The online metacognitive control of decisions

D. Lee[3,4], J. Bénon[1], W. Hopper[1], M. Verdeil[1], M. Pessiglione[1], F. Vinckier[1], S. Bouret[1], M. Rouault[1], R. Lebouc[1], G. Pezzulo[4], C. Schreiweis[1], E. Burguière[1], J. Daunizeau[1,2]

[1] Paris Brain Institute, Paris, France

[2] ETH, Zurich, Switzerland

[3] Tel-Aviv University, Tel-Aviv, Israel

[4] Institute of Cognitive Sciences and Technologies, Rome, Italy

Address for correspondence:

Jean Daunizeau

Motivation, Brain and Behavior Group

Paris Brain Institute (ICM)

47, bd de l'Hôpital, 75013, Paris, France.

Tel: +33 1 57 27 43 26

Mail: jean.daunizeau@icm-institute.org

Web: http://sites.google.com/site/jeandaunizeauswebsite

Abstract

Difficult decisions typically involve mental effort, which signals the subjective cost of processing decision-relevant information. But how does the brain regulate mental effort? A possibility is that the brain optimizes a resource allocation problem, whereby the amount of invested resources optimally balances its expected cost (i.e. effort) and benefit. Our working assumption is that subjective decision confidence serves as the benefit term of the resource allocation problem, hence the “metacognitive” nature of decision control (Lee & Daunizeau, 2021). In this work, we present a computational model for the *online metacognitive control of decisions* or oMCD. Formally, oMCD is a Markov Decision Process that optimally solves the ensuing resource allocation problem under agnostic assumptions about the inner workings of the underlying decision system. We disclose its main properties, when coupled with two standard decision processes (namely: the ideal observer and the attribute integration cases). Importantly, we show that oMCD reproduces most established empirical results in the field of value-based decision making. Finally, we discuss the possible connexions of the model with most prominent neurocognitive theories about mental effort, and highlight potential extensions.

Introduction

There is no free lunch: obtaining reward typically requires investing effort. This holds even for mental tasks, which may involve mental effort for achieving success (in terms of, e.g., mnesic or attentional performance). Nevertheless, we sometimes invest very few mental effort, eventually rushing decisions and falling for all sorts of cognitive biases (Kahneman, 2011). So how does the brain regulate mental effort? A possibility is to understand mental effort regulation in terms of a resource allocation problem: namely, identifying the amount of cognitive resources that optimizes a cost/benefit tradeoff (Musslick et al., 2015; Shenhav et al., 2013, 2017). In this context, mental effort simply signals the subjective cost of investing resources, whose aversiveness is balanced by the anticipated performance gain. In conjunction with simple optimality principles, this idea has proven fruitful for understanding the relationship between mental effort and peoples' performance in various cognitive tasks, in particular those that involve cognitive control (Griffiths et al., 2015; Lieder et al., 2018). Recently, it was adapted to the specific case of value-based decision making, and framed as a self-contained computational model: namely, the Metacognitive Control of Decisions or MCD (Lee & Daunizeau, 2021).

Central to this work is the notion that decision confidence serves as the benefit term of the resource allocation problem, hence the “metacognitive” nature of decision control. Formally, confidence derives from the uncertainty of subjective value representations, which evolve over decision time as the brain processes more value-relevant information. In brief, low confidence induces a latent demand for mental effort: the brain refines uncertain value representations by deploying cognitive resources, until they reach the optimal confidence/effort trade-off. Interestingly, this mechanism was shown to explain the (otherwise surprising) phenomenon of choice-induced preference change (Lee and Daunizeau, 2020). More importantly, the MCD model makes quantitative out-of-sample predictions about many features of value-based decisions, including decision time, subjective feeling of effort, choice confidence and changes of mind. These predictions have already been tested in a systematic manner, using a dedicated behavioral paradigm (Lee and Daunizeau, 2021). Despite its remarkable prediction accuracy, the original derivation of the model suffered from one main simplifying but limiting approximation: it assumes that MCD is operating in a purely *prospective* manner, i.e. the MCD controller commits to a mental effort investment that is identified prior to the decision. In principle, this can be done by anticipating the prospective benefit (in terms of confidence gain) and cost of effort, given a prior or default representation of option values that would rely on fast/automatic/effortless processes (Lopez-Persem et al., 2016). The issue here, is twofold. First, it cannot explain variations in decision features (e.g., response times, choice confidence, etc) that occur in the absence of changes in default preferences. Second, it is somehow suboptimal, as it neglects *reactive* processes, which enables the MCD controller to re-evaluate – and improve on- the decision to stop or continue allocating resources, as new information comes in and value representations are updated (Tajima et al., 2016, 2019a). This work addresses these limitations, effectively proposing an "online" variant of MCD which we coin oMCD.

