Optimal policy for value-based decision-making

Tajima, Satohiro; Drugowitsch, Jan; Pouget, Alexandre

How to cite

TAJIMA, Satohiro, DRUGOWITSCH, Jan, POUGET, Alexandre. Optimal policy for value-based decision-making. In: Nature communications, 2016, vol. 7, p. 12400. doi: 10.1038/ncomms12400

This publication URL: https://archive-ouverte.unige.ch/unige:86850

Publication DOI: 10.1038/ncomms12400

nature COMMUNICATIONS

ARTICLE

Received 7 Oct 2015 | Accepted 29 Jun 2016 | Published 18 Aug 2016

DOI: 10.1038/ncomms12400

OPEN

Optimal policy for value-based decision-making

Satohiro Tajima[1,*], Jan Drugowitsch[1,2,*] & Alexandre Pouget[1,3,4]

For decades now, normative theories of perceptual decisions, and their implementation as drift diffusion models, have driven and significantly improved our understanding of human and animal behaviour and the underlying neural processes. While similar processes seem to govern value-based decisions, we still lack the theoretical understanding of why this ought to be the case. Here, we show that, similar to perceptual decisions, drift diffusion models implement the optimal strategy for value-based decisions. Such optimal decisions require the models' decision boundaries to collapse over time, and to depend on the *a priori* knowledge about reward contingencies. Diffusion models only implement the optimal strategy under specific task assumptions, and cease to be optimal once we start relaxing these assumptions, by, for example, using non-linear utility functions. Our findings thus provide the much-needed theory for value-based decisions, explain the apparent similarity to perceptual decisions, and predict conditions under which this similarity should break down.

[1] Département des Neurosciences Fondamentales, University of Geneva, Rue Michel-Servet 1, Genève 1211, Switzerland. [2] Department of Neurobiology, Harvard Medical School, 220 Longwood Avenue, Boston, Massachusetts 02115, USA. [3] Department of Brain and Cognitive Sciences, University of Rochester, Rochester, NY USA. [4] Gatsby Computational Neuroscience Unit, University College of London, London, UK. * These authors contributed equally to this work. Correspondence and requests for materials should be addressed to A.P. (email: alexandre.pouget@unige.ch).

In everyday ambiguous and noisy environments, decision-making requires the accumulation of evidence over time. In perceptual decision-making tasks (for example, discriminating a motion direction), choices and reaction times are well-fit by drift diffusion models (DDMs)[1–3]. These models represent the accumulated belief informed by sensory evidence as the location of a diffusing particle that triggers a decision once it reaches one of two decision boundaries. DDMs are known to implement theoretically optimal algorithms, such as the sequential likelihood ratio test[4–6] and more general algorithms that handle varying task difficulty[7].

Recently, DDMs have been shown to also describe human behaviour in value-based decisions, where subjects compare the endogenous values of rewarding items (for example, deciding between two lunch options). This suggests that humans perform value-based decisions by computations similar to those used for standard perceptual decision (such as visual discrimination of random dot motion directions). In this case, the DDMs are driven only by the difference in item values, and thus predict the choices to be insensitive to the absolute values of the compared items (Fig. 1)[8–10]. In particular, relying only on the relative value means that it might take on average longer to decide between two equally good options than between items of very different values.

This raises an important question: do DDMs indeed implement the optimal strategy for value-based decisions? Intuitively, absolute values should also influence the decision strategy, such that relying only on relative values appears suboptimal. In particular, it seems unreasonable to wait for a long time to decide between two nearly similar highly rewarding options. Nonetheless, DDMs or related models are generally better at explaining human behaviour than alternative models. For example, race models (RMs) assume independent 'races' to accumulate evidence for individual options. Once one of these races reaches a decision criterion the corresponding choice is triggered[11,12]. Even though RMs are sensitive to absolute choice values and as such predict more rapid choices for higher rewarded options, they neither fit well human behaviour in perceptual decision-making tasks[12] nor in value-based decision tasks in which decisions are usually better described by relying only on relative values[13,14]. Does this mean that humans use DDMs even though these model implement suboptimal strategies, or that DDMs indeed implement the optimal strategy for value-based choices? What is clear is that we need to understand (i) what the optimal strategy for value-based decisions is, (ii) why the value-based and perceptual decision seem to be fitted by the same class of models (DDMs) despite the qualitative difference between these tasks and (iii) to which degree value-based and the perceptual decisions differ in terms of their normative computational strategies.

In this paper, we derive the theoretically optimal strategy for value-based decisions, and show that this strategy is in fact equivalent to a particular class of DDMs that feature 'collapsing boundaries' whose distance shrinks over time. We show that the exact shape of these boundaries and the associated average reaction times depend on average-reward magnitudes even if decisions within individual trials are only guided by the relative reward between choice options. Finally, we highlight the difference between value-based and standard perceptual decisions, reveal specific conditions under which the optimality of DDMs are violated, and show how to reconcile the ongoing debate on whether decision makers are indeed using collapsing decision boundaries. In contrast to previous work that assumed particular *a priori* mechanisms underlying value-based choices, such as RMs or DDMs[15–18], our work instead deduces optimal decision-making mechanisms based solely on a description of the information available to the decision maker. Thus, the use of diffusion models for value-based choices is not an *a priori* assumption of our work, but rather a result that follows from the normative decision-making strategy.

Results

Problem setup and aim. Consider a decision maker choosing between options that yield potentially different rewards (or 'values'), as, for example, choosing between two lunch menu options in the local restaurant. If the decision maker knew these rewards precisely and immediately then she should instantly choose the more rewarding option. However, in realistic scenarios, the reward associated with either option is uncertain *a priori*. This uncertainty might, for example, arise if she has *a priori* limited information about the choice options. Then, it is better to gather more evidence about the reward associated with the compared options before committing to a choice (for example, when choosing among lunch menus, we can reduce uncertainty about the value of either menu by contemplating the composition of each menu course separately and how these separate courses complement each other). However, how much evidence should

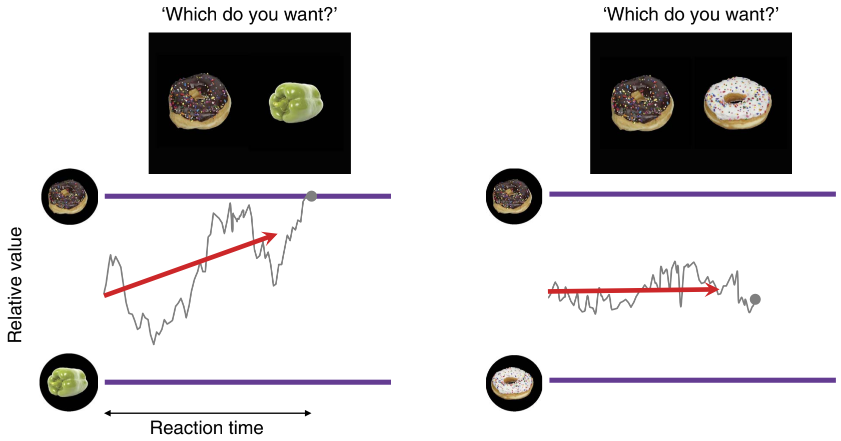

Figure 1 | DDMs for value-based decisions. The purple lines represent the decision boundaries. The red arrows indicate the mean drift rate. The grey fluctuating traces illustrate sample trajectories of a particle that drifts in the space of relative value between two given options. (left) If an option is preferred (that is, yields higher reward) than the other, the mean drift is biased toward the boundary of preferred option, making the particle to hit the decision boundary within a relatively short time. (right) However, if the given options are equally good, DDM assumes a mean drift without any bias, requiring much longer time for the particle to hit either decision boundary—even if both options are highly rewarding.

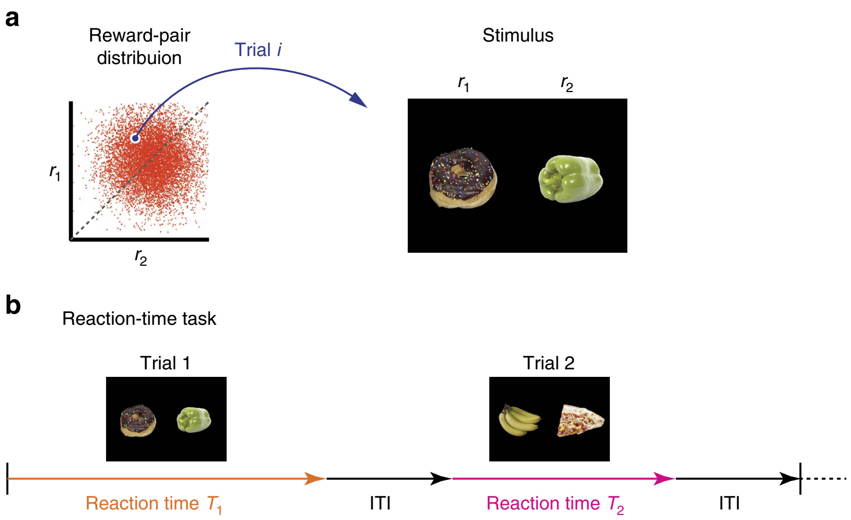

Figure 2 | Typical value-based decision-making task. (**a**) (left) Prior distribution from which the rewards for a pair of options are sampled. r_1 and r_2 indicate the reward magnitudes for individual options (for example, objects presented on left- and right-hand side of the screen in each trial). (right) A typical visual stimulus in each trial. (**b**) Reaction-time task. In each trial, the decision maker is presented with a pair of options. The decision maker reports her choice as soon as she has decided which of the two options she prefers. The reaction time (T_i) can vary among trials, and individual trials are separated by a fixed inter-trial interval.

we accumulate before committing to a choice? Too little evidence might result in the choice of the lower-rewarding option (the less appreciated lunch menu), whereas long evidence accumulation comes at the cost of both time and effort (for example, missing the passing waiter yet another time). In what follows, we formalize how to best tradeoff speed and accuracy of such choices, and then derive how the decision maker ought to behave in such scenarios. We first introduce each component of the decision-making task in its most basic form, and discuss generalizations thereof in later sections.

We assume that, at the beginning of each trial, the two options have associated true rewards, z_1 and z_2, which are each stochastically drawn from separate normal distributions with a fixed mean $\bar{z}_j$ for option $j\in\{1,2\}$ and common variance σ_z^2. These true rewards are unknown to the decision maker, as they are never observed directly. Instead, we assume that the decision maker observes some momentary evidence with mean $\bar{z}_j\delta t, \delta x_{j,i} \sim \mathcal{N}(z_j\delta t, \sigma^2\delta t)$ for both options $j\in\{1,2\}$ simultaneously in small time-steps i of duration δt. Note that variability (and associated ambiguity) of the momentary evidence can arise through noise sources that are both internal or external to the decision maker—sources that we discuss in more detail further below.

