (and lowering r' towards zero) or *vice versa*. Here we describe its stochastic steady-state response to constant visual inputs $I=f(c)$ and $I'=f(c')$ and for arbitrary initial conditions of e, e', r and r' (**Fig. 10**). Note that, for purposes of this analysis, evidence activity e, e' was *not* subject to feedback suppression.

The choice is random when the input is ambiguous, $I \simeq I'$, but quickly becomes deterministic with growing input bias $|I - I'| > 0$. Importantly, the choice is consistently determined by visual input for all initial conditions. The 75% performance level is reached for biases $|I - I'| \approx 0.04$ to 0.06.

Mutual inhibition w_{comp} controls the width of the ambiguous region around $I = I'$, and self-excitation w_{coop} ensures a categorical decision even for small $I, I' \simeq 0$. The balance between feedforward excitation w_{exc} and inhibition w_{inh} eliminates decision failures for all but the largest values of $I, I' > 0.7$ and reduces the dependence of sensitivity to differential input $|I - I'|$ on total input $I + I'$.

For particularly high values of input I , $I' > 0.7$, no categorical decision is reached and activities of both r and r' grow above 0.5. In the full model, such inconclusive outcomes are eliminated by feedback suppression.

Deterministic dynamics

In the deterministic limit of $N \rightarrow \infty$, fractional pool activity x equals its expectation $\langle x \rangle$ and the relaxation dynamics of Eq. 2 becomes

$$\tau_x \frac{dx}{dt} = -x + x_\infty$$

with characteristic time $\tau_x = \frac{1}{\nu^+ + \nu^-} = \Upsilon(\Delta u)$ and asymptotic values $x_\infty = \frac{\nu^+}{\nu^+ + \nu^-} = \Phi(\Delta u)$, where Δu is the potential difference. Input-dependencies of characteristic time and of asymptotic value follow from Eq. 1:

$$\Upsilon(s) = \frac{1}{\nu} \text{sech} \frac{\Delta u}{2}, \quad \Phi(s) = [1 + e^{-\Delta u}]^{-1}$$

Evidence pools

The relaxation dynamics of evidence pools is given by Eq. 2 and Eq. 3'. As shown in the next section, reversals occur when evidence difference $|e - e'|$ reaches a reversal threshold Δ_{rev} . For example, a dominance period of evidence e' begins with $e'_{start} = e_{start} + \Delta_{rev}$ and ends when the concurrent habituation of e' and recovery of e have inverted the situation to $e_{end} = e'_{end} + \Delta_{rev}$ (**Fig. 11a**). Once the deterministic limit has settled into a limit cycle, all dominance periods start from, and end at, the same evidence levels.

If pool R' has just become dominant, so that $r' \approx 1$ and $r \approx 0$, the state-dependent potential differences are

$$u_e \simeq w_{vis} f(c)$$

$$u_{e'} \simeq w_{vis} f(c') - w_{coop}$$

and the deterministic development is

$$\tau_e \frac{de}{dt} = -e + e_\infty, \quad \tau_{e'} \frac{de'}{dt} = -e' + e'_\infty$$

with asymptotic values

$$e_{\infty} = \Phi(u_e + u_e^0), \quad e'_{\infty} = \Phi(u_{e'} + u_e^0)$$

and characteristic times

$$\tau_e = \Upsilon(u_e + u_e^0), \quad \tau_{e'} = \Upsilon(u_{e'} + u_e^0)$$

The starting points of the development, e_{rev} and e'_{rev} (dashed lines in **Fig. 11ab**), depend mostly on total input $c+c'$ and only little on input difference $c-c'$. Accordingly, for a given level of total input $c+c'$, the situation is governed by the distance between asymptotic evidence levels $\Delta_{\infty}=e_{\infty}-e'_{\infty}$ and by characteristic times $\tau_e, \tau_{e'}$.

The dependence on input bias $c-c'$ of effective potential Δu_e+u_e , characteristic time τ_e , and asymptotic value e_{∞} is illustrated in **Fig. 11ced**. The potential range of relaxation is $e_{rev} \rightarrow e_{\infty}$ and $e'_{rev} \rightarrow e_{\infty}$, where reversal levels e_{rev} and e'_{rev} can be obtained numerically.

Dominance durations depend more sensitively on the slower of the two concurrent processes, as it sets the pace of the combined development. The initial rates ρ_e and $\rho_{e'}$ after a reversal of the two opponent relaxations

$$\rho_e = \frac{d|e|}{dt} = \frac{|e_{rev} - e_{\infty}|}{\tau_e}, \quad \rho_{e'} = \frac{d|e'|}{dt} = \frac{|e'_{rev} - e'_{\infty}|}{\tau_{e'}}$$

provide a convenient proxy for relative rate. As shown in **Fig. 11f**, when stronger-input evidence e dominates, recovery of weaker-input evidence (red up arrow on blue background) is slower than habituation of stronger-input evidence (blue down arrow on blue background). Conversely, when weaker-input evidence e' dominates, recovery of stronger-input evidence (blue up arrow on red background) is slower than habituation of weaker-input evidence (red down arrow on red background). *In short, dominance durations always depend more sensitively on recovery of the currently non-dominant evidence.*

If the two evidence populations E, E' have equal and opposite potential differences, $\Delta u_e = -\Delta u_{e'}$, then they also have equal and opposite activation and inactivation rates (Eq. 1)

$$\nu_e^+ = \nu_{e'}^- = \nu^+, \quad \nu_e^- = \nu_{e'}^+ = \nu^-$$

and identical characteristic times τ_e (recovery of E) and $\tau_{e'}$ (habituation of E'). In this special case, the two processes may be combined and the development of evidence difference