Models and methods

As we will see below, deriving an optimal reactive variant of MCD requires specific mathematical developments, which falls under the frame of Markov decision processes (Feinberg & Shwartz, 2012). But before we describe the oMCD model, let us first recall the prospective variant of MCD.

1. The prospective MCD model

Disclaimer: this section is a summary of the mathematical derivation of the MCD model, which has already been published (Lee & Daunizeau, 2021).

Let z be the amount of cognitive (e.g., executive, mnemonic, or attentional) resources that serve to process value-relevant information. Allocating these resources will be associated with both a benefit $B(z)$, and a cost $C(z)$. As we will see, both are increasing functions of z: $B(z)$ derives from the refinement of internal representations of subjective values of alternative options or actions that compose the choice set, and $C(z)$ quantifies how aversive engaging cognitive resources is (mental effort). In line with the framework of *expected value of control* (Musslick et al., 2015; Shenhav et al., 2013), we assume that the brain chooses to allocate the amount of resources $\hat{z}$ that optimizes the following cost-benefit trade-off:

$$\hat{z} = \arg\max_{z} E\left[B(z) - C(z)\right] \tag{1}$$

where the expectation accounts for predictable stochastic influences that ensue from allocating resources (this will be clarified below). Here, the benefit term is simply given by $B(z) = R \times P_c(z)$, where $P_c(z)$ is choice confidence and the weight R is analogous to a reward and quantifies the importance of making a confident decision. As we will see, $P_c(z)$

plays a pivotal role in the model, in that it captures the efficacy of allocating resources for processing value-relevant information. So, how do we define choice confidence?

We assume that the subjective evaluation of alternative options in the choice set is uncertain. In other words, the internal representations of values V_i of alternative options are probabilistic. Such a probabilistic representation of value can be understood in terms of, for example, an uncertain prediction regarding the to-be-experienced value of a given option. Without loss of generality, the probabilistic representation of option value takes the form of Gaussian probability density functions $p(V_i)=N(\mu_i,\sigma_i)$, where μ_i and σ_i are the mode and the variance of the probabilistic value representation, respectively (and i indexes alternative options in the choice set). This allows us to define choice confidence P_c as the probability that the (predicted) experienced value of the (to be) chosen item is higher than that of the (to be) unchosen item. When the choice set is composed of two alternatives, P_c is given by:

$$P_c \approx s\left(\frac{\pi|\Delta\mu|}{\sqrt{3(\sigma_1+\sigma_2)}}\right) \qquad (2)$$

where $s(x)=1/1+e^{-x}$ is the standard sigmoid mapping, and we assume that the choice follows the sign of the preference $\Delta\mu=\mu_1-\mu_2$. Note that Equation (2) derives from a moment-matching approximation to the Gaussian cumulative density function (Daunizeau, 2017).

We assume that the brain valuation system automatically generates uncertain estimates of options' value (Lebreton et al., 2009, 2015), before cognitive effort is invested in decision making. In what follows, μ_i^0 and σ_i^0 are the mode and variance of the ensuing prior value representations (we treat them as inputs to the MCD model). They yield an initial confidence level P_c^0. Importantly, this prior or default preference neglects existing value-relevant information that would require cognitive effort to be retrieved and processed (Lopez-Persem et al., 2016).

Now, how does the system anticipate the benefit of allocating resources to the decision process? Recall that the purpose of allocating resources is to process (yet unavailable) value-relevant information. The critical issue is thus to predict how both the uncertainty σ_i and the modes μ_i of value representations will eventually change, before having actually allocated the resources (i.e., without having processed the information). In brief, allocating resources essentially has two impacts: (i) it decreases the uncertainty σ_i , and (ii) it perturbs the modes μ_i in a stochastic manner.