Before observing any momentary evidence, we assume that the decision maker holds a normally distributed belief $z_j \sim \mathcal{N}(\bar{z}_j, \sigma_z^2)$ with mean $\bar{z}_j$ and variance σ_z^2, which are, respectively, the mean and variance of the distribution from which the reward are being drawn from at the beginning of each trial. In other words, this a priori belief corresponds to the actual distribution from which the true rewards are drawn (that is, the decision maker uses the correct generative model), and entails that option j is most likely to yield reward $\bar{z}_j$, but might also yield other rewards, with the spread of rewards around $\bar{z}_j$ controlled by the level of uncertainty σ_z^2 about z_j. For now, we only consider the case in which the amounts of reward associated with both options are uncorrelated and, on average, the same ($\bar{z}_1 = \bar{z}_2$). In terms of choosing between lunch menu options, either menu would *a priori* yield the same reward, and the true rewards of either menu option are independently of each other drawn from the aforementioned normal distribution (Fig. 2a). Later, we discuss the consequences of a correlation between true option values.

As soon as being presented with sensory evidence $\delta x_{j,i}$, the decision maker accumulates further information about the rewards associated with either choice option. This momentary evidence $\delta x_{j,i}$ reveals noisy information about the true reward z_j, such that each additional piece of momentary evidence reduces the uncertainty about this reward. We emphasize that neither of the true rewards is ever observed without noise. As a result, the decision maker needs to accumulate evidence to reduce uncertainty about the underlying true rewards by averaging out the noise. Longer evidence accumulation results in a better average and lower associated uncertainty.

The noise in the momentary evidence itself can have both internal and external sources. External sources constitute the potentially stochastic nature of stimuli, perceptual noise, ambiguity and incomplete knowledge. For example, having not yet read the main course and dessert of a particular menu option causes uncertainty about the option's value due to incomplete knowledge. Internal sources could result from uncertain memory, or value inference that extends over time. One example for such value inference would be to sequentially contemplate the value of different features of a particular menu course over time.

Formally, after observing the value-related evidence $\delta x_j(0:t)$ from time 0 (onset of momentary evidence) to some time t, the decision-maker's posterior belief about the true reward, z_j, of option j is given by

$$z_j \mid \delta x_j(0:t) \sim \mathcal{N}\left(\frac{\sigma^2/t}{\sigma^2/t+\sigma_z^2}\bar{z}_j + \frac{\sigma_z^2}{\sigma^2/t+\sigma_z^2}\frac{x_j(t)}{t}, \frac{\sigma_z^2\sigma^2/t}{\sigma^2/t+\sigma_z^2}\right). \quad (1)$$

The posterior mean is an evidence-weighted combination of the *a priori* mean $\bar{z}_j$and the time-averaged accumulated evidence $x_j(t = n\delta t) = \sum_{i=1}^{n}\delta x_{j,i}$, and the posterior variance (that is uncertainty) decreases monotonically with time (see Methods section). Due to uncertainty in the momentary evidence, the accumulated evidence $x_j(t)$ itself describes a stochastic process. Here, and in contrast to other models of decision-making (both perceptual[19,20] and value-based[15,16]), all stochasticity in the accumulated evidence results from ambiguity in the momentary evidence itself, rather than from noise in the mechanisms that implement the decision-making process. In other words, the process responsible for the accumulation of the evidence is assumed to be noiseless, an assumption consistent with recent neurophysiological recordings.[21]

What are the costs and rewards that the decision maker incurs during the course of her decisions? In terms of costs we assume that the decision maker pays a cost c per second of accumulating evidence, from onset of the choice options until an option is chosen. This cost could, for example, be an explicit cost for delayed choices, or represent the effort induced by evidence accumulation. In the context of choosing between lunch menus, this cost might arise from missing the passing waiter yet again, or from being late for a post-lunch meeting. Choosing option j is associated with experiencing some reward r_j that is a function of the true reward z_j associated with this option, as, for example, when experiencing reward for consuming the lunch. For now, we assume experienced and true reward to be equivalent, that is $r_j = z_j$. For a single choice, the overall aim of the decision maker is to maximize expected reward minus expected cost,

$$\langle r_j \mid \delta x_j(0:T) \rangle - c\langle T \rangle, \tag{2}$$

where the expectation is across choices j and evidence accumulation times T, given the flow of evidence δx_j (0:T) from time 0 to T. We first derive the optimal behaviour, or 'policy', that maximizes this objective function for single, isolated choices and later generalize it to the more realistic scenario in which the total reward in a long consecutive sequence is maximized.

Optimal decisions with DDMs with collapsing boundaries. To find the optimal policy, we borrow tools from dynamic programming (DP). One of these tools is the 'value function', which can be defined recursively through Bellman's equation. In what follows, we show that the optimal policy resulting from this value function is described by two time-dependent parallel bounds in the two-dimensional space of current estimates of the true option rewards. These bounds are parallel with unity slopes, approach each other over time and together form a bound on the difference of reward estimates. This difference is efficiently inferred by diffusion models, such that DDMs can implement the optimal strategy for value-based decision-making.

Bellman's equation for optimal value-based decision-making. To define the value function, assume that the decision maker has accumulated some evidence about the option rewards for some time t. Given this accumulated evidence, the value function returns the total reward the decision maker expects to receive when following the optimal policy. This value includes both the cost for evidence accumulation from time t onwards and the reward resulting from the final choice. The expected rewards, $\hat{r}_j(t) = \langle r_j \mid \delta x(0:t) \rangle$, and elapsed time t are sufficient statistics of the accumulated evidence (see Methods section), such that the value function is defined over these quantities. At each point in time t during evidence accumulation we can either commit to a choice or accumulate more evidence and choose later. When committing to a choice, it is best to choose the option associated with the higher expected reward, such that the total expected reward $V_\mathrm{d}(\hat{r}_1, \hat{r}_2)$ for choosing immediately is given by the value for 'deciding', $V_\mathrm{d}(\hat{r}_1, \hat{r}_2) = \max\{\hat{r}_1, \hat{r}_2\}$ (Fig. 3a). When accumulating more evidence for a small duration δt, in contrast, the decision maker observes additional evidence on which she updates her belief about the true rewards while paying accumulation cost $c\delta t$. At this stage, she expects to receive a total reward of $V(t+\delta t,\ \hat{r}_1(t+\delta t),\ \hat{r}_2(t+\delta t))$. Therefore, the total expected reward for accumulating more evidence is given by the value for 'waiting', $\langle V(t+\delta t,\ \hat{r}_1(t+\delta t),\ \hat{r}_2(t+\delta t)) \rangle - c\delta t$ (Fig. 3b), where the expectation is over the distribution of future expected rewards, $\hat{r}_1(t+\delta t)$ and $\hat{r}_2(t+\delta t)$, given that they are $\hat{r}_1$ and $\hat{r}_2$ at time t (see Methods section for an expression of this distribution). The decision maker ought to only accumulate more evidence if doing so promises more total reward, such that the value function can be written recursively in a form called Bellman's equation (Fig. 3a-c,e; see Supplementary Note 1 for formal derivation),

$$\begin{aligned} V(t, \hat{r}_1, \hat{r}_2) = \max\{&V_\mathrm{d}(\hat{r}_1, \hat{r}_2), \\ &\langle V(t+\delta t, \hat{r}_1(t+\delta t), \hat{r}_2(t+\delta t)) \mid \hat{r}_{1,2}(t) = \hat{r}_{1,2} \rangle - c\,\delta t\}. \end{aligned} \tag{3}$$

With knowledge of the value function, optimal choices are performed as follows. Before having accumulated any evidence, the subjective expected reward associated with option j equals the mean of the prior belief, $\hat{r}_j = \bar{z}_j$, such that the total expected reward at this point is given by $V(0, \hat{r}_1 = \bar{z}_1, \hat{r}_2 = \bar{z}_2)$. Once evidence is accumulated, $\hat{r}_1$ and $\hat{r}_2$ evolve over time, reflecting the accumulated evidence and associated updated belief of the true reward of the choice options. It remains advantageous to accumulate evidence as long as the total expected reward for doing so is larger than that for deciding immediately. As soon as deciding and waiting become equally valuable, that is, $V_\mathrm{d}(\hat{r}_1, \hat{r}_2) = \langle V(t+\delta t,\ \hat{r}_1(t+\delta t),\ \hat{r}_2(t+\delta t)) \rangle - c\delta t$, it is best to choose option j associated with the higher rewarded expected rewarded $\hat{r}_j$. This optimal policy results in two decision boundaries in $(\hat{r}_1, \hat{r}_2)$-space that might change with time (Fig. 3f). In-between these boundaries it remains advantageous to accumulate more evidence, but as soon as either boundary is reached, the associated option ought to be chosen.

Parallel optimal decision boundaries. For the task setup considered above, the decision boundaries take a surprisingly simple shape. When plotted in the $(\hat{r}_1, \hat{r}_2)$-space of estimated option rewards for some fixed time t, the two boundaries are always parallel to the diagonal $\hat{r}_1 = \hat{r}_2$ (Fig. 3f). Furthermore, they are always above and below this diagonal, reflecting that the diagonal separates the regions in which the choice of either option promises more reward. Here, we provide an informal argument why this is the case.

The argument relies on the fact that, for each time t, the decision boundaries are determined by the intersection between the value for deciding and that for waiting (Fig. 3c,d). Both of these values share the property that, in lines parallel to the diagonal, they are linearly increasing with slope one. Formally, both functions satisfy $f(t, \hat{r}_1 + C, \hat{r}_2 + C) = f(t, \hat{r}_1, \hat{r}_2) + C$ for any fixed time t, reward estimates $\hat{r}_1$ and $\hat{r}_2$, and arbitrary scalar C. This implies that, if they intersect at some point $\left(\hat{r}_1^*, \hat{r}_2^*\right)$, thus forming part of the decision boundary, they will intersect at the whole line $\left(\hat{r}_1^* + C, \hat{r}_2^* + C\right)$ that is parallel to the diagonal (Fig. 3c,e,f). Therefore both decision boundaries are parallel to the diagonal.