Figure 11. Exponential habituation and recovery of evidence activities. Dominance durations depend on distance between asymptotic values and on characteristic times. **(ab)** Development of evidence e (blue) and e' (red), over two successive dominance periods. Input $c=0.9375$ is stronger, input $c' = 0.0625$ weaker. Activities recover, or habituate, exponentially until reversal threshold Δ_{rev} is reached. Thin curves extrapolate to the respective asymptotic values, e_∞ and e'_∞ . **(a)** Evidence e' (with weaker input c') is dominant. Incrementing input c to non-dominant evidence e shortens dominance T' . **(b)** Evidence e (with stronger input c) is dominant. Incrementing input c to e extends dominance T . **(c-f)** Contrast-dependence of relaxation dynamics, as a function of differential contrast $c-c'$, for $c+c'=1$. Values when evidence e is dominant (dom , thick solid curves), and when it is non-dominant (sup , thick dotted curves). Values for e' are mirror symmetric (about vertical midline $c-c'=0$). **(c)** Effective potential $\Delta u_e + u_e$. **(d)** Characteristic time τ_e . **(e)** Relaxation range $e_{rev} \rightarrow e_\infty$ (bottom left, thin curves e_{rev} , thick curves e_∞). **(f)** Effective rate ρ_e of development. Symbols and arrows correspond to subfigures **(ab)** and represent recovery (up arrow) or habituation (down arrow) of stronger-input evidence (blue) or weaker-input evidence (red). Underlying color patches indicate dominance of stronger-input evidence (blue patches) or of weaker-input evidence (red patches). Dominance durations depend more sensitively on the slower development, with smaller ρ , which generally is the recovery of non-dominant evidence (up arrows).

$\Delta e = e - e'$ is

$$\tau_{\Delta} \frac{d\Delta e(t)}{dt} = -\Delta e(t) + \Delta_{\infty}$$

$$\tau_{\Delta} = \frac{1}{\nu^{+} + \nu^{-}}, \quad \Delta_{\infty} = \frac{\nu^{+} - \nu^{-}}{\nu^{+} + \nu^{-}}$$

Starting from $\Delta e(0) = -\Delta_{rev}$, we consider the first-passage-time of $\Delta e(t)$ through $+\Delta_{rev}$. If the crossing is certain (*i.e.* when $\frac{\nu^{+}}{\nu^{+} + \nu^{-}} > \Delta_{rev}$), the first-passage-time T writes

$$T = \tau_{\Delta} \ln \left(\frac{\Delta_{\infty} + \Delta_{rev}}{\Delta_{\infty} - \Delta_{rev}} \right) \quad (7)$$

A similar hyperbolic dependence obtains also in all other cases. When the distance between asymptotic levels Δ_{∞} falls below the reversal threshold Δ_{rev} , dominance durations become infinite and reversals cease.

The hyperbolic dependence of dominance durations, illustrated in **Fig. 11d**, has an interesting implication. Consider the point of equidominance, at which both dominance durations are equal and of moderate duration. Increasing the difference between image contrasts (*e.g.*, increasing $c \rightarrow c + \Delta c$ and decreasing $c' \rightarrow c' - \Delta c$ increases Δ_{∞} during the dominance of e and decreases it during the dominance of e' . Due to the hyperbolic dependence, longer dominance periods lengthen more ($T \rightarrow T + \Delta T$) than shorter dominance periods shorten ($T' \rightarrow T' - \Delta T'$), consistent with the contemporary formulation of Levelt III [15].

Decision pools

We wish to analyze steady-state conditions for decision pools R, R' , as illustrated in **Fig. 5ab**. From Eq. 4, we can write

$$r_{\infty} = \phi(e, e', r_{\infty}, r'), \quad r'_{\infty} = \phi(e, e', r, r'_{\infty}),$$

Under certain conditions – in particular, for sufficient self-coupling w_{coop} – the steady-state equations admit more than one solution: a low-activity fixed point with $r_{\infty} \simeq 0$, and a high-activity fixed point with $r_{\infty} \simeq 1$. Importantly, the low-activity fixed point can be destabilized as the activation of the different population changes.

Consider a non-dominant decision pool R with fractional activity $r = n_r/N \approx 0$ and its dominant rival pool R' with fractional activity $r' = n_{r'}/N \approx 1$. The steady-state condition then becomes

$$\begin{aligned} r_\infty &\simeq \phi[w_{coop}(r_\infty - x_{eff})] \\ x_{eff} &= \frac{w_{comp} - w_{exc}e + w_{inh}(e + e') - u_r}{w_{coop}} \end{aligned}$$

For certain values $x_{eff} \leq x_{crit}$, the low-activity fixed point becomes unstable, causing a sudden upward activation of pool R and eventually a perceptual reversal. We call the steady-state r_∞ at the point of disappearance r_{crit} .

We can now define a threshold Δ_{rev} in terms of the value of evidence bias $\Delta e = e - e'$ which ensures that $x_{eff} \geq x_{crit}$:

$$\begin{aligned} \Delta_{rev} &\geq \frac{2}{w_{exc}} (w_{comp} - x_{crit}w_{coop} - u_r) - \\ &\quad - \frac{2}{w_{exc}} (w_{exc} - 2w_{inh}) \frac{e + e'}{2} \end{aligned}$$

We find that the threshold value Δ_{rev} decreases linearly with average evidence $\bar{e} = \frac{e+e'}{2}$, so that higher evidence activity necessarily entails lower thresholds (dashed red line in **Fig. 5c**).

For $w_{coop} = 15.21$, we find $x_{crit} = 0.24006$, $r_{crit} = 0.0708$, and $\Delta_{rev} = 0.4554 - 1.1564 \bar{e}$.