The former impact derives from assuming that the amount of information that will be processed increases with the amount of allocated resources. Here, this implies that the variance $\sigma_i\left(z\right)$ of a given probabilistic value representation decreases in proportion to the amount of allocated resources, i.e.:

$$\sigma_i\left(z\right)=\frac{1}{\frac{1}{\sigma_i^0}+\beta z} \tag{3}$$

where σ_i^0 is the prior variance of the representation (before any effort has been allocated), and β controls the efficacy with which resources increase the precision of the value representation. As we will see below, Equation (3) has the form of a Bayesian update of the belief's precision in a Gaussian-likelihood model, where the precision of the likelihood term is βz . More precisely, β is the precision increase that follows from allocating a unitary amount of resources z . In what follows, we will refer to β as the "*type #1 effort efficacy*".

The latter impact follows from acknowledging the fact that the system cannot know how processing more value-relevant information will affect its preference before having allocated the corresponding resources. Let δ_i be the change in the position of the mode of the i^{th} value representation, having allocated an amount z of resources. The direction of the mode's perturbation δ_i cannot be predicted because it is tied to the information that is yet to be processed. However, a tenable assumption is to consider that the magnitude of the perturbation increases with the amount of information that will be processed. This reduces to stating that the variance of δ_i increases in proportion to z , i.e.:

$$\begin{aligned} &\mu_i(z) = \mu_i^0 + \delta_i \\ &\quad \delta_i \sim N(0, \gamma z) \end{aligned} \tag{4}$$

where μ_i^0 is the mode of the value representation before any effort has been allocated, and γ controls the relationship between the amount of allocated resources and the variance of the perturbation term δ . The higher γ, the greater the expected perturbation of the mode for a given amount of allocated resources. In what follows, we will refer to γ as the "*type #2 effort efficacy*". Note that Equation 4 treats the impact of future information processing as a non-specific random perturbation on the mode of the prior value representation. Our justification for this assumption is twofold: (i) it is simple, and (ii) and it captures the idea that the MCD controller is agnostic about how the allocated resources will be used by the underlying valuation/decision system. We will see that, in spite of this, the MCD controller can still make quantitative predictions regarding the expected benefit of allocating resources.

Now, predicting the net effect of resource investment onto choice confidence (from Equations (3) and (4)) is not entirely trivial. On the one hand, allocating effort will increase the precision of value representations, which mechanically increases choice confidence, all other things being equal. On the other hand, allocating effort can either increase or decrease the absolute difference $\left|\Delta\mu(z)\right|$ between the modes (and hence increase or decrease choice confidence). This, in fact, depends upon the direction of the perturbation term δ, which is a priori unknown. Having said this, it is possible to derive the *expected* response of the absolute difference between the modes that would follow from allocating an amount z of resources, in terms of its mean and variance:

$$\begin{cases} E\left[\left|\Delta\mu\right| \middle| z\right] = 2\sqrt{\dfrac{\gamma z}{\pi}}\exp\left(-\dfrac{\left|\Delta\mu^0\right|^2}{4\gamma z}\right) + \Delta\mu^0\left(2\times s\left(\dfrac{\pi\Delta\mu^0}{\sqrt{6\gamma z}}\right) - 1\right) \\ V\left[\left|\Delta\mu\right| \middle| z\right] = 2\gamma z + \left|\Delta\mu^0\right|^2 - E\left[\left|\Delta\mu\right| \middle| z\right]^2 \end{cases} \tag{5}$$

where we have used the expression for the first-order moment of the so-called "folded normal distribution". Importantly, $E\left[|\Delta\mu|\,|\,z\right]$ is always greater than $|\Delta\mu^0|$ and increases monotonically with z (as is $V\left[|\Delta\mu|\,|\,z\right]$). In other words, allocating resources is expected to increase the value difference, despite the fact that the impact of the perturbation term can go either way.