How can we guarantee that the values for both deciding and waiting are linearly increasing in lines parallel to the diagonal? For the value for deciding, $V_\mathrm{d}(\hat{r}_1, \hat{r}_2) = \max\{\hat{r}_1, \hat{r}_2\}$, this is immediately obvious from its definition (Fig. 3a and caption). Showing the same for the value for waiting requires more work, and is done by a backwards induction argument in time (see Methods section for details). Intuitively, after having accumulated evidence about reward for a long time ($t \to \infty$), the decision maker expects to gain little further insight by any additional evidence. Therefore, deciding is better than waiting, such that the value function will be that for deciding, $V(t, \hat{r}_1, \hat{r}_2) = V_\mathrm{d}(\hat{r}_1, \hat{r}_2)$, which, as previously mentioned, is linearly increasing in lines parallel to the diagonal, providing the base case. Next, it can be shown that, if the value function at time $t + \delta t$ is linearly increasing in lines parallel to the diagonal, then so is the value of waiting at time t, and, as a consequence, also the value function at

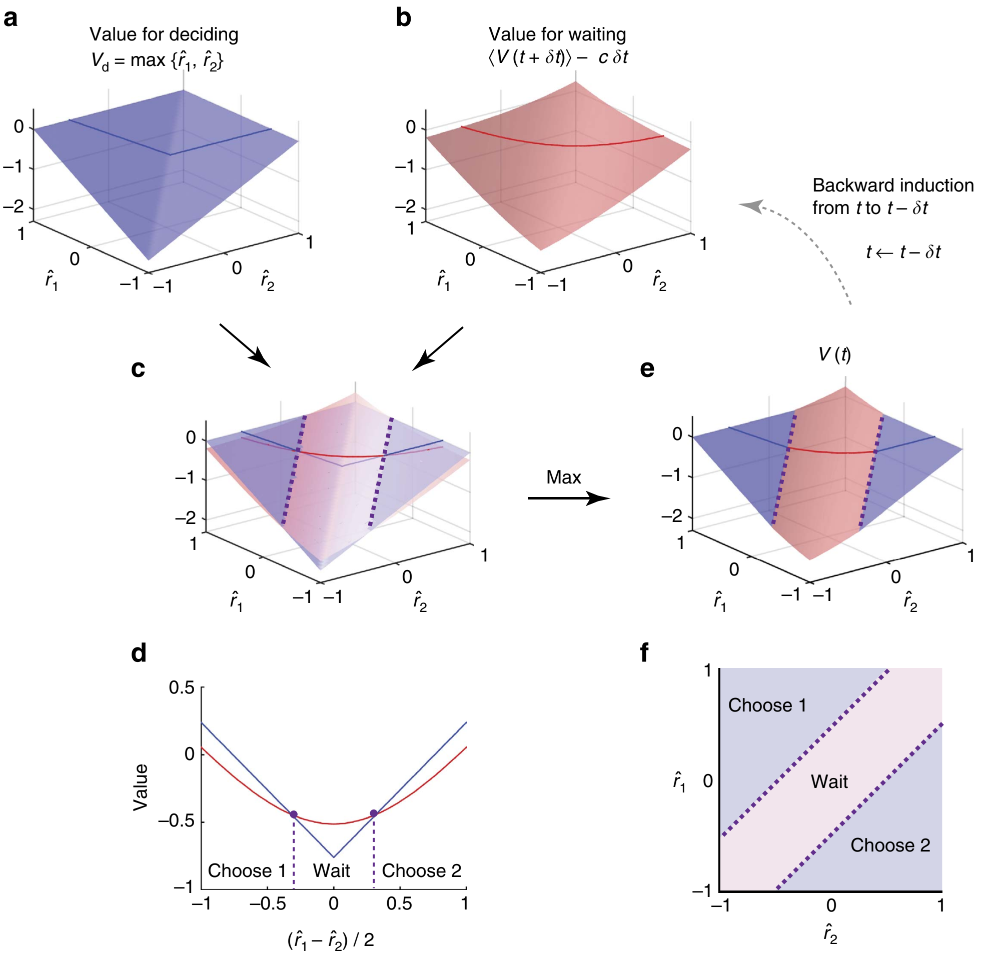

Figure 3 | Finding the decision boundaries in value-based decision. (**a**) The expected values of choosing either option are defined as a two-dimensional function (surface), $\max\{\hat{r}_1,\hat{r}_2\}$, of a pair of reward estimates, $(\hat{r}_1,\hat{r}_2)$, at time t. The dark coloured line shows the section at $\hat{r}_1+\hat{r}_2=0.5$. (**b**) Similarly, the value surface for 'waiting' (that is, the expected value after observing new evidence for a short period δt, subtracted cost for waiting $c\delta t$) is defined as a function of $(\hat{r}_1,\hat{r}_2)$. Note that, around the diagonal, $\hat{r}_1=\hat{r}_2$, the value for waiting is smoother than that for choosing due to the uncertainty about future evidence. (**c**,**d**) The value surfaces for choosing and waiting superimposed, and their sections at $\hat{r}_1+\hat{r}_2=0.5$. The decision boundaries (dotted lines) are determined by points in the space of reward estimates in which the value for 'deciding' (blue) equals that for waiting (red). In the region where waiting has a higher value than choosing either option (blue below red curve/surface), the decision maker postpones the decision to accumulate more evidence; otherwise, she chooses the option that is expected to give the higher reward. Because the relationship between the two value surfaces is translational symmetric in terms of mean reward $\frac{\hat{r}_1+\hat{r}_2}{2}$, their intersections are parallel and do not depend on this mean reward. (**e**) The expected value $V(t)$ is given by the maximum of the values for choosing and waiting. This surface determines the value for waiting (**b**) at the next-earlier time step, $t-\delta t$. (**f**) Decision boundaries and associated choices shown in the two-dimensional $(\hat{r}_1,\hat{r}_2)$representation. Note that the two boundaries are always parallel to the diagonal, $\hat{r}_1=\hat{r}_2$. This is because the both value functions (for deciding and for waiting) are linearly increasing with slope one in lines parallel to the diagonal (**a**,**b**). For the value for deciding, for example, below the diagonal we have $\hat{r}_1>\hat{r}_2$, such that $V_d(\hat{r}_1,\hat{r}_2)=\hat{r}_1$, and therefore $V_d(\hat{r}_1+C,\hat{r}_2+C)=\hat{r}_1+C=V_d(\hat{r}_1,\hat{r}_2)+C$, where C is an arbitrary scalar. The value for waiting can be shown to have the same property.

time t—essentially because the uncertainty about how the reward estimate evolves over time is shift-invariant (does not depend on current expected rewards, $(\hat{r}_1,\hat{r}_2)$; see Methods section). The value function at time t is the maximum over the value for deciding and that for waiting. As both increase linearly in lines parallel to the diagonal, so does this value function, $V(t,\hat{r}_1,\hat{r}_2)$ (Fig. 3c,e). This completes the inductive step.

To summarize, an induction argument backward in time shows that both the values for deciding and waiting increase linearly in lines parallel to the diagonal for all t. As a consequence, the decision boundaries, which lie on the intersection between these two values, are parallel to this diagonal for all times t. In Supplementary Methods, we demonstrate the same property with an argument that does not rely on induction. In both cases, the argument requires, for any fixed t, a stochastic temporal evolution of our expected reward estimates that is shift-invariant with respect to our current estimates $(\hat{r}_1,\hat{r}_2)$. In other words, for any estimates $(\hat{r}_1,\hat{r}_2)$, the decision maker expects them to evolve in exactly the same way. This property holds for the task setup described above and some generalizations thereof (Supplementary Note 1), but might be violated under certain, more complex scenarios, as described further below.

Optimal decisions with collapsing boundaries, and by diffusion models. A consequence of parallel decision boundaries is that optimal choices can be performed by tracking only the difference in expected option rewards, $\hat{r}_1-\hat{r}_2$, rather than both $\hat{r}_1$ and $\hat{r}_2$ independently. To see this, consider rotating these boundaries in $(\hat{r}_1,\hat{r}_2)$-space by $-45°$ such that they come to be parallel to the horizontal axis in the new $(\hat{r}_1+\hat{r}_2,\hat{r}_2-\hat{r}_1)$-space (Fig. 4a,b). After the rotation they bound $\hat{r}_1-\hat{r}_2$ and are independent of $\hat{r}_1+\hat{r}_2$.

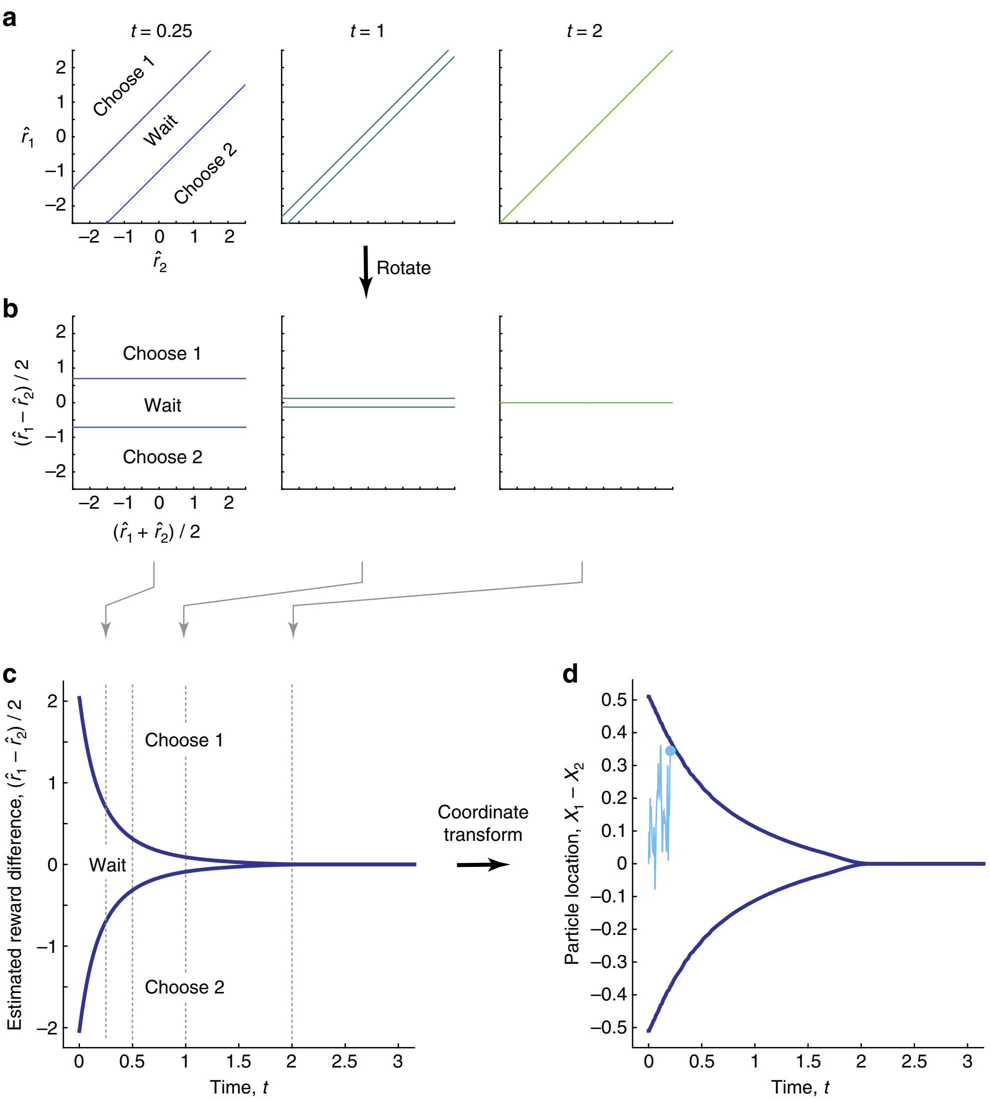

Figure 4 | Optimal decision boundaries computed numerically by DP. (**a**) The decision boundaries and choices at three representative time points ($t = 0.5$, 1 and 2 s), shown in the space of reward estimates $(\hat{r}_1, \hat{r}_2)$. (**b**) The same as **a**, but shown in a rotated space, $(\frac{\hat{r}_1+\hat{r}_2}{2}, \frac{\hat{r}_1-\hat{r}_2}{2})$. (**c**) The decision boundaries in terms of the expected reward difference, $\frac{\hat{r}_1-\hat{r}_2}{2}$, as functions of time. The distance between boundaries decreases as time elapses, and collapses to zero at some point. (**d**) The same decision boundaries shown in the space of accumulated evidence, which is represented by the particle location in a DDM. The cyan trace is a sample particle trajectory, representing evidence accumulation and the subsequent choice of option 1.