Potential landscape

In **Fig. 5b**, we illustrate the steady-state condition $r_\infty = \phi[w_{coop}(r_\infty - x_{eff})]$ in terms of an effective potential landscape $U(x)$. The functional form of this landscape was be obtained by integrating ‘restoring force’ $F(x)$ over activity x :

$$\begin{aligned} F(x) &= \Phi[w_{coop}(x - x_{eff})] - x \\ U(x) &= - \int_{x_{eff}}^x F(x') dx' \end{aligned}$$

Stochastic dynamics

Poisson-like variability

The discretely stochastic process $x(t) \in \{0, 1/N, 2/N, \dots, 1\}$ has a continuously stochastic ‘diffusion limit’, $x_{diff}(t)$, for $N \rightarrow \infty$, with identical mean $\langle x_{diff} \rangle = \langle x \rangle$ and variance $\langle x_{diff}^2 \rangle -$

Figure 12. Birth-death dynamics of evidence pools ensures Gamma-like distribution and ‘scaling property’ (invariance of distribution shape). **(a)** Representative examples for the time-development of evidence bias $\Delta e = e - e'$ between reversals (*i.e.*, between $-\Delta_{rev}$ and approximately $+\Delta_{rev}$). **(b)** Dominance distributions for $c=c'=0.0625$ (blue), $c=c'=0.25$ (green), and $c=c'=1.0$ (yellow). Distribution mean μ changes approximately threefold, but coefficient of variation c_v and skewness γ_1 are nearly invariant (inset), largely preserving distribution shape. **(c)** Development of expectation $\langle \Delta x \rangle$ between reversals (schematic). Left: a Poisson-variable process, such as the difference Δx between two birth-death processes. Mean $\langle \Delta x \rangle$ grows linearly with t (lines, with slopes μ, μ') and variance $\langle (\Delta x - \langle \Delta x \rangle)^2 \rangle$ grows linearly with $\sqrt{t}$ (dashed curves, with scaling factors σ, σ'). Constants μ and σ change with stimulus contrast (blue and red). Proportionality $\mu \propto \sigma^2$, ensures constant dispersion of Δx at threshold ($\delta_x = \delta'_x$), and, consequently, a dispersion of threshold-crossing times that grows linearly with mean threshold-crossing time ($\delta_t/t_{rev} = \delta'_t/t'_{rev} = const$), preserving distribution shape. Right: a process with constant variance, $\sigma = \sigma'$. Dispersion of Δx at threshold increases with threshold-crossing time ($\delta'_x > \delta_x$) and dispersion of threshold-crossing times grows supra-linearly with mean threshold-crossing time ($\delta_t/t_{rev} < \delta'_t/t'_{rev}$), broadening distribution shape.

$\langle x_{diff} \rangle^2 = \langle x^2 \rangle - \langle x \rangle^2$. This diffusion limit is a Cox-Ingersoll process and its dynamical equation

$$\dot{x}_{diff} = (1 - x_{diff}) \nu^+ - x_{diff} \nu^- + \sqrt{\frac{(1 - x_{diff}) \nu^+ + x_{diff} \nu^-}{N}} \xi(t),$$

where $\xi(t)$ is white noise, reveals that its increments $N \dot{x}_{diff}$ (and thus also the increments of the original discrete process) exhibit Poisson-like variability. Specifically, in the low-activity regime, $x_{diff} \ll 1$, both mean and variance of increments approximate activation rate $N\nu^+$:

$$\begin{aligned} \langle N \dot{x} \rangle &= \langle N \dot{x}_{diff} \rangle \simeq N\nu^+ \\ \langle N^2 \dot{x}^2 \rangle - \langle N \dot{x} \rangle^2 &= \langle N^2 \dot{x}_{diff}^2 \rangle - \langle N \dot{x}_{diff} \rangle^2 \simeq N\nu^+. \end{aligned}$$

Gamma-distributed first-passage times

When the input to a pool of bistable variables undergoes a step-change, the active fraction $x(t)$ transitions stochastically between old and new steady-states, x_{∞}^{old} and x_{∞}^{new} (set by old and new input values, respectively). The time that elapses until fractional activity crosses an intermediate ‘threshold’ level θ ($x_{\infty}^{old} < \theta < x_{\infty}^{new}$) is termed a ‘first-passage time’. In a low-threshold regime, birth-death processes exhibit a particular and highly unusual distribution of first-passage times.

Specifically, the distribution of first-passage times assumes a characteristic, Gamma-like shape for a wide range of value triplets $x_{\infty}^{old}, \theta, x_{\infty}^{new}$ [18]: skewness γ_1 takes a stereotypical value $\gamma_1 \simeq 2c_v$, the coefficient of variation c_v remains constant (as long as the distance between x_{∞}^{old} and θ remains the same), whereas the distribution mean may assume widely different values. This Gamma-like distribution shape is maintained even when shared input changes during the transition (*e.g.*, when bistable variables are coupled to each other) [18].

Importantly, only a birth-death process (*e.g.*, a pool of bistable variables) guarantees a Gamma-like distribution of first-passage times under different input conditions [19]. Many other discretely stochastic processes (*e.g.*, Poisson process) and continuously stochastic processes (*e.g.*, Wiener, Ornstein-Uhlenbeck, Cox-Ingersoll) produce inverse Gaussian distributions with $\gamma_1 \simeq 3c_v$. Models combining competition, adaptation, and noise can produce Gamma-like distributions, but require different parameter values for every input condition

(see **Methods: Alternative model**).