Equation 5 now enables us to derive the expected confidence level $\bar{P}_c(z) \triangleq E\left[P_c|z\right]$ that would result from allocating the amount of resource z:

$$\bar{P}_c(z) \approx s\left(\frac{\lambda E\left[|\Delta\mu|\,|\,z\right]}{\sqrt{1+\frac{1}{2}\left(\lambda^2 V\left[|\Delta\mu|\,|\,z\right]\right)^{\frac{3}{4}}}}\right) \quad (6)$$

where $\lambda = 1\big/\sqrt{3\left(\sigma_1(z)+\sigma_2(z)\right)}$. Of course, $\bar{P}_c(0) = P_c^0$, i.e. investing no resources yields no confidence gain. Moreover, the expected choice confidence $\bar{P}_c(z)$ always increase with z, irrespective of the efficacy parameters, as long as $\beta \neq 0$ or $\gamma \neq 0$. Equation 6 is important, because it quantifies the expected benefit of resource allocation, before having processed the ensuing value-relevant information.

To complete the cost-benefit model, we simply assume that the cost of allocating resources to the decision process increases monotonically with the amount of resources, i.e.:

$$C(z) = \alpha z^{\nu} \quad (7)$$

where α determines the effort cost of allocating a unitary amount of resources z (we refer to α as the "effort unitary cost"), and ν effectively controls the range of resource investments that result in noticeable cost variations (we refer to ν as the "cost power").

Finally, the MCD-optimal resource allocation $\hat{z}$ is identified by replacing Equations (5), (6) and (7) into Equation (1). Note that this implicitly assumes that the allocation of resources is similar for all alternative options in the choice set.

2. Online MCD: optimal stopping rule

We now augment this model, by assuming that the MCD controller is re-evaluating the decision to stop or continue allocating resources, as value representations are updated and online confidence changes. This makes the ensuing *oMCD* model a reactive extension of the above "purely prospective" MCD model, which relieves the system from the constraint of effort investment pre-commitment.

Let t be the decision time. Without loss of generality, we assume that there is a linear relationship between deliberation time and resource investment, i.e.: $z=\kappa t$, where κ is the amount of resources that is spent per unit of time. We refer to κ as “effort intensity”. By convention, we assume that the maximal decision time T (the so-called *temporal horizon*) corresponds to the exhaustion of all available resources. This implies that $T=1/\kappa$ (because we consider normalized resources amounts).

Now, at time t, the system holds probabilistic value representations with modes μ_t and variance $\sigma(t)$. These value representations yield the following confidence level (cf. Equation (2) above):

$$P_c\left(\Delta\mu_t,t\right)\approx s\left(\frac{\pi\left|\Delta\mu_t\right|}{\sqrt{3\left(\sigma_1(t)+\sigma_2(t)\right)}}\right) \tag{8}$$

where we have used some abuse of notation with time indices.

This confidence level can be greater or smaller than the initial confidence level P_c^0, because new information regarding option values has been assimilated since then. Of course, the system will anticipate that investing additional resources will increase its confidence (on average). But this may not always overcompensate the cost of spending more resources on the decision. Thus, what should the decision to stop or continue look like, in order to maximize the expected cost-benefit tradeoff? It turns out that this problem is one of *optimal* *stopping*, which is a special case of Markov Decision Processes (Feinberg & Shwartz, 2012; Papadimitriou & Tsitsiklis, 1987). As we will see, it can be solved recursively (backward in time) using Bellman's optimality principle (Bellman, 1957).

Let $a_t \in \{0,1\}$ be the action that is taken at time t, where $a_t = 0$ (resp. $a_t = 1$) means that the system stops (resp. continues) deliberating. Let $Q(a_t, \Delta\mu_t, t)$ be the discounted benefit that the decision system would obtain at time t:

$$Q(a_t, \Delta\mu_t, t) = \begin{cases} R \times P_c(\Delta\mu_t, t) - \alpha(\kappa t)^\nu & \text{if } a_t = 0 \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

where the cost of effort investment $\alpha(\kappa t)^\nu$ has been rewritten in terms of decision time.

A time t, the optimal stopping policy derives from a comparison between the discounted benefit of stopping now (i.e. $Q(0, \Delta\mu_t, t)$) and some (yet undefined) threshold value $\omega(t)$, which may depend upon decision time. Let $\pi_\omega(t)$ be the stopping policy that is induced by the threshold $\omega(t)$:

$$\pi_\omega(t) = \begin{cases} 1 \text{ if } Q(0, \Delta\mu_t, t) \geq \omega(t) \\ 0 \text{ otherwise} \end{cases} \tag{10}$$

Finding the optimal stopping policy π_ω^* thus reduces to finding the optimal threshold $\omega^*(t)$.