For Gaussian *a priori* rewards (Fig. 2a), numerical solutions reveal that the distance between the two boundaries decreases over time, resulting in 'collapsing boundaries' (Fig. 4c) that can be explained as follows. In the beginning of the decision, the true option rewards are highly uncertain due to a lack of information. Hence, every small piece of additional evidence will make the running reward estimates substantially more certain. This makes it worth to withhold decisions by far-separated decision boundaries (Fig. 4c for small t). Once a significant amount of evidence is accumulated, further evidence will barely increase certainty about the true rewards. Thus, it becomes more preferable to decide quickly rather than to withhold choice for an insignificant increase in choice accuracy (even for similar reward estimates, $\hat{r}_2 - \hat{r}_1 \approx 0$, and residual uncertainty about which option yields the higher reward). The narrowing boundary separation ensures such rapid decisions (Fig. 4c for large t).

We can further simplify the optimal decision procedure by implementing the computation of the expected option reward difference by a diffusion model. As long as $\bar{z}_1 = \bar{z}_2$, such an implementation remains statistically optimal, as the diffusing particle, $x(t) \equiv x_1(t) - x_2(t)$, (recall that $x_j(t = n\delta t) = \sum_{i=1}^{n} \delta x_{j,i}$) and elapsed time t form a set of sufficient statistics of the posterior $r_1(t) - r_2(t) | \delta \boldsymbol{x}(0{:}t)$ over this difference (see Methods section). Furthermore, $x_j(t)$ can be interpreted as the sample path of a particle that diffuses with variance σ^2 and drifts with rate z_j. For this reason, $x(t)$ diffuses with variance $2\sigma^2$ and drifts with rate $z_1 - z_2$, thus forming the particle in a diffusion model that performs statistically optimal inference. The same mapping between expected reward difference and diffusing particle allows us to map the optimal boundary on reward into boundaries on $x(t)$ (Fig. 4c,d). Therefore, models as simple as diffusion models can implement optimal value-based decision-making.

Moving from single choices to a sequence thereof. So far we have focused on single choices in which the decision maker trades off the expected reward received for this choice with the cost associated with accumulating evidence about the true option

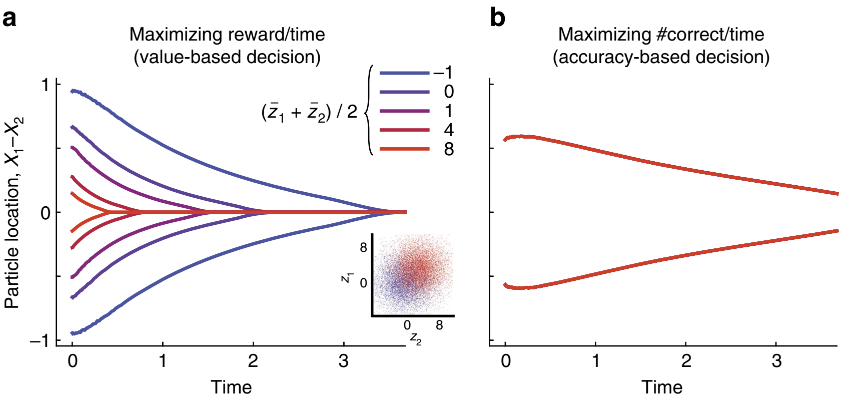

Figure 5 | Effects of prior and task demands on the speed of boundary collapse. (**a**) The decision boundaries for value-based decisions that maximizes the reward per unit time. The average *a priori* reward $\frac{\bar{z}_1+\bar{z}_2}{2}$ is varied from -1 to 8, while keeping $\bar{z}_1 = \bar{z}_2$. The inset shows two examples of prior distributions (blue and red: mean reward 0 and 4, respectively). The figure illustrates that the optimal decision boundaries depend on the prior knowledge about the average reward across trials. (**b**) The decision boundaries that maximizes the number of correct response per unit time for accuracy-based decisions. (The 'correct rate' in value-based decisions can be defined, for example, as the probability of choosing the more preferable option.) These boundaries do not vary with $\frac{\bar{z}_1+\bar{z}_2}{2}$ are all plotted on top of each other. Here, the decision boundaries were derived with the same dynamic-programming procedure as for the value-based case, except for that the rewards were assumed to be binary, and only one if the decision maker correctly identified the option with the larger 'reward' cue z_j (see Methods section). In contrast to the reward-rate maximization strategy for value-based decisions (**a**), the decision strategy maximizing the correct rate is invariant to the absolute values of mean reward/evidence strength, thus demonstrating a qualitative difference between value-based and perceptual decision-making in terms of the optimal strategy. In addition, the optimal boundaries in the value-based case approach each other more rapidly over time than for perceptual decisions. The faster boundary collapse for value-based decisions is consistent across a broad range of mean absolute rewards, showing that the distinction in boundary dynamics is not just due to the difference in expected reward rates, but reflecting a qualitative difference between the geometries of value functions in these two tasks.

rewards. This setup assumes a single choice and, besides the accumulation cost, infinite time to perform it. In realistic scenarios, however, such choices are usually embedded within a sequence of similar choices. Here, we consider how such embedding influences the form of the optimal policy.

Maximizing the reward rate across choices. We assume that each choice within the sequence follow the previous single-choice setup. That is, after onset of the choice options, the decision maker pays a cost c per second for accumulating evidence about the true option rewards. At choice, she receives the true reward associated with the chosen option. The choice is followed by a (possibly stochastic) waiting time of t_w seconds on average, after which two new choice options appear and new evidence is accumulated. The true reward associated with either option is before choice option onset drawn according to the previously described Gaussian prior (Fig. 2a), such that these rewards remain constant within individual choices, but vary across consecutive choices. Rather than maximizing the total expected reward for each individual choice, we assume that the aim is to maximize the total expected reward within a fixed time period, independent of how many choices are performed within this period. To avoid boundary effects, we assume the period duration to be close-to-infinite, such that maximizing the total expected reward within this period becomes equivalent to maximizing the reward rate ρ, given by

$$\rho = \frac{\langle r_j \mid \delta x_j(0:T) \rangle - c\langle T \rangle}{t_w + \langle T \rangle}, \qquad (4)$$

where the expectation is, as for equation (2), across choices j and evidence accumulation times T, given the flow of evidence. Here, it is critical that we fix the time period while leaving open the number of choices that can be performed. If we instead were to fix the number of choices while leaving open the time to make them, it again becomes optimal to maximize the total expected reward for each of these choices separately, such that the optimal policy for each such choice is the same as that for single, isolated choices.

Infinite choice sequences make using the standard value function difficult. This value function returns the total expected reward for all current and future choices when starting from the current state. For an infinite number of such future choices, the value function might thus become infinite. One way to avoid this is to use instead the 'average-adjusted value' function, which—in addition to an accumulation cost—penalizes the passage of some time duration δt by $-\rho\delta t$, where ρ is the reward rate. This reward rate is by equation (4) the total reward received (including accumulation costs) per second, averaged over the whole choice sequence. Penalizing the value function by this reward rate makes explicit the implicit loss of rewards due to potential future choices that the decision maker misses out on when accumulating too much evidence for the current choice. This penalization allows us to treat all choices in the sequence as if they were the same, unique choice. A further consequence of this penalization is that the value function for accumulating more evidence for some duration δt undergoes a more significant change, as accumulating this evidence now comes at a cost $-(c+\rho)\delta t$ instead of the previous $-c\delta t$ (see Methods section for the associated Bellman equation). For positive reward rates, $\rho > 0$, this cost augmentation implies more costly evidence accumulation such that it becomes advantageous to accumulate less evidence than for single, isolated choices. This change is implemented by decision boundaries that collapse more rapidly (shown formally in Supplementary Note 1, see also Supplementary Fig. 1). Thus, collapsing decision boundaries implement the optimal policy for both single choices and sequences of choices, with the only difference that these boundaries collapse more rapidly for the latter. The duration of inter-choice waiting t_w modulates this difference, as with $t_w \rightarrow \infty$, the reward rate described by equation (4) reduces to the expected reward for single, isolated choices, equation (2). Therefore the policy for single trials is a special case of that for maximizing the reward rate in which the waiting time between consecutive choices becomes close-to-infinite.

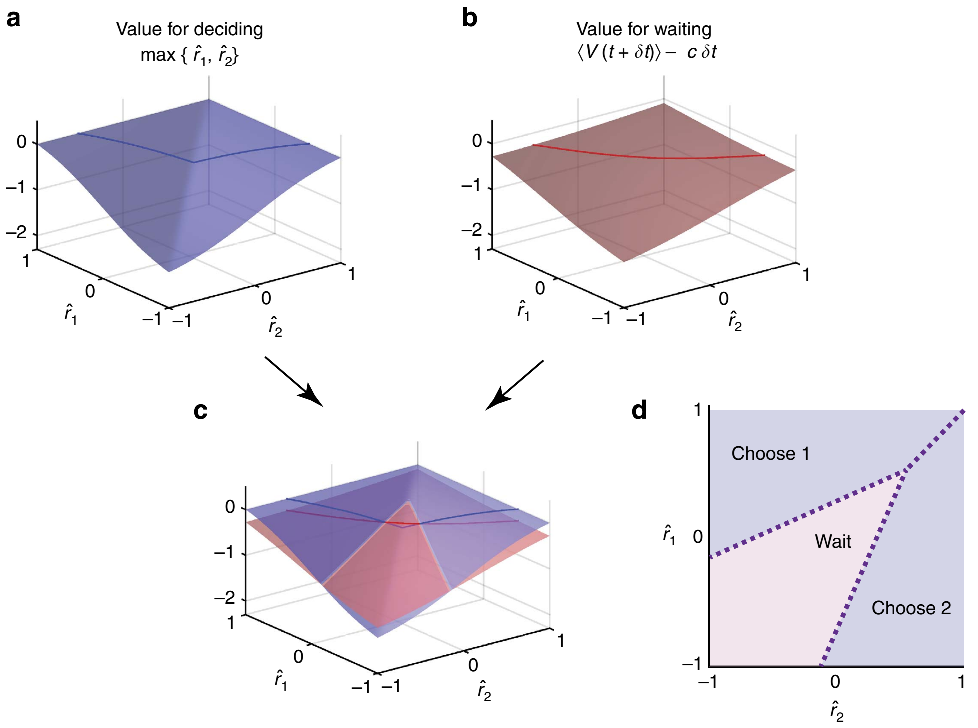

Figure 6 | In some scenarios the optimal policy becomes even more complex than two parallel boundaries in the space of expected reward estimates. This property might, for example, break down if the utility that the decision maker receives from her choices is not the reward itself but instead a non-linear function of this reward. If this utility grows sub-linearly in the reward, as is frequently assumed, the decision boundaries approach each other with increasing expected reward, as higher rewards yield comparably less utility. In such circumstances, optimal choices require tracking of both expected reward estimates, $\hat{r}_1$ and $\hat{r}_2$, independently rather than only their difference. To demonstrate this, here we assumed a saturating utility function, $Utility = u(r)$, which saturates at $r \to \infty$ and $r \to -\infty$. This could be the case, for example, if rewards vary over a large range over which the subjectively perceived utility follows a non-linear saturating function of this reward. (In this figure, u is modelled with a tangent hyperbolic function, but the exact details of the functional form do not qualitatively change the results). The logic of the different panels follows that of Fig. 2. (**a**) The value function surface for choosing either of two options. (**b**) The value surfaces for postponing decision to accumulate more evidence for a period of δt. (**c**) The two value surfaces superimposed. (**d**) The decision boundary and choice represented in the two-dimensional space of $(\hat{r}_1, \hat{r}_2)$. Note that the distance between decision boundaries is narrower in the regime where estimated rewards are high on average, resembling 'RMs'[11,12], which are more sensitive to absolute reward magnitudes than DDMs.