Scaling property

In the present model, first-passage times reflect the concurrent developments of two opponent birth-death processes (pools of $N=25$ binary variables). Dominance periods begin with newly non-dominant evidence e *well below* newly dominant evidence e' , $\Delta e = e - e' \simeq -\Delta_{rev}$, and end with the former *well above* the latter, $\Delta e \simeq +\Delta_{rev}$ (**Fig. 12a**). The combination of two small pools with $N=25$ approximates a single large pool with $N=50$. When image contrast changes, distribution shape remains nearly the same, with a coefficient of variation $c_v \simeq 0.6$ and a Gamma-like skewness $\gamma_1 \simeq 2$, even though mean μ of first-passage times changes substantially (**Fig. 12b**).

This ‘scaling property’ (preservation of distribution shape) is owed to the Poisson-like variability of birth-death processes (see above, **Fig. 12c**). Poisson-like variability implies that accumulation rate μ and dispersion rate σ^2 are proportional, $\mu \sim \sigma^2$. This proportionality ensures that activity at threshold disperses equally widely for different accumulation rates (*i.e.*, for different input strengths), preserving the shape of first-passage-time distributions [19].

Characteristic times

As mentioned previously, the characteristic times of pools of bistable variables are not fixed but vary with input (Eq. 2). In our model, the characteristic times of evidence activities lengthen with increasing input contrast and shorten with feedback suppression (**Fig. 13a**). Characteristic times are reflected also in the temporal autocorrelation, which averages over periods of dominance and non-dominance alike. Autocorrelation times lengthen with increasing input contrast, both in absolute terms and relative to the average dominance duration (**Fig. 13b**).

Importantly, the autocorrelation time of *mean* evidence activity $\bar{e} = (e + e')/2$ is even longer, particularly for high input contrast (**Fig. 13c**). The reason is that spontaneous fluctuations of $\bar{e}$ are constrained not only by birth-death dynamics, but additionally by the

Figure 13. Characteristic times of evidence activity. **(a)** Characteristic times $\tau_e, \tau_{e'}$ for different image contrast $c=c'$, when evidence pool is dominant (*dom*) and non-dominant (*sup*). **(b)** Auto-correlation of evidence activity e, e' as a function of image contrast (color) and latency, expressed in multiples of average dominance duration $\langle T \rangle$. **(c)** Autocorrelation of joint evidence activity $\bar{e}=(e + e')/2$ as a function of image contrast (color) and latency. Note that autocorrelation time lengthens substantially for high image contrast.

reversal dynamics that keeps evidence activities e and e' close together (*i.e.*, within reversal threshold Δ_{rev}). As a result, the characteristic timescale of spontaneous fluctuations of $\bar{e}$ lengthens with input contrast. The amplitude of such fluctuations also grows with contrast (not shown).

The slow fluctuations of $\bar{e}$ induce mirror-image fluctuations of reversal threshold Δ_{rev} and thus are responsible for the serial dependency of reversal sequences (see **Deterministic dynamics: Decision pools**).

Burstiness

The proposed mechanism predicts that reversal sequences include episodes with several successive short (or long) dominance periods. It further predicts that this inhomogeneity increases with image contrast. Such an inhomogeneity may be quantified in terms of a ‘burstiness index’ (BI), which compares the variability of the mean for sets of n successive

Figure 14. Burstiness of reversal sequences predicted by model and confirmed by experimental observations. **(a)** Burstiness index BI (mean) for n successive dominance periods in experimentally observed reversal sequences, for contrasts 12.5% (green), 50% (yellow) and 100% (red). **(b)** Burstiness index for reversal sequences generated by model (mean $\pm$ std).

periods to the expected variability for randomly shuffled reversal sequences. In both model and experimental observations, this index rises far above chance (over broad range of n) for high image contrast (**Fig. 14**). The degree of inhomogeneity expressed by the model at high image contrast is comparable to that observed experimentally, even though the model was neither designed nor fitted to reproduce non-stationary aspect of reversal dynamics. This correspondence between model and experimental observation compellingly corroborates the proposed mechanism.

-
- [1] S. Albert, K. Schmack, P. Sterzer, and G. Schneider. A hierarchical stochastic model for bistable perception. *PLOS Comput. Biol.*, 13(11):e1005856, 2017. doi: 10.1371/journal.pcbi.1005856.
 - [2] P. Ashwin and L. Aureliu. A low-dimensional model of binocular rivalry using winnerless competition. *Physica D*, 239:529–538, 2010.
 - [3] F. Attneave. Multistability in perception. *Scientific American*, 225(6):63–71, 1971.
 - [4] M. Bar. The proactive brain: memory for predictions. *Phil. Trans. R. Soc. Lond. B*, 364(1521):1235–1243, 2009.
 - [5] H. D. Barlow, R. Narasimhan, and A. Rosenfeld. Visual pattern analysis in machines and animals. *Science.*, 177(49):576–75, 1972.
 - [6] J. Beck, W. Ma, R. Kiani, T. Hanks, A. Churchland, J. Roitman, M. Shadlen, P. Latham, and A. Pouget. Probabilistic population codes for bayesian decision making. *Neuron*, 60: 1142–52, 2008.
 - [7] R. Blake and N. K. Logothetis. Visual competition. *Nat. Rev. Neurosci.*, 3:13–21, 2002.
 - [8] R. R. Blake, R. Fox, and C. McIntyre. Stochastic properties of stabilized-image binocular rivalry alternations. *J. Exp. Psychol.*, 88(3):327–332., 1971.
 - [9] R. Bogacz. Optimal decision-making theories: linking neurobiology with behaviour. *Trends Cogn. Sci. (Regul. Ed.)*, 11:118–25, 2007.
 - [10] R. Bogacz, E. Brown, J. Moehlis, P. Holmes, and J. D. Cohen. The physics of optimal decision making: a formal analysis of models of performance in two-alternative forced-choice tasks. *Psychol Rev*, 113(4):700–765, 2006.
 - [11] Y. Bonnef, D. Sagi, and A. Karni. A transition between eye and object rivalry determined by stimulus coherence. *Vision Res.*, 41:981989, 2001.
 - [12] A. Borsellino, A. De Marco, A. Allazetta, S. Rinesi, and B. Bartolini. Reversal time distribution in the perception of visual ambiguous stimuli. *Kybernetik*, 10:139–44, 1972.
 - [13] J. W. Brascamp, R. van Ee, W. R. Pestman, and A. V. van den Berg. Distributions of alternation rates in various forms of bistable perception. *J. Vision*, 5(4):287–98, 2005.