By definition, at $t = T$, the system is stopping its deliberation, irrespective of its current discounted benefit $Q(0, \Delta\mu_T, T)$. By convention, the optimal threshold $\omega^*(T)$ can thus be written as:

$$\begin{aligned} \omega^*(T) &= \min_{\Delta\mu_T} Q(0, \Delta\mu_T, T) \\ &= Q(0, 0, T) \\ &= R/2 - \alpha(\kappa T)^\nu \end{aligned} \tag{11}$$

Now, at $t = T-1$, the discounted benefit $Q(0,\Delta\mu_{T-1},T-1)$ of stopping now can be compared to the expected discounted benefit $E\left[Q(0,\Delta\mu_T,T)\middle|\Delta\mu_{T-1}\right]$ of stopping at time $t = T$, conditional on the current value mode difference $\Delta\mu_{T-1}$:

$$\begin{aligned} E\left[Q(0,\Delta\mu_T,T)\middle|\Delta\mu_{T-1}\right] &= R\times E\left[P_c\ (\Delta\mu_T,T)\middle|\Delta\mu_{T-1}\right]-\alpha(\kappa T)^{\nu} \\ &\approx R\times s\left(\frac{\lambda(T)\times E\left[\left|\Delta\mu_T\right|\middle|\Delta\mu_{T-1}\right]}{\sqrt{1+\frac{1}{2}\left(\lambda(T)^2\times V\left[\left|\Delta\mu_T\right|\middle|\Delta\mu_{T-1}\right]\right)^{3/4}}}\right)-\alpha(\kappa T)^{\nu} \end{aligned} \quad (12)$$

where moments of the absolute value mode difference follow Equation 5 (for a unitary time increment):

$$\begin{cases} E\left[\left|\Delta\mu_T\right|\middle|\Delta\mu_{T-1}\right] = 2\sqrt{\frac{\gamma\kappa}{\pi}}\exp\left(-\frac{\left|\Delta\mu_{T-1}\right|^2}{4\gamma\kappa}\right)+\Delta\mu_{T-1}\left(2\times s\left(\frac{\pi\Delta\mu_{T-1}}{\sqrt{6\gamma\kappa}}\right)-1\right) \\ V\left[\left|\Delta\mu_T\right|\middle|\Delta\mu_{T-1}\right] = 2\gamma\kappa+\left|\Delta\mu_{T-1}\right|^2-E\left[\left|\Delta\mu_T\right|\middle|\mu_{T-1}\right]^2 \\ \lambda(T) = \frac{1}{\sqrt{3\left(\sigma_1(T)+\sigma_2(T)\right)}} \\ \sigma_i(T) = \frac{1}{\frac{1}{\sigma_i\ (T-1)}+\beta\kappa} \end{cases} \quad (13)$$

Here, the optimal decision is simply to stop if $Q(0,\Delta\mu_{T-1},T-1) \geq E\left[Q(0,\Delta\mu_T,T)\middle|\Delta\mu_{T-1}\right]$, and to continue otherwise. Note that both $Q(0,\Delta\mu_{T-1},T-1)$ and $E\left[Q(0,\Delta\mu_T,T)\middle|\Delta\mu_{T-1}\right]$ are deterministic functions of $\Delta\mu_{T-1}$. More precisely, they are both monotonically increasing with $\Delta\mu_{T-1}$ (see Figure 1 below), because current confidence and expected future confidence monotonically increase with $\Delta\mu_{T-1}$. Critically however, these functions have a different offset at $\Delta\mu_{T-1} = 0$, i.e.: $Q(0,0,T-1) < E\left[Q(0,\Delta\mu_T,T)\middle|\Delta\mu_{T-1} = 0\right]$ as long as $\gamma > 0$. In addition, they reach a different plateau at the infinite value mode difference limit, i.e.: $\lim\limits_{\Delta\mu_{T-1}\to\infty} Q(0,\Delta\mu_{T-1},T-1) > \lim\limits_{\Delta\mu_{T-1}\to\infty} E\left[Q(0,\Delta\mu_T,T)\middle|\Delta\mu_{T-1}\right]$ as long as $\alpha > 0$. This is important, because this implies that there exists a value mode difference $\Delta\mu^*_{T-1}$ such that $Q\left(0,\Delta\mu^*_{T-1},T-1\right)=E\left[Q\left(0,\Delta\mu_T,T\right)\middle|\Delta\mu^*_{T-1}\right]$. The discounted benefit at that point is the optimal threshold at $t=T-1$, i.e.: $\omega^*\left(T-1\right)=E\left[Q\left(0,\Delta\mu_T,T\right)\middle|\Delta\mu^*_{T-1}\right]$.