Dependency of the policy on the prior distribution of reward. As shown above, optimal value-based decisions are achieved by accumulating only the difference of reward estimates, as implementable by DDMs. However, this does not mean that the absolute reward magnitudes have no effect on the decision strategy; they affect the decision boundary shape. Figure 5a shows how the optimal decision boundaries depend on the mean of the *a priori* belief about the true rewards across trials. When both options are likely to be highly rewarding on average, the boundaries should collapse more rapidly to perform more choices within the same amount of time. In the light of a guaranteed high reward, this faster collapse promotes saving time and effort of evidence accumulation. The boundary shape does not change for trial-by-trial variations in true rewards (which are *a priori* unknown) for the same prior, but only when the prior itself changes. This sensitivity to the prior and associated average rewards also differentiates reward rate-maximizing value-based decision-making from decisions that aim at maximizing the reward for single, isolated choices (Supplementary Note 1), and from classic paradigms of perceptual decision-making (Fig. 5b, see also Discussion section). To summarize, for value-based decisions that maximize the reward rate, the *a priori* belief about average-reward magnitudes affect the strategy (and, as a consequence, the average reaction time) by modulating the speed of collapse of the decision boundaries, even if choices within individual decisions are only guided by the relative reward estimates between options.

The limits of diffusion models for value-based decisions. For all scenarios we have considered so far, diffusion models can implement the optimal decision-making policy. Here, we discuss that this is still the case for some, but not all generalizations of the task. For some tasks, the optimal policy won't even be representable by parallel boundaries in the $(\hat{r}_1, \hat{r}_2)$-space of expected reward estimates. This is, for example, the case when the prior/likelihood distributions of reward/evidence are correlated in a particular way (see Methods section and Supplementary Note 1), or when the utility function is non-linear (see Fig. 6 for an example).

Thus, diffusion models only seem to implement the optimal decision strategy under very constrained circumstances. However, even beyond these circumstances, diffusion models might not be too far off from achieving close-to-optimal performance, but their loss of reward remains to be evaluated in general circumstances. Laboratory experiments could satisfy conditions for diffusion models to be close-to-optimal even in the presence of a non-linear utility function. Such experiments often use moderate rewards (for example, moderately valued food items, rather than extreme payoffs) in which case a potentially non-linear utility would be well-approximated by a linear function within the tested range of rewards.

Discussion

We have theoretically derived the optimal behaviour for value-based decision-making with noisy evidence about rewards. Our analysis revealed that the optimal strategy in a natural problem setup (where values are linear in rewards) reduces to a DDM with time-varying boundaries. This result provides a theoretical basis for why human decision makers seem to feature behaviour in such tasks that, just as in accuracy-based (conventional perceptual) decisions, is well captured by DDMs—despite the profound qualitative difference in task structures (for example, a two-dimensional value functions for value-based tasks, but not for accuracy-based ones). Furthermore, we found that the optimal strategy does not always reduce to DDMs if we assume non-linear relationships between value and reward (Fig. 6), predicting that human behaviour may deviates from DDMs in specific experimental conditions (perceived utility following a non-linear saturating function of this reward; Fig. 6d); interestingly, such decision boundary structure might be better approximated by 'correlated RMs'[11,12].

Simultaneous to our work, another theoretical study by Fudenberg *et al.* (unpublished work[22]) has recently focused on optimal evidence accumulation and decision-making for value-based decisions. This study provides a more in-depth mathematical characterization of the optimal policy implemented by diffusion model with collapsing boundaries. Their analysis, however, is restricted to single, isolated choices, and—unlike us—does not consider policy changes for reward rate maximization, nor non-linear utility functions that invalidate the use of diffusion models.

Whether human and animal use collapsing decision boundaries is a topic of debate in the recent accuracy-based[23] and value-based[9] decision-making studies. Interestingly, a recent meta-analysis study reports that whether subject uses collapsing boundaries varies strongly across tasks and individuals[23]. Our theory suggests that the optimal boundary dynamics is sensitive to task demands (for example, reward-rate maximization or correct-rate maximization) as well as the absolute mean reward magnitude (in contrast to perceptual decision-making; see Supplementary Note 2). Thus, subjects might switch their decision strategies depending on those experimental factors, emphasizing the need to carefully control these factors in further studies.

Still, in both daily lives and laboratory experiments, humans can sometimes take a long time to decide between two valuable options, which might reflect suboptimal behaviour or an insufficiently fast collapse of the bound. For instance, a recent empirical study by Oud *et al.*[24] reports slower-than-optimal value-based and perceptual choices of human decision makers in a reward rate maximization setting. These slow choices might arise, however, from incompletely or incorrectly learned priors (Supplementary Note 3), and warrant further investigation. Another slowing factor is insufficient time pressure induced by, for example, fixing the number of choices instead of the total duration of the experiment. In this case, the slow reaction times may not reflect a suboptimal strategy. For example, Milosavljevic *et al.*[9] have found that subjects can take a surprisingly long time to decide between two high-valued items but, in this experiment, subjects had to perform a fixed number of choices without any time constraint. Their reward at the end of the experiment was determined by drawing one item among all the items selected by the subject[9]. With such a task design, there is no explicit incentive for making fast choices and, therefore, the optimal strategy does allow for long reaction times. All of the above cases highlight that the seeming irrationality of slow choices between two high-valued options might in fact reflect a completely rational strategy under contrived laboratory settings. Thus, by revealing the optimal policy for value-based decisions, the present theory provides a critical step in studying the factors that determine our decisions about values.

What do collapsing boundaries in diffusion models tell us about the neural mechanisms involved in such decisions? Previous studies concerning perceptual decisions have linked such boundary collapse to a neural 'urgency signal' that collectively drives neural activity towards a constant threshold[7,25]. However, note that in such a setup even a constant (that is, non-collapsing) diffusion model bound realizes a collapsing bound in the decision maker's posterior belief[7]. Analogously, a constant diffusion model bound in our setup realizes a collapsing bound on the value estimate difference. Furthermore, how accumulated evidence is exactly coded in the activity of individual neurons or neural populations remains unclear (for example, compare refs 6,26), and even less is known about value encoding. For these reasons we promote diffusion models for behavioural predictions, but for now refrain from directly predicting neural activity and associated mechanisms. Nonetheless, our theory postulates what kind of information ought to be encoded in neural populations, and as such can guide further empirical research in neural value coding.

Methods

Structure of evidence and evidence accumulation. Here, we assume a slightly more general version of the task than the one we discuss throughout most of the main text, with a correlated prior and a correlated likelihood. Further below we describe how this version relates to the one in the main text. In particular, we assume the prior over true rewards, given by vector $\boldsymbol{z} \equiv (z_1, z_2)^T$, to be a bivariate Gaussian, $\boldsymbol{z} \sim \mathcal{N}(\bar{\boldsymbol{z}}, \boldsymbol{\Sigma}_z)$, with mean $\bar{\boldsymbol{z}}$ and covariance $\boldsymbol{\Sigma}_z$. In each small time step i of duration δt, the decision maker observes some momentary evidence $\delta \boldsymbol{x}_i \equiv \left(\delta x_{1,i}, \delta x_{2,i}\right)^T \sim \mathcal{N}(\boldsymbol{z}\delta t, \boldsymbol{\Sigma}\delta t)$ that informs her about these true rewards. After accumulating evidence for some time $t = n\delta t$, her posterior belief about the true rewards is found by Bayes' rule, $p(\boldsymbol{z} \mid \delta\boldsymbol{x}(0:t)) \propto p(\boldsymbol{z}) \prod_{i=1}^{n} p(\delta\boldsymbol{x}_i \mid \boldsymbol{z})$, and results in

$$\boldsymbol{z} \mid \delta\boldsymbol{x}(0:t) \sim \mathcal{N}\left(\boldsymbol{\Sigma}(t)\left(\boldsymbol{\Sigma}_z^{-1}\bar{\boldsymbol{z}} + \boldsymbol{\Sigma}^{-1}\boldsymbol{x}(t)\right), \boldsymbol{\Sigma}(t)\right),$$

where we have defined $\boldsymbol{x}(t) = \sum_{i=1}^{n} \delta\boldsymbol{x}_i$ as the sum of all momentary evidence up to time t, and $\boldsymbol{\Sigma}(t) = \left(\boldsymbol{\Sigma}_z^{-1} + t\boldsymbol{\Sigma}^{-1}\right)^{-1}$ as the posterior covariance (hereafter, when $\boldsymbol{\Sigma}(t)$ is a function of time it denote the posterior covariance, rather than the covariance of evidence, $\boldsymbol{\Sigma}$). For the case that experienced reward $\boldsymbol{r} \equiv (r_1, r_2)^T$ equals true reward $\boldsymbol{z}$, that is $\boldsymbol{r} = \boldsymbol{z}$, the mean estimated option reward $\hat{\boldsymbol{r}}(t) = \langle \boldsymbol{z} \mid \delta\boldsymbol{x}(0:t)\rangle$ is the mean of the above posterior.