- [14] J. W. Brascamp, R. van Ee, A. J. Noest, R. H. A. H. Jacobs, and A. V. Van den Berg. The time course of binocular rivalry reveals a fundamental role of noise. *J. Vision*, 6(11):1244–1256, 2006.
- [15] J. W. Brascamp, P. C. Klink, and J. M. Levelt. The ‘Laws’ of binocular rivalry : 50 years of Levelt’s propositions. *Vision Research.*, 109:20–37, 2015.
- [16] J. W. Brascamp, C. S. Qian, D. Z. Hambrick, and M. W. Becker. Individual differences point to two separate processes involved in the resolution of binocular rivalry. *J Vis*, 19(12):15, 2019.
- [17] J. Braun and M. Mattia. Attractors and noise: Twin drivers of decisions and multistability. *Neuroimage*, 52(3):740–751, 2010.
- [18] R. Cao, J. Braun, and M. Mattia. Stochastic accumulation by cortical columns may explain the scalar property of multi-stable perception. *Phys. Rev. Lett.*, 113(9):098103, 2014.
- [19] R. Cao, A. Pastukhov, M. Mattia, and J. Braun. Collective activity of many bistable assemblies reproduces characteristic dynamics of multistable perception. *Journal of Neuroscience*, 36(26):6957–6972, 2016.
- [20] M. Carandini and D. J. Heeger. Normalization as a canonical neural computation. *Nat Rev Neurosci*, 13(1):51–62, 2011.
- [21] M. M. Churchland, B. M. Yu, J. P. Cunningham, L. P. Sugrue, M. R. Cohen, G. S. Corrado, W. T. Newsome, A. M. Clark, P. Hosseini, B. B. Scott, D. C. Bradley, M. A. Smith, A. Kohn, J. A. Movshon, K. M. Armstrong, T. Moore, S. W. Chang, L. H. Snyder, S. G. Lisberger, N. J. Priebe, I. M. Finn, D. Ferster, S. I. Ryu, G. Santhanam, M. Sahani, and K. V. Shenoy. Stimulus onset quenches neural variability: a widespread cortical phenomenon. *Nat Neurosci*, 13(3):369–378, 2010.
- [22] P. Cisek and J. F. Kalaska. Neural mechanisms for interacting with a world full of action choices. *Annu. Rev. Neurosci.*, 33:269–98, 2010.
- [23] B. P. Cohen, C. C. Chow, and S. Vattikuti. Multi-scale variability in neuronal competition. *Communications Biology*, 2(1):319–330, 2019.
- [24] A. Compte, C. Constantinidis, J. Tegner, S. Raghavachari, M. Chafee, P. Goldman-Rakic, and X. Wang. Temporally irregular mnemonic persistent activity in prefrontal neurons of

- monkeys during a delayed response task. *J Neurophysiol*, 90:3441–3454, 2003.
- [25] L. Cossell, M. F. Iacaruso, D. R. Muir, R. Houlton, E. N. Sader, H. Ko, S. B. Hofer, and T. D. Mrsic-Flogel. Functional organization of excitatory synaptic strength in primary visual cortex. *Nature*, 518(7539):399–403, 2015.
 - [26] Y. Cui, L. D. Liu, J. M. McFarland, C. C. Pack, and D. A. Butts. Inferring cortical variability from local field potentials. *J Neurosci*, 36(14):4121–4135, 2016.
 - [27] P. Dayan. A hierarchical model of binocular rivalry. *Neural Computation*, 10(5):1119–35, jul 1998.
 - [28] A. De Marco, P. Penengo, A. Trabucco, A. Borsellino, F. Carlini, M. Piani, and M. T. Tuccio. Stochastic models and fluctuations in reversal time of ambiguous figures. *Perception*, 6(6):645–656, 1977.
 - [29] G. Deco and E. Hugues. Neural network mechanisms underlying stimulus driven variability reduction. *PLoS Comput Biol*, 8(3):e1002395, 2012.
 - [30] G. Deco and E. T. Rolls. Attention, short-term memory, and action selection: a unifying theory. *Prog. Neurobiol.*, 76(4):236–256, 2005.
 - [31] G. Deco, L. Scarano, and S. Soto-Faraco. Weber’s law in decision making: integrating behavioral data in humans with a neurophysiological model. *J Neurosci*, 27(42):11192–11200, 2007.
 - [32] S. L. Denham, D. Farkas, R. van Ee, M. Taranu, Z. Kocsis, M. Wimmer, D. Carmel, and I. Winkler. Similar but separate systems underlie perceptual bistability in vision and audition. *Scientific Reports*, 8(7106):1–10, 2018.
 - [33] E. Diaz-Caneja. Sur l’alterance binoculaire. *Ann. Occul. (Paris)*, pages 721–731, 1928.
 - [34] J. Ditterich, M. E. Mazurek, and M. N. Shadlen. Microstimulation of visual cortex affects the speed of perceptual decisions. *Nat Neurosci*, 6(8):891–898, 2003.
 - [35] P. Dixit, J. Wagoner, C. Weistuch, S. Presse, K. Ghosh, and K. Dill. *J. Chem. Phys.*, 148(1):010901, 2018.
 - [36] B. Doiron and A. Litwin-Kumar. Balanced neural architecture and the idling brain. *Front Comput Neurosci*, 8:56, 2014.