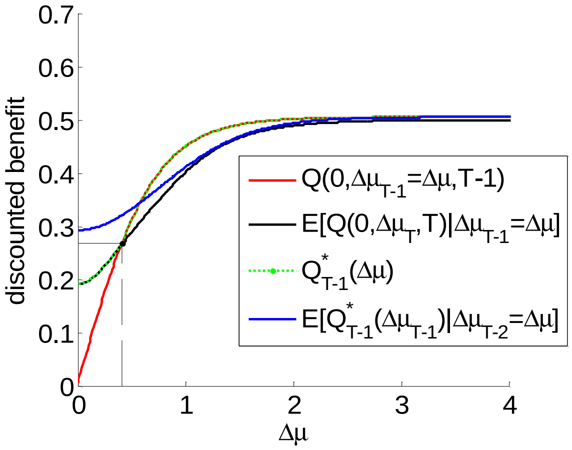

Figure 1: derivation of oMCD's optimal stopping strategy. Discounted benefits (y-axis) are plotted against the value mode difference $\Delta\mu$ (x-axis). The red and black lines show the current discounted benefit $Q\left(0,\Delta\mu_{T-1},T-1\right)$ if the system was stopping at $t=T-1$, and the expected discounted benefit $E\left[Q\left(0,\Delta\mu_T,T\right)\middle|\Delta\mu_{T-1}\right]$ at $t=T-1$, respectively (when setting $\Delta\mu_{T-1}=\Delta\mu$). The dotted green line shows the optimal discounted benefit $Q^*_{T-1}\left(\Delta\mu_{T-1}\right)$, which is simply the maximum over the two above benefit functions for each value mode difference $\Delta\mu_{T-1}$. Finally, the blue line shows the expected optimal discounted benefit $E\left[Q^*_{T-1}\left(\Delta\mu_{T-1}\right)\middle|\Delta\mu_{T-2}\right]$ at $t=T-1$ (when setting $\Delta\mu_{T-2}=\Delta\mu$). See main text.

Now, let us move one step backward in time, at $t=T-2$. Here again, the optimal strategy is to stop if the current discounted benefit $Q\left(0,\mu_{T-2},T-2\right)$ is higher than the expected future discounted benefit $E\left[Q\left(a_{T-1},\Delta\mu_{T-1},T-1\right)\middle|\Delta\mu_{T-2}\right]$, conditional on $\Delta\mu_{T-2}$. However, the latter now depends upon a_{T-1}, i.e. whether the system will later decide to stop or to continue:

$$E\left[Q\left(a_{T-1},\Delta\mu_{T-1},T-1\right)\middle|\Delta\mu_{T-2}\right]=\begin{cases}E\left[Q\left(0,\Delta\mu_{T-1},T-1\right)\middle|\Delta\mu_{T-2}\right] & \text{if } a_{T-1}=1\\ E\left[E\left[Q\left(0,\Delta\mu_T,T\right)\middle|\Delta\mu_{T-1}\right]\middle|\Delta\mu_{T-2}\right] & \text{otherwise}\end{cases} \quad (14)$$

The optimal policy cannot be directly identified from Equation (14). This is where we resort to Bellman's optimality principle: namely, whatever the current state and decision are, the remaining decisions of an optimal policy must also constitute an optimal policy with regard to the state resulting from the current decision (Bellman, 1957). Practically speaking, the derivation of the optimal threshold at $t=T-2$ is done under the constraint that oMCD's next decision follows the optimal policy, i.e. $a_{T-1}=\pi_{\omega}(T-1)$. Let $Q_t^*(\Delta\mu_t)$ be the expected discounted benefit evaluated under the optimal policy at time t, where the expectation is taken over all the sources of remaining uncertainty, when conditioning on the current value mode difference $\Delta\mu_t$. In what follows, we refer to $Q_t^*(\Delta\mu_t)$ as the "optimal discounted benefit". Under Bellman's optimality principle, the optimal strategy at $t=T-1$ is to stop if the current discounted benefit $Q(0,\mu_{T-2},T-2)$ is higher than the expected optimal discounted benefit $E\left[Q_{T-1}^*(\Delta\mu_{T-1})\middle|\Delta\mu_{T-2}\right]$.