Expected future reward estimates. Finding the optimal policy by solving Bellman's equation requires computing the distribution of expected future rewards $\hat{\boldsymbol{r}}(t+\delta t)$ given the current expected rewards $\hat{\boldsymbol{r}}(t)$. Assuming a small δt such that the probability of an eventual boundary crossing becomes negligible, we can find this distribution by the marginalization

$$p(\hat{\boldsymbol{r}}(t+\delta t) \mid \hat{\boldsymbol{r}}(t)) = \int p(\hat{\boldsymbol{r}}(t+\delta t) \mid \delta\boldsymbol{x}(t+\delta t), \hat{\boldsymbol{r}}(t)) p(\delta\boldsymbol{x}(t+\delta t) \mid \hat{\boldsymbol{r}}(t)) \mathrm{d}\delta\boldsymbol{x}(t+\delta t).$$

As $\hat{\boldsymbol{r}}(t+\delta t)$ is the mean of the posterior of $\boldsymbol{z}$ after having accumulated evidence up to time $t+\delta t$, it is given by

$$\hat{r}(t+\delta t) = \boldsymbol{\Sigma}(t+\delta t)\boldsymbol{\Sigma}(t)^{-1}\hat{\boldsymbol{r}}(t) + \boldsymbol{\Sigma}(t+\delta t)\boldsymbol{\Sigma}^{-1}\delta\boldsymbol{x}(t+\delta t),$$

where we have used $\boldsymbol{x}(t+\delta t) = \boldsymbol{x}(t) + \delta\boldsymbol{x}(t+\delta t)$ and $\boldsymbol{x}(t) = \boldsymbol{\Sigma}\left(\boldsymbol{\Sigma}(t)^{-1}\hat{\boldsymbol{r}}(t) - \boldsymbol{\Sigma}_z^{-1} \mid \boldsymbol{z}\right)$, following from the definition of $\hat{\boldsymbol{r}}(t)$. Furthermore, by the generative model for the momentary evidence we have $\delta\boldsymbol{x}(t+\delta t) \mid \boldsymbol{z} \sim \mathcal{N}(\boldsymbol{z}\delta t, \boldsymbol{\Sigma}\delta t)$, and our current posterior is $\boldsymbol{z} \mid \hat{\boldsymbol{r}}(t) \sim \mathcal{N}(\hat{\boldsymbol{r}}(t), \boldsymbol{\Sigma}(t))$, which, together, gives $\delta\boldsymbol{x}(t+\delta t) \mid \hat{\boldsymbol{r}}(t) \sim \mathcal{N}(\hat{\boldsymbol{r}}(t)\delta t, \boldsymbol{\Sigma}\delta t + \boldsymbol{\Sigma}(t)\delta t^2)$. With these components, the marginalization results in

$$\hat{\boldsymbol{r}}(t+\delta t) \mid \hat{\boldsymbol{r}}(t) \sim \mathcal{N}\left(\hat{\boldsymbol{r}}(t), \boldsymbol{\Sigma}(t+\delta t)\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}(t+\delta t)\delta t\right),$$

where we have only kept terms of order δt or lower. An extended version of this derivation is given in Supplementary Note 1.

More specific task setups. Here, we consider two more specific task setups. In the first one, the prior covariance is proportional to the likelihood covariance, that is $\boldsymbol{\Sigma}_z = \alpha\boldsymbol{\Sigma}$. This causes the posterior $\boldsymbol{z}$ to be given by

$$\boldsymbol{z} \mid \delta\boldsymbol{x}(0:t) \sim \mathcal{N}\left(\frac{\alpha^{-1}}{\alpha^{-1}+t}\bar{\boldsymbol{z}} + \frac{t}{\alpha^{-1}+t}\frac{\boldsymbol{x}(t)}{t}, \frac{1}{\alpha^{-1}+t}\boldsymbol{\Sigma}\right).$$

In this case, the posterior mean becomes independent of the covariance, and is a weighted mixture of prior and accumulated evidence. The distribution over expected future reward estimates becomes $\hat{\boldsymbol{r}}(t+\delta t) \mid \hat{\boldsymbol{r}}(t) \sim \mathcal{N}(\hat{\boldsymbol{r}}(t), (\alpha^{-1}+t)^{-2}\boldsymbol{\Sigma})$. In terms of choosing among lunch menus, a positively correlated prior could correspond to differently skilled cooks working on different days, such that the true rewards associated with the different options fluctuate jointly. A correlated likelihood might correspond to fresh produce in one menu option predicting the same in the other menu option. If the likelihood covariance is proportional to that of the prior, diffusion models still implement the optimal choice policy.

In the second more specific setup we assume both prior and likelihood to be uncorrelated, with covariance matrices given by $\boldsymbol{\Sigma}_z = \sigma_z^2 \boldsymbol{I}$ and $\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{I}$. This is the setup discussed throughout most of the work, and results in an equally uncorrelated posterior $\boldsymbol{z}$, that is for option j given by equation (1). The distribution over expected future reward estimates is also uncorrelated, and for option j is given by $\hat{r}_j(t+\delta t) \mid \hat{r}_j(t) \sim \mathcal{N}\left(\hat{r}_j(t), \sigma^{-2}\left(\sigma_z^{-2} + t\sigma^{-2}\right)^{-2}\right)$.

A more general scenario than the ones we have discussed so far is that both the decision-maker's *a priori* belief about the true rewards, as well as the likelihood of the momentary evidence about these rewards are correlated, but the prior covariance is not proportional to the likelihood covariance. Once prior covariance and likelihood covariance are not proportional to each other anymore, diffusion models fail to implement the optimal policy. Even then, the optimal policy in the $(\hat{r}_1, \hat{r}_2)$-space of expected reward estimates is still given by two boundaries parallel to the identity line, such that we can again only bound the difference between these estimates. However, these are bounds on expected reward estimate differences, and not on a diffusing particle. Mapping the estimates into a single diffusing particle requires combining them linearly with combination weights that change over time, which is incompatible with the standard diffusion model architecture (although it can be implemented by an extended diffusion model as shown in (ref. 27). Thus, parallel decision boundaries on expected reward estimates do not automatically imply that diffusion models can implement optimal decisions.

Evidence accumulation and decisions with diffusion models. For the class of tasks in which the decision boundaries in $(\hat{r}_1, \hat{r}_2)$ are parallel to the diagonal, the optimal policy can be represented by two boundaries, $\xi_1(t)$ and $\xi_2(t)$, on the expected reward difference $\Delta\hat{r}(t) = \hat{r}_1(t) - \hat{r}_2(t)$, such that evidence is accumulated as long as $\xi_2(t) < \Delta\hat{r}(t) < \xi_1(t)$, and option 1 (option 2) is chosen as soon as $\Delta\hat{r}(t) \geq \xi_1(t)$ $(\Delta\hat{r}(t) \leq \xi_2(t))$. To implement this policy with diffusion models, we need to find a possibly time-dependent function $x(t) = f(\Delta\hat{r}(t), t)$ that maps the expected reward difference $\Delta\hat{r}(t)$ into a drifting/diffusing particle $\mathrm{d}x(t) = \mu \mathrm{d}t + \sigma_x \mathrm{d}W_t$, that drifts with drift μ and diffuses with variance σ_x^2, and where $\mathrm{d}W_t$ is a Wiener process. Such a mapping allows us to find the boundaries $\theta_{j\in\{1,2\}}(t) = f\left(\xi_j(t), t\right)$, that implement the same policy by bounding particle $x(t)$.

For the general case of a correlated prior and correlated likelihood, as discussed further above, we have $\hat{\boldsymbol{r}}(t) = \boldsymbol{\Sigma}(t)\boldsymbol{\Sigma}_z^{-1}\bar{\boldsymbol{z}} + \boldsymbol{\Sigma}(t)\boldsymbol{\Sigma}^{-1}\boldsymbol{x}$, where $\boldsymbol{x}(t)$ drifts and diffuses according to $\boldsymbol{x}(t) \sim \mathcal{N}(\boldsymbol{z}t, \boldsymbol{\Sigma}t)$. Using $\boldsymbol{x}(t) = \boldsymbol{z}t + \sqrt{t}\boldsymbol{\Gamma}\boldsymbol{\eta}$, where $\boldsymbol{\eta} = (\eta_1, \eta_2)^T \sim \mathcal{N}(0, \boldsymbol{I})$, and a $\boldsymbol{\Gamma}$ that satisfies $\boldsymbol{\Gamma}\boldsymbol{\Gamma}^T = \boldsymbol{\Sigma}$, we find

$$\Delta\hat{r}(t) = (a_1(t) - a_2(t))t + \sqrt{t}((b_{11}(t) - b_{21}(t))\eta_1 + (b_{12}(t) - b_{22}(t))\eta_2),$$

with $a_j(t)$ denoting the elements of vector $\boldsymbol{a}(t) = t^{-1}\boldsymbol{\Sigma}(t)\boldsymbol{\Sigma}_z^{-1}\bar{z} + \boldsymbol{\Sigma}(t)\boldsymbol{\Sigma}^{-1}\boldsymbol{z}$ and $b_{ij}(t)$ being the elements of matrix $\boldsymbol{B}(t) = \boldsymbol{\Sigma}(t)\boldsymbol{\Sigma}^{-1}\boldsymbol{\Gamma}$. The above describes a diffusion process with drift and diffusion that vary over time in different ways. Therefore, we cannot find a function $f(\cdot, t)$ that maps $\Delta\hat{r}(t)$ into a diffusing particle with constant drift and diffusion. As a result, we cannot use diffusion models for optimal decision-making in this case.

One reason for this incompatibility is that the posterior covariance changes from prior covariance, $\boldsymbol{\Sigma}(0) = \boldsymbol{\Sigma}_z$, to likelihood covariance, $\lim_{t\to\infty}\boldsymbol{\Sigma}(t) = t^{-1}\boldsymbol{\Sigma}$, over time and influences the relation between drift and diffusion. If we set the prior covariance proportional to the likelihood covariance, that is $\boldsymbol{\Sigma}_z = \alpha\boldsymbol{\Sigma}$, then we can find a mapping to diffusion models. Using the mean of the posterior $\boldsymbol{z}$ from the previous section, we find that $\hat{\boldsymbol{r}}(t) = (\alpha^{-1} + t)^{-1}\left((t^{-1}\alpha^{-1}\bar{\boldsymbol{z}} + \boldsymbol{z})t + \sqrt{t}\boldsymbol{\Gamma}\boldsymbol{\eta}\right)$, which results in the expected reward difference

$$\Delta\hat{r}(t) = \frac{1}{\alpha^{-1} + t}\left((t^{-1}\alpha^{-1}(\bar{z}_1 - \bar{z}_2) + z_1 - z_2)t + \sqrt{t}((\gamma_{11} - \gamma_{21})\eta_1 + (\gamma_{12} - \gamma_{22})\eta_2)\right),$$

where γ_{ij} are the elements of $\boldsymbol{\Gamma}$. Now, as long as $\bar{z}_1 = \bar{z}_2$ (a priori, both options have the same true reward), we can use the mapping $f(\Delta\hat{r}(t), t) = (\alpha^{-1} + t)\Delta\hat{r}(t)$ to map the boundaries in the diffusion model space, which features a particle that drifts with drift $\mu = z_1 - z_2$ and diffuses with variance $\sigma_x^2 = (\gamma_{11} - \gamma_{21})^2 + (\gamma_{12} - \gamma_{22})^2$.