- [37] T. A. Engel, N. A. Steinmetz, M. A. Gieselmann, A. Thiele, T. Moore, and K. Boahen. Selective modulation of cortical state during spatial attention. *Science*, 354(6316), 2016.
- [38] J. Fiser, C. Chiu, and M. Weliky. Small modulation of ongoing cortical dynamics by sensory input during natural vision. *Nature*, 431(7008):573–578, 2004.
- [39] R. Fox and J. Herrmann. Stochastic properties of binocular rivalry alternations. *Percept. Psychophys.*, 2:432–6, 1967.
- [40] A. Funamizu, B. Kuhn, and K. Doya. Neural substrate of dynamic Bayesian inference in the cerebral cortex. *Nature Neuroscience.*, 19(12):1682–1689, 2016.
- [41] G. Gigante, M. Mattia, J. Braun, and P. Del Giudice. Bistable perception modeled as competing stochastic integrations at two levels. *PLoS Comput. Biol.*, 5:e1000430, 2009.
- [42] J. Gold and M. Shadlen. The neural basis of decision making. *Annu. Rev. Neurosci.*, 30: 535–74, 2007.
- [43] J. Gold and A. Stocker. Visual decision-making in an uncertain and dynamic world. *Annu. Rev. Vis. Sci.*, 3:227–50, 2017.
- [44] R. L. Gregory. Perception as hypotheses. *Phil. Trans. R. Soc. Lond. Biol. Sci.*, 290(1038): 181–197, 1980.
- [45] P. Hanggi, P. Talkner, and M. Borkovec. Reaction-rate theory: fifty years after kramers. *Rev. Mod. Phys.*, 62:251–341, 1990.
- [46] J. Hohwy, A. Roepstorff, and K. Friston. Predictive coding explains binocular rivalry: an epistemological review. *Cognition*, 108(3):687–701, 2008.
- [47] C. Huang and B. Doiron. Once upon a (slow) time in the land of recurrent neuronal networks... *Curr Opin Neurobiol*, 46:31–38, 2017.
- [48] M.-S. Kang. Size matters: a study of binocular rivalry dynamics. *Journal of vision*, 9(1): 17.1–11, 2009.
- [49] S. Karlin and J. L. McGregor. Ehrenfest Urn Models. *J. Appl. Probab.*, 2(2):352–376, 1965.
- [50] Y.-J. Kim, M. Grabowecky, and S. Suzuki. Stochastic resonance in binocular rivalry. *Vision Res.*, 46:392–406, 2006.
- [51] P. C. Klink, R. van Ee, and R. J. a. van Wezel. General validity of levelt’s proportions reveals common computational mechanisms for visual rivalry. *PLoS ONE*, 3(10):1–9, 2008.

- [52] C. Koch and S. Ullman. Shifts in selective visual attention: towards the underlying neural circuitry. *Hum. Neurobiol.*, 4(4):219–227, 1985.
- [53] K. Koffka. *Principles of Gestalt Psychology*. Harcourt Brace, 1935.
- [54] I. Kovacs, T. Papathomas, and A. Yang, M Feheer. When the brain changes its mind: Interocular grouping during binocular rivalry. *Proc. Natl Acad. Sci. USA*, 93:550815511, 1997.
- [55] H. Kramers. Brownian motion in a field of force and the diffusion model of chemical reactions. *Physica*, 7:284–304, 1940.
- [56] N. Kriegeskorte and P. K. Douglas. Cognitive computational neuroscience. *Nat Neurosci*, 21(9):1148–1160, 2018.
- [57] K. Krug. Coding perceptual decisions: from single units to emergent signaling properties in cortical circuits. *Annu. Rev. Vis. Sci.*, 6(6):xxx–xxx, 2020.
- [58] G. La Camera, A. Fontanini, and L. Mazzucato. Cortical computations via metastable activity. *Curr Opin Neurobiol*, 58:37–45, 2019.
- [59] C. R. Laing and C. C. Chow. A spiking neuron model for binocular rivalry. *J. Comput. Neurosci.*, 12:39–53, 2002.
- [60] M. Larkum. A cellular mechanism for cortical associations : an organizing principle for the cerebral cortex. *Trends in Neurosciences*, 36(3):141–151, 2013.
- [61] T. S. Lee and D. Mumford. Hierarchical Bayesian inference in the visual cortex. *J. Opt. Soc. Am. A*, 20(7):1434–1448, 2003.
- [62] W.-C. A. Lee, V. Bonin, M. Reed, B. J. Graham, G. Hood, K. Glattfelder, and R. C. Reid. Anatomy and function of an excitatory network in the visual cortex. *Nature*, 532(7599):370–374, 2016.
- [63] S. R. Lehky. An astable multivibrator model of binocular rivalry. *Perception*, 17:215–28, 1988.
- [64] A. Leopold and N. K. Logothetis. Stable perception of visually ambiguous patterns. *Nat. Neurosci.*, 5:605–9, 2002.
- [65] D. A. Leopold and N. K. Logothetis. Multistable phenomena: changing views in perception. *Trends Cogn. Sci.*, 3(7):254–264, 1999.