Now, at time $t=T-1$, the optimal discounted benefit is given by:

$$Q_{T-1}^*(\Delta\mu_{T-1}) \square \max\left\{Q(0,\Delta\mu_{T-1},T-1),E\left[Q(0,\Delta\mu_T,T)\middle|\Delta\mu_{T-1}\right]\right\} \tag{15}$$

Note that $Q_{T-1}^*(\Delta\mu_{T-1})$ is just another function of $\Delta\mu_{T-1}$ (cf. dotted green curve in Figure 1). This means that the only source of stochasticity in $Q_{T-1}^*(\Delta\mu_{T-1})$ comes from μ_{T-1}, which can nonetheless be predicted (with some uncertainty), given the current value mode μ_{T-2}. In turn, this makes the expected optimal discounted benefit $E\left[Q_{T-1}^*(\Delta\mu_{T-1})\middle|\Delta\mu_{T-2}\right]$ a deterministic function of $\Delta\mu_{T-2}$. Here again, there is a critical value mode difference $\Delta\mu_{T-2}^*$ such that $Q(0,\Delta\mu_{T-2}^*,T-2)=E\left[Q_{T-1}^*(\Delta\mu_{T-1})\middle|\Delta\mu_{T-2}\right]$. The discounted benefit at that point is the optimal threshold $\omega^*(T-2)$ at $t=T-2$.

In fact, the reasoning is the same for all times $t<T-1$:

First, the expected optimal discounted benefit obeys the following backward recurrence relationship (Bellman equation for all $t<T-1$):

$$E\left[Q_t^*\left(\Delta\mu_t\right)\middle|\Delta\mu_{t-1}\right]=E\left[\max\left\{Q\left(0,\Delta\mu_t,t\right),E\left[Q_{t+1}^*\left(\Delta\mu_{t+1}\right)\middle|\Delta\mu_t\right]\right\}\middle|\Delta\mu_{t-1}\right] \quad (16)$$

This equation is solved recursively backward in time, starting at the expected discounted benefit at $t=T-1$, as given in Equation (12).

Second, the optimal threshold at time t is given by:

$$\omega^*\left(t\right)=Q\left(0,\Delta\mu_t^*,t\right) \quad (17)$$

where $\Delta\mu_t^*$ is the critical value mode difference, i.e. $\Delta\mu_t^*$ is such that:

$$Q\left(0,\Delta\mu_t^*,t\right)=E\left[Q_{t+1}^*\left(\Delta\mu_{t+1}\right)\middle|\Delta\mu_t=\Delta\mu_t^*\right] \quad (18)$$

Since the discounted benefit is a deterministic function of decision confidence, the oMCD-optimal threshold $\omega^*\left(t\right)$ for discounted benefits can be transformed into an oMCD-optimal confidence threshold $\omega_P^*\left(t\right)$, as follows:

$$\omega_P^*\left(t\right)=\frac{\omega^*\left(t\right)+\alpha\left(\kappa t\right)^\nu}{R} \quad (19)$$

This closes the derivation of oMCD's optimal stopping policy.

Note that the derivation of oMCD's optimal stopping policy requires prior information regarding the upcoming decision: namely, prior moments of value representations, type #1 and #2 effort efficacies, decision importance, unitary effort cost and cost power. This means that oMCD implicitly includes a *prospective* component, which is used to decide how to optimally *react* to a particular (stochastic) internal state of confidence. In other terms, oMCD is a mixed prospective/reactive policy.

Figure 2 below shows a representative instance of oMCD's optimal stopping strategy, from 500 Monte-Carlo simulations (using decision parameters R=1, α=0.2, β=1, γ=4, κ=1/100, υ=0.5, σ_0=1 and $\Delta\mu_0$=0).