The setup that is discussed throughout the main text becomes even simpler, with a diagonal prior covariance, $\boldsymbol{\Sigma}_z = \sigma_z^2\boldsymbol{I}$, and a diagonal likelihood covariance $\boldsymbol{\Sigma} = \sigma^2\boldsymbol{I}$. Using the mean of the posterior in equation (1), and again assuming $\bar{z}_1 = \bar{z}_2$, a similar argument as before shows that the mapping $f(\Delta\hat{r}(t), t) = \left(\sigma^2/\sigma_z^2 + t\right)\Delta\hat{r}(t)$ allows us to implement optimal decision-making with a diffusion model with drift $\mu = z_1 - z_2$ and diffusion variance $\sigma_x^2 = 2\sigma^2$.

Bellman's equation for single isolated trials. The rationale behind optimal decision-making in single, isolated trials is explained in the main text and is here repeated only briefly. At each point in time t after onset of the choice options, the decision maker performs the action that promises the largest sum of expected rewards from that point onwards (including the cost for accumulating evidence). Given that at this time the decision maker holds a posterior belief over values with sufficient statistics $(t, \hat{r}_1, \hat{r}_2)$, the sum of expected rewards is denoted by the value function $V(t, \hat{r}_1, \hat{r}_2)$. The available actions are to either choose option one or two, or to accumulate more evidence and decide later. Deciding immediately, the decision maker would choose the option that is expected to yield higher reward, such that the value associated with deciding is $V_\mathrm{d}(\hat{r}_1, \hat{r}_2) = \max\{\hat{r}_1, \hat{r}_2\}$. Accumulating evidence for another time period δt comes at cost $c\delta t$ but is expected to yield reward $V(t + \delta t, \hat{r}_1(t + \delta t), \hat{r}_2(t + \delta t))$. Here, the expectation is over how the sufficient statistics are expected to evolve when accumulating more evidence, and is given by the bivariate Gaussian $\hat{\boldsymbol{r}}(t + \delta t) \mid \hat{\boldsymbol{r}}(t)$ that we have derived further above. Thus, the value for waiting is $\langle V(t + \delta t, \hat{r}_1(t + \delta t), \hat{r}_2(t + \delta t))\rangle - c\delta t$. At any time t, the decision maker chooses the action associated with the higher value, which leads to Bellman's equation, as given by equation (3) in the main text. This equation on one hand defines the value function, and on the other hand determines the optimal policy: as long as the value for waiting dominates, the decision maker ought to accumulate more evidence. Once the value for deciding becomes larger, it is best to choose the option that is expected to yield the higher reward.

Bellman's equation for reward rate maximization. In order to find Bellman's equation and the associated optimal policy that maximizes the reward rate, we borrow concepts from average-reward DP (refs 7,28). We do so to avoid that the value function associated with the first trial becomes infinite if this trial is followed by an infinite number of trials that, in total, promise infinite reward. Average-reward DP penalizes the passage of some time δt by cost $\rho\delta t$, where ρ is the reward rate, equation (4), which equals the average expected reward per unit time. With this additional cost, and the value function turns into the 'average-adjusted value' function $\tilde{V}(t, \hat{r}_1, \hat{r}_2, \rho)$, which is the same for each trial in the sequence, and is defined as follows. Immediate decisions are expected to be rewarded by $\max\{\hat{r}_1, \hat{r}_2\}$, followed by some waiting time t_w that comes at cost ρt_w. After this waiting time, the decision maker holds belief $(t, \hat{r}_1, \hat{r}_2) = (0, \bar{z}_1, \bar{z}_2)$ (recall that $\bar{z}_j$ denotes the prior mean for option j) at the onset of the next trial, and therefore expects reward $\tilde{V}(0, \bar{z}_1, \bar{z}_2, \rho)$ in this trial. Thus, the value for decision immediately is given by $\max\{\hat{r}_1, \hat{r}_2\} - \rho t_w + \tilde{V}(0, \bar{z}_1, \bar{z}_2, \rho)$. The value for accumulating more evidence is the same as for single, isolated trials (see previous section), only that the cost increases from $c\delta t$ to $(c + \rho)\delta t$. Bellman's equation is again given by taking the maximum over all values. In contrast to single, isolated trials, the policy arising from Bellman's equation is invariant to global shifts in the value function. That is, we can add some constant C to the average-adjusted value associated with all sufficient statistics, such that $\tilde{V}(t, \hat{r}_1, \hat{r}_2, \rho) \to \tilde{V}(t, \hat{r}_1, \hat{r}_2, \rho) + C$, and would recover the same policy[28]. As a result, we can arbitrarily fix the average-adjusted value for one such statistic, and all other values follow accordingly. For convenience, we choose $\tilde{V}(0, \bar{z}_1, \bar{z}_2, \rho) = 0$, which results in Bellman's Equation

$$\tilde{V}(t, \hat{r}_1, \hat{r}_2, \rho) = \max\left\{\begin{array}{c} \tilde{V}_\mathrm{d}(\hat{r}_1, \hat{r}_2, \rho), \\ \langle\tilde{V}(t + \delta t, \hat{r}_1(t + \delta t), \hat{r}_2(t + \delta t), \rho) \mid \hat{r}_{1,2}(t) = \hat{r}_{1,2}\rangle - (c + \rho)\delta t \end{array}\right\},$$

where $\tilde{V}_\mathrm{d}(\hat{r}_1, \hat{r}_2, \rho)$ is given by $\tilde{V}_\mathrm{d}(\hat{r}_1, \hat{r}_2, \rho) = \max\{\hat{r}_1, \hat{r}_2\} - \rho t_w$. This also gives us a recipe to find $\rho : \tilde{V}(0, \bar{z}_1, \bar{z}_2, \rho) = 0$ will only hold for the correct ρ, such that we can compute $\tilde{V}(0, \bar{z}_1, \bar{z}_2, \rho)$ for some arbitrary ρ, and then adjust ρ until $\tilde{V}(0, \bar{z}_1, \bar{z}_2, \rho) = 0$ holds. This is guaranteed to provide the desired solution, as $\tilde{V}(0, \bar{z}_1, \bar{z}_2, \rho)$ is strictly decreasing in ρ as long as $t_w > 0$ (rather than $t_w = 0$; see Supplementary Note 1).

Bellman's equation for maximizing the correct rate. We now move to assuming that, all that matters to the decision maker is to identify the higher rewarded option, irrespective of the associated true reward. To do so, we abolish the identity between true and experienced reward, z and r, and instead assume that an experienced reward of $r = R_\mathrm{corr}$ is associated with choosing option j if $z_j > z_i$, $i \neq j$, and a reward of $r = r_\mathrm{incorr}$ with the alternative choice. This captures the case of maximizing the correct rate for value-based decisions, and also relates closely to simpler perceptual decisions in which the decision maker only gets rewarded for correct choices (for example, ref. 29), as long as the momentary evidence is well-approximated by a Gaussian. Evidence accumulation in this setup remains unchanged from before, as the posterior $z \mid \delta\boldsymbol{x}(0:t)$ contains all information required to compute the expected experienced reward. This posterior is fully specified by the sufficient statistics $(t, \bar{z}_1, \bar{z}_2)$, where we have defined $\bar{z}_j = z_j \mid \delta\boldsymbol{x}(0:t)$.

The value function for single, isolated trials changes in two ways. First, it is now defined over $(t, \bar{z}_1, \bar{z}_2)$ instead of $(t, \hat{r}_1, \hat{r}_2)$ (previously we had $(t, \hat{r}_1, \hat{r}_2) = (t, \bar{z}_1, \bar{z}_2)$, which does not hold anymore). Second, the value for deciding changes as follows. When choosing option one, the decision maker receives reward R_corr with probability $p(z_1 > z_2 \mid \delta\boldsymbol{x}(0:t))$ and reward R_incorr with probability $p(z_1 \leq z_2 \mid \delta\boldsymbol{x}(0:t))$. Thus, the expected reward associated with this choice is $R_\mathrm{corr}p(z_1 > z_2 \mid \delta\boldsymbol{x}(0:t)) + R_\mathrm{incorr}p(z_1 \leq z_2 \mid \delta\boldsymbol{x}(0:t))$. The expected reward for option two is found analogously, and results in the value for deciding

$$V_\mathrm{d}(t, \hat{z}_1, \hat{z}_2) = \max\left\{\begin{array}{c} R_\mathrm{corr}p(z_1 > z_2 \mid \delta\boldsymbol{x}(0:t)) + R_\mathrm{incorr}p(z_1 \leq z_2 \mid \delta\boldsymbol{x}(0:t)), \\ R_\mathrm{corr}p(z_2 > z_1 \mid \delta\boldsymbol{x}(0:t)) + R_\mathrm{incorr}p(z_2 \leq z_1 \mid \delta\boldsymbol{x}(0:t)) \end{array}\right\}.$$

As the posterior $\boldsymbol{z} \mid \delta\boldsymbol{x}(0:t)$ is Gaussian in all task setups we have considered, the probabilities in the above expression are cumulative Gaussian functions that are functions of the sufficient statistics $(t, \bar{z}_1, \bar{z}_2)$. Besides these two changes, the value function and associated Bellman Equation remain unchanged (see Supplementary Note 1 for explicit expressions).

Moving from single, isolated trials to maximizing the correct rate over sequences of trials requires the same changes as when moving from single trials to maximizing the reward rate. In particular, the value function turns into the average-adjusted value function that penalizes the passage of some time δt by $\rho \delta t$, where ρ is now the correct rate rather than the reward rate. The correct rate is still the average experienced reward (minus accumulation cost) per unit time, but—due to the changed definition of experienced reward—does not anymore relate to the true reward, but only if the option associated with the larger associated true reward was correctly identified. This causes the value for deciding to be additionally penalized by ρt_w. The value for waiting some more time δt to accumulate more evidence incurs an additional cost $\rho \delta t$, but remains unchanged otherwise. The average-adjusted value function is again invariant under addition of a constant, such we choose $\tilde{V}(0, \bar{z}_1, \bar{z}_2, \rho) = 0$. This fully specifies the value function and associated Bellman equation, which is provided in Supplementary Note 1.

Linearity of value function for waiting. Here, we show that value function for waiting increases linearly in line parallel to the diagonal within the $(\hat{r}_1, \hat{r}_2)$-space, which is required to show that the optimal decision boundaries are parallel to the diagonal. We will do so by a backwards induction argument in time. The base case for the induction argument relies on the shape of the value function for large times, $t \to \infty$. For such times, the decision maker incurs a large cost for accumulating evidence up until that time, and also expects to gain little further insight into the true rewards when accumulating more evidence. As a consequence, at such times it will always be better to decide immediately rather than to accumulate more evidence. Therefore, the value function will be given by the value for deciding, $V(t, \hat{r}_1, \hat{r}_2) = V_{\mathrm{d}}(\hat{r}_1, \hat{r}_2)$, which, as discussed in the previous paragraph, is linearly increasing in lines parallel to the diagonal.