- [66] W. J. M. Levelt. *On binocular rivalry*. Van Gorkum Comp., Leiden, 1965.
- [67] A. Litwin-Kumar and B. Doiron. Slow dynamics and high variability in balanced cortical networks with clustered connections. *Nat. Neurosci.*, 15(11):1498–1505, 2012.
- [68] N. Logothetis, D. Leopold, and D. Sheinberg. What is rivaling during binocular rivalry? *Nature*, 380:621624, 1996.
- [69] R. D. Luce. *Response time: their role in inferring elementary mental organization*. New York: Oxford University Press, 1986.
- [70] W. Ma, J. Beck, P. Latham, and A. Pouget. Bayesian inference with probabilistic population codes. *Nat. Neurosci.*, 9:1432–8, 2006.
- [71] W. J. Ma, J. M. Beck, and A. Pouget. Spiking networks for Bayesian inference and choice. *Curr. Opin. Neurobiol.*, 18(2):217–222, 2008.
- [72] A. Maier, M. Wilke, N. K. Logothetis, and D. Leopold. Perception of temporally interleaved ambiguous patterns. *Curr. Biol.*, 13(13):1076–85, 2003.
- [73] E. Marcos, S. Tsujimoto, M. Mattia, and A. Genovesio. A network activity reconfiguration underlies the transition from goal to action. *Cell Rep.*, 27(10):2909–2920.e4, 2019. doi: 10.1016/j.celrep.2019.05.021.
- [74] M. Mattia, P. Pani, G. Mirabella, S. Costa, P. Del Giudice, and S. Ferraina. Heterogeneous attractor cell assemblies for motor planning in premotor cortex. *J. Neurosci.*, 33(27):11155–68, 2013.
- [75] L. Mazzucato, A. Fontanini, and G. La Camera. Dynamics of multistable states during ongoing and evoked cortical activity. *J Neurosci*, 35(21):8214–8231, 2015.
- [76] L. Mazzucato, G. La Camera, and A. Fontanini. Expectation-induced modulation of metastable activity underlies faster coding of sensory stimuli. *Nat Neurosci*, 22(5):787–796, 2019.
- [77] P. Miller. Itinerancy between attractor states in neural systems. *Curr Opin Neurobiol*, 40: 14–22, 2016.
- [78] M. B. Mirza, R. A. Adams, C. D. Mathys, and K. J. Friston. Scene Construction , Visual Foraging , and Active Inference. *Frontiers in Computational Neuroscience.*, 10(56):1–16, 2016.

- [79] R. Moreno-Bote, J. Rinzel, and N. Rubin. Noise-induced alternations in an attractor network model of perceptual bistability. *J. Neurophysiol.*, 98:1125–39, 2007.
- [80] R. Moreno-Bote, A. Shpiro, J. Rinzel, and N. Rubin. Alternation rate in perceptual bistability is maximal at and symmetric around equi-dominance. *J. Vision*, 10(11):1, 2010.
- [81] D. R. Muir and R. J. Douglas. From neural arbors to daisies. *Cereb. Cortex*, 21(5):1118–1133, May 2011.
- [82] T. Murata, N. Matsui, S. Miyauchi, Y. Kakita, and T. Yanagida. Discrete stochastic process underlying perceptual rivalry. *Neuroreport*, 14(10):1347–52, 2003.
- [83] L. A. Necker. Observations on some remarkable phenomena seen in Switzerland; and an optical phenomenon which occurs on viewing of a crystal or geometrical solid. *Philosophical Magazine*, 1:329–337, 1832.
- [84] M. Okun, N. a. Steinmetz, L. Cossell, M. F. Iacaruso, H. Ko, P. Barthó, T. Moore, S. B. Hofer, T. D. Mrsic-Flogel, M. Carandini, and K. D. Harris. Diverse coupling of neurons to populations in sensory cortex. *Nature*, 512(7553):511–515, 2015.
- [85] T. Parr and K. J. Friston. The active construction of the visual world. *Neuropsychologia*, 104:92–101, 2017.
- [86] T. Parr and K. J. Friston. Uncertainty , epistemics and active inference. *J. R. Soc. Interface.*, 14(20170376):1–10, 2017.
- [87] X. T. Parr, X. M. B. Mirza, H. Cagnan, and X. K. J. Friston. Dynamic Causal Modelling of Active Vision. *J. Neurosci.*, 39(32):6265–6275, 2019.
- [88] A. Pastukhov and J. Braun. Perceptual reversals need no prompting by attention. *J. Vision*, 7(10):5.1–17, 2007.
- [89] A. Pastukhov, P. Garcia-Rodriguez, J. Haenicke, A. Guillamon, G. Deco, and J. Braun. Cumulative history quantifies the role of neural adaptation in multistable perception. *Front. Comput. Neurosci.*, 7:17, 2013.
- [90] J. Pearl. *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. Morgan Kaufmann, 1988.
- [91] G. Pezzulo, F. Rigoli, and K. J. Friston. Hierarchical active inference: a theory of motivated control. *Trends in Cognitive Sciences*, 22(4):294–306, 2018.

- [92] A. Ponce-Alvarez, V. Nacher, R. Luna, A. Riehle, and R. Romo. Dynamics of cortical neuronal ensembles transit from decision making to storage for later report. *J Neurosci*, 32(35):11956–11969, 2012.
- [93] A. Pouget, J. M. Beck, W. J. Ma, and P. E. Latham. Probabilistic brains : knowns and unknowns. *Nature Neuroscience*, 16(9), 2013.
- [94] S. Pressé, K. Gosh, J. Lee, and K. A. Dill. Principles of maximum entropy and maximum caliber in statistical physics. *Reviews of Modern Physics*, 85:1115–41, 2013.
- [95] D. Pressnitzer and J. M. Hupe. Temporal dynamics of auditory and visual bistability reveal common principles of perceptual organization. *Curr. Biol.*, 16(13):1351–7, 2006.
- [96] V. S. Ramachandran and S. M. Anstis. Perceptual organization in multistable apparent motion. *Perception*, 144:135–143, 1985.
- [97] R. P. N. Rao and D. H. Ballard. Predictive coding in the visual cortex : a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, 2(1):79–87, 1999.
- [98] R. Ratcliff and P. L. Smith. A comparison of sequential sampling models for two-choice reaction time. *Psychol. Rev.*, 111(2):333–367, 2004.
- [99] E. L. Rich and J. D. Wallis. Decoding subjective decisions from orbitofrontal cortex. *Nat Neurosci*, 19(7):973–980, 2016.
- [100] I. Rock. *The Logic of Perception*. MIT Press., 1983.
- [101] R. Rosenbaum, M. A. Smith, A. Kohn, J. E. Rubin, and B. Doiron. The spatial structure of correlated neuronal variability. *Nat Neurosci*, 20(1):107–114, 2017.
- [102] E. Rubin. Figure and ground. In D. C. Beardslee and M. Wertheimer, editors, *Readings in Perception*, pages 194–203. Van Nostrand, 1958.
- [103] M. T. Schaub, Y. N. Billeh, C. A. Anastassiou, C. Koch, and M. Barahona. Emergence of slow-switching assemblies in structured neuronal networks. *PLoS Comput Biol*, 11(7):e1004196, 2015.
- [104] P. Schrater and R. Sundaeswara. Theory and dynamics of derceptual bistability. *Adv. Neural Inf. Proc. Sys.*, 19:1217–1224, 2006.

- [105] J. L. Schwartz, N. Grimault, J. M. Hupe, B. C. Moore, and D. Pressnitzer. Multistability in perception: binding sensory modalities, an overview. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, 367(1591):896–905, 2012.
- [106] J. Seely and C. C. Chow. Role of mutual inhibition in binocular rivalry. *J. Neurophysiol.*, 106(5):2136–50, 2011.
- [107] S. Shipp. Neural Elements for Predictive Coding. *Front. Psychol.*, 7(1792):1–21, 2016.
- [108] A. Shpiro, R. Moreno-Bote, N. Rubin, and J. Rinzel. Balance between noise and adaptation in competition models of perceptual bistability. *J. Comput. Neurosci.*, pages 37–54, 2009.
- [109] M. Stern, H. Sompolinsky, and L. F. Abbott. Dynamics of random neural networks with bistable units. *Phys. Rev. E*, 90, Dec 2014.
- [110] R. Sundareswara and P. R. Schrater. Perceptual multistability predicted by search model for Bayesian decisions Rashmi Sundareswara. *J. Vis.*, 8:1–19, 2008.
- [111] M. M. Taylor and K. D. Ladridge. Stochastic processes in reversing figure perception. *Percept. Psychophys.*, 16(1):9–27, 1974.
- [112] J. Ternus. Untersuchungen zur Lehre von der Gestalt. *Psychologische Forschung*, 7:81–136, 1926.
- [113] I. Tsuda. Toward an interpretation of dynamic neural activity in terms of chaotic dynamical systems. *Behav. Brain Sci.*, 24:793–810, 2001.
- [114] R. Van Ee. Dynamics of perceptual bi-stability for stereoscopic slant rivalry and a comparison with grating, house-face, and Necker cube rivalry. *Vision Res.*, 45:29–40, 2005.
- [115] R. van Ee. Stochastic variations in sensory awareness are driven by noisy neuronal adaptation: evidence from serial correlations in perceptual bistability. *J. Opt. Soc. Am. A*, 26(12):2612–22, 2009.
- [116] N. G. van Kampen. *Stochastic processes in physics and chemistry*. North-Holland Physics Publishing, Amsterdam, 1981.
- [117] A. Veliz-Cuba, Z. P. Kilpatrick, and K. Josic. Stochastic models of evidence accumulation in changing environments. *SIAM Review*, 58(2):264–289, 2016.
- [118] H. von Helmholtz. *Handbuch der Physiologischen Optik*. Leopold Voss, 1867.

- [119] P. Walker. Stochastic properties of binocular rivalry alternations. *Percept. & Psychophys.*, 18(6):467–473, 1975.
- [120] X.-J. Wang. Probabilistic decision making by slow reverberation in cortical circuits. *Neuron*, 36:955–68, 2002.
- [121] X.-J. Wang. Decision making in recurrent neuronal circuits. *Neuron*, 60(2):215–234, 2008.
- [122] X.-J. Wang. Neural dynamics and circuit mechanisms of decision-making. *Curr. Opin. Neurobiol.*, 22(6):1039–1046, 2012.
- [123] B. Wark, A. Fairhall, and F. Rieke. Timescales of inference in visual adaptation. *Neuron*, 61(5):750–61, 2009.
- [124] V. Weilhhammer, H. Stuke, G. Hesselmann, P. Sterzer, and K. Schmack. A predictive coding account of bistable perception - a model-based fMRI study. *PLoS Computational Biology.*, 13(5):1–21, 2017.
- [125] M. Wertheimer. Zeitschrift für Psychologie mit Zeitschrift für angewandte Psychologie. *Experimentelle Studien über das Sehen von Bewegung*, 61:161–165, 1912.
- [126] C. Wheatstone. Contributions to the Physiology of Vision . –Part the First . On Some Remarkable , and Hitherto Unobserved , Phenomena of Binocular Vision. *Philosophical Transactions of the Royal Society of London.*, 128(1838):371–394, 1838.
- [127] H. R. Wilson. Computational evidence for a rivalry hierarchy in vision. *Proc. Natl. Acad. Sci. U.S.A.*, 100:14499–503, 2003.
- [128] H. R. Wilson. Minimal physiological conditions for binocular rivalry and rivalry memory. *Vision Res.*, 47:2741–50, 2007.
- [129] S. C.-h. Yang, D. M. Wolpert, and M. Lengyel. Theoretical perspectives on active sensing. *Curr. Opin. Behav. Sci.*, 11:100–108, 2018.
- [130] A. Yuille and D. Kersten. Theoretical perspectives on active sensing. *Trends Cogn. Sci.*, 10(7):301–308, 2006.