The inductive step will show that, if the value function at time $t + \delta t$ is linearly increasing in lines parallel to the diagonal, then so it the value of waiting at time t, and, as a consequence, also the value function at time t. The value of waiting at time t is given by $V(t + \delta t, \hat{r}_1(t + \delta t), \hat{r}_2(t + \delta t)) - c\delta t$, where the expectation is over future expected rewards $\hat{r}_1(t + \delta t)$ and $\hat{r}_2(t + \delta t)$, and reflects the uncertainty about how the reward estimate evolves over time. For our case, the distribution describing this uncertainty is a bivariate Gaussian (as described in the previous sections), centred on the current expected rewards, $(\hat{r}_1, \hat{r}_2)$, and with a covariance that only depends on t. Its shift-invariant shape causes the expectation $V(t + \delta t, \hat{r}_1(t + \delta t), \hat{r}_2(t + \delta t))$ to be a smoothed version of $V(t + \delta t, \hat{r}_1, \hat{r}_2)$ that, as $V(t + \delta t, \hat{r}_1, \hat{r}_2)$, linearly increase in lines parallel to the diagonal. The value of waiting is this expectation shifted by the constant momentary cost $-c\delta t$, and therefore also has this property (Fig. 3b). This establishes that, if the value function at time $t + \delta t$ is linearly increasing in lines parallel to the diagonal, then so is the value of waiting at time t. The value function at time t is the maximum over the value for deciding and that for waiting. As both increase linearly in lines parallel to the diagonal, so does this value function, $V(t, \hat{r}_1, \hat{r}_2)$ (Fig. 3c,e). This completes the inductive step.

The induction argument shows that both value for deciding as well as that for waiting increases linearly with slope one in lines parallel to the diagonal for all t. This immediately means that, if they intersect at some point $\left(\hat{r}_1^*, \hat{r}_2^*\right)$, then they will intersect at the whole line $\left(\hat{r}_1^* + C, \hat{r}_2^* + C\right)$ that is parallel to the diagonal (Fig. 3c). As a consequence, the decision boundaries, which lie on the intersection between these two values, are parallel to this diagonal for all times t. See Supplementary Note 1 for the proof of the same property with an argument that does not rely on induction.

Finding the optimal policy numerically. To find the optimal policy for the above cases numerically, we computed the value function by backward induction[30], using Bellman's equation. Bellman's equation expresses the value function at time t as a function of the value function at time $t + \delta t$. Therefore, if we know the value function at some time T, we can compute it at time $T - \delta t$, then $T - 2\delta t$, and so on, until time $t = 0$. We usually chose some large T, significantly beyond the time horizon of interest, at which we set $V(T, \hat{z}_1, \hat{z}_2) = V_d(\hat{z}_1, \hat{z}_2)$, independent of the value at any $t > T$. For any time $t \leq T$, we represented the value function over the remaining two parameters $(\hat{z}_1, \hat{z}_2)$ (or $(\hat{r}_1, \hat{r}_2)$ in the value-based task) numerically over an equally space two-dimensional grid. This grid allowed us to compute the integral that represents the expectation over the future value numerically by the two-dimensional convolution between future value Vt + δt,1t + δt,z2t + δt and transition probability distribution $\mathrm{P}(\hat{\boldsymbol{z}}(t + \delta t) \mid \hat{\boldsymbol{z}}(t))$. For any such time, the optimal decision boundaries were found on this grid by the intersection of the value for deciding and that for waiting. We handled boundary effects in space and time by significantly extending the grid beyond the area of interest and cropping the value function after fully computing it over the extended range.

In the reward rate and correct rate case, computing the value function requires knowledge of the corresponding rate ρ. This ρ was unknown, but could be found by the condition $\hat{V}(0, \bar{z}_1, \bar{z}_2, \rho) = 0$. $\tilde{V}(0, \bar{z}_1, \bar{z}_2, \rho)$ is strictly decreasing in ρ (Supplementary Note 1), such that we could initially assume an arbitrary ρ for which we computed $\tilde{V}(0, \bar{z}_1, \bar{z}_2, \rho)$. The correct ρ was then found by iterating the computation of $\tilde{V}(0, \bar{z}_1, \bar{z}_2, \rho)$ within a root finding procedure until $\tilde{V}(0, \bar{z}_1, \bar{z}_2, \rho) = 0$.

The following parameters were used to generate the figures. We set the prior mean as $\bar{z}_1 = \bar{z}_2 = 0$, except for Fig. 5 where we varied $\bar{z}_1 + \bar{z}_2$ while fixing $\bar{z}_1 = \bar{z}_2$. The prior variance was $\sigma_z^2 = 16$, and observation noise $\sigma_x^2 = 4$, for both options. We used a grid spanning $-10 \leq \hat{z}_1 \leq 10$ and $-10 \leq \hat{z}_2 \leq 10$, in steps of 0.4 in both dimensions. The maximum time to consider was set to $T = 5$ s, with time-steps of size $\delta t = 0.005$ s for backward induction. To focus on the effect of reward rate, we assumed no explicit cost of evidence accumulation, $c = 0$ and a waiting time t_w set to 0.5 s.

Data availability. The authors declare that the data supporting the findings of this study are available within the article and its Supplementary Information File.

References

1. Link, S. W. & Heath, R. A. A sequential theory of psychological discrimination. *Psychometrika* **40,** 77–105 (1975).
2. Ratcliff, R. A theory of memory retrieval. *Psychol. Rev.* **85,** 59–108 (1978).
3. Gold, J. I. & Shadlen, M. N. Neural computations that underlie decisions about sensory stimuli. *Trends Cogn. Sci.* **5,** 10–16 (2001).
4. Wald, A. Sequential tests of statistical hypotheses. *Ann. Math. Stat* **16,** 117–186 (1945).
5. Wald, A. & Wolfowitz, J. Optimum character of the sequential probability ratio test. *Ann. Math. Stat.* **19,** 326–339 (1948).
6. Kira, S. *et al.* A neural implementation of wald's sequential probability ratio test. *Neuron* **85,** 861–873 (2015).
7. Drugowitsch, J., Moreno-Bote, R., Churchland, A. K., Shadlen, M. N. & Pouget, A. The cost of accumulating evidence in perceptual decision making. *J. Neurosci.* **32,** 3612–3628 (2012).
8. Krajbich, I., Armel, C. & Rangel, A. Visual fixations and the computation and comparison of value in simple choice. *Nat. Neurosci.* **13,** 1292–1298 (2010).
9. Milosavljevic, M., Malmaud, J., Huth, A., Koch, C. & Rangel, A. The drift diffusion model can account for the accuracy and reaction time of value-based choices under high and low time pressure. *Judgm. Decis. Mak.* **5,** 437–449 (2010).
10. Krajbich, I. & Rangel, A. Multialternative drift-diffusion model predicts the relationship between visual fixations and choice in value-based decisions. *Proc. Natl Acad. Sci. USA* **108,** 13852–13857 (2011).
11. Vickers, D. Evidence for an accumulator model of psychophysical discrimination. *Ergonomics* **1,** 37–58 (1970).
12. Teodorescu, A. R. & Usher, M. Disentangling decision models: from independence to competition. *Psychol. Rev.* **120,** 1–38 (2013).
13. Basten, U., Biele, G., Heekeren, H. R. & Fiebach, C. J. How the brain integrates costs and benefits during decision making. *Proc. Natl Acad. Sci. USA* **107,** 21767–21772 (2010).
14. Louie, K., Khaw, M. W. & Glimcher, P. W. Normalization is a general neural mechanism for context-dependent decision making. *Proc. Natl Acad. Sci. USA* **110,** 6139–6144 (2013).
15. Pirrone, A., Stafford, T. & Marshall, J. a. R. When natural selection should optimize speed-accuracy trade-offs. *Front. Neurosci.* **08,** 1–5 (2014).
16. Pais, D. *et al.* A mechanism for value-sensitive decision-making. *PLoS ONE* **8,** e73216 (2013).
17. Gao, J., Tortell, R. & McClelland, J. L. Dynamic integration of reward and stimulus information in perceptual decision-making. *PLoS ONE* **6,** 5–7 (2011).
18. Feng, S., Holmes, P., Rorie, A. & Newsome, W. T. Can monkeys choose optimally when faced with noisy stimuli and unequal rewards? *PLoS Comput. Biol.* **5,** e1000284 (2009).
19. Wang, X. Probabilistic decision making by slow reverberation in cortical circuits. *Neuron* **36,** 955–968 (2002).
20. Wang, X. J. Decision making in recurrent neuronal circuits. *Neuron* **60,** 215–234 (2008).
21. Brunton, B. W., Botvinick, M. M. & Brody, C. D. Rats and humans can optimally accumulate evidence for decision-making. *Science* **340,** 95–98 (2013).
22. Fudenberg, D., Strack, P. & Strzalecki, T. Stochastic choice and optimal sequential sampling (2015); Available at SSRN: http://ssrn.com/abstract=2602927 or http://dx.doi.org/10.2139/ssrn.2602927.
23. Hawkins, G. E., Forstmann, B. U., Wagenmakers, E.-J., Ratcliff, R. & Brown, S. D. Revisiting the evidence for collapsing boundaries and urgency signals in perceptual decision-making. *J. Neurosci.* **35,** 2476–2484 (2015).
24. Oud, B. *et al.* Irrational time allocation in decision-making. *Proc. R. Soc. B Biol. Sci* **283,** 20151439 (2016).
25. Churchland, A. K., Kiani, R. & Shadlen, M. N. Decision-making with multiple alternatives. *Nat. Neurosci.* **11,** 693–702 (2008).
26. Beck, J. M. *et al.* Probabilistic population codes for Bayesian decision making. *Neuron* **60,** 1142–1152 (2008).
27. Drugowitsch, J., Deangelis, G. C., Klier, E. M., Angelaki, D. E. & Pouget, A. Optimal multisensory decision-making in a reaction-time task. *Elife* **2014,** 1–19 (2014).
28. Mahadevan, S. Average reward reinforcement learning: foundations, algorithms, and empirical results. *Mach. Learn.* **22,** 159–195 (1996).
29. Kim, J. N. & Shadlen, M. N. Neural correlates of a decision in the dorsolateral prefrontal cortex of the macaque. *Nat. Neurosci.* **2,** 176–185 (1999).
30. Brockwell, A. E. & Kadane, J. B. A gridding method for bayesian sequential decision problems. *J. Comput. Graph. Stat.* **12,** 566–584 (2003).

Acknowledgements

This work was partially supported by Swiss National Foundation #31003A_143707 and a grant from the Simons Foundation (#325057).

Author contributions

S.T., J.D. and A.P. conceived the study. S.T. and J.D. developed the theory and conducted the mathematical analysis. S.T. performed the simulations. S.T., J.D. and A.P. interpreted the results and wrote the paper.

Additional information

Supplementary Information accompanies this paper at http://www.nature.com/naturecommunications

Competing financial interests: The authors declare no competing financial interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

How to cite this article: Tajima, S. *et al.* Optimal policy for value-based decision-making. *Nat. Commun.* 7:12400 doi: 10.1038/ncomms12400 (2016).