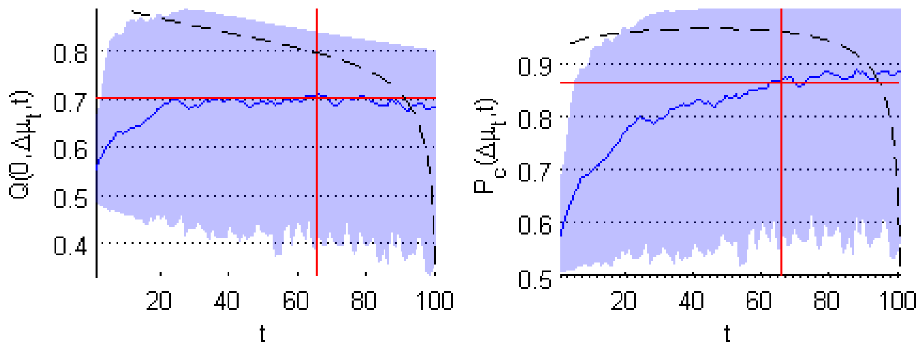

Figure 2: oMCD's optimal stopping strategy. Left panel: The black dotted line shows the oMCD-optimal discounted benefit threshold. The blue line and shaded area depict the mean and standard deviation of discounted benefit dynamics (over the 500 Monte-Carlo simulations), respectively. The vertical red line indicates the optimal resource allocation as obtained from the prospective variant of MCD, and the horizontal red line depicts the corresponding expected discounted benefit level. **Right panel**: The black dotted line shows the oMCD-optimal confidence threshold. The blue line and shaded area depict the mean and standard deviation of decision confidence (over the same Monte-Carlo simulations).

First, the prospective variant of MCD correctly identifies the decision time that maximizes the expected discounted benefit. However, one can see that oMCD's optimal stopping rule would, in most cases, demand higher confidence than its prospective variant. Second, one can see that oMCD's discounted benefit threshold $\omega^*(t)$ is monotonically decreasing with decision time. This, in fact, trivially derives from the max operator that constitutes the recursive relationship in Equation (16). However, the corresponding confidence threshold $\omega_P^*(t)$ may not be a monotonic function of decision time (cf. right panel in Figure 2). In fact, one can think of the shape of the confidence threshold (over time) as oMCD's prospective component, which is determined by the decision parameters. In the theoretical results section, we will investigate how the decision parameters (prior moments of value representations, type #1 and #2 effort efficacies, decision importance, unitary cost of effort and cost power) influence oMCD's policy. But let us first relate the MCD framework to standard decision processes.

3. How does MCD relate to standard decision processes?

By itself, the MCD framework does not commit to any specific assumption regarding how value-relevant information is processed. Nevertheless, the properties of decisions that are controlled through MCD actually depend upon how probabilistic value representations change over time. In what follows, we focus on two specific scenarios of value representation update, and disclose their connection with MCD.

- The ideal observer case.

Let us first consider the so-called *ideal observer* case, i.e. Bayesian inference on a hidden value signal. Note that, in this case, the optimal stopping rule - for maximizing expected reward rate - reduces to a specific instance of so-called *drift-diffusion decision* models with decaying bounds (Tajima et al., 2016, 2019b).

Assume that, at each time point, the decision system receives one partially unreliable copy y_t of the (hidden) value V of each alternative option. More precisely, y_t is a noisy input signal that is centered around V, i.e.: $y_t = V + \varepsilon_t$, where the noise term ε_t is gaussian with zero mean and variance Σ (and we have dropped the option indexing for notational simplicity). One may think of Σ as measuring the (lack of) reliability of the input value signal. This induces the following likelihood function for the hidden value: $p\left(y_t \middle| V\right) = N\left(V, \Sigma\right)$. Finally, assume that the decision system holds a Gaussian prior belief about the hidden options' value, i.e.: $p\left(V\right) = N\left(\mu_0, \sigma_0\right)$, where μ_0 and σ_0 are the corresponding prior mean and variance. At time *t*, an ideal (Bayesian) observer would assimilate the series of noisy signals to derive a probabilistic (posterior) representation $p\left(V \middle| y_1, \ldots, y_t\right) = N\left(\mu_t, \sigma_t\right)$ of options' value with the following mean and variance:

$$\begin{cases} \mu_t = \mu_0 + \tilde{\delta}_t \\ \sigma_t = \dfrac{1}{\dfrac{1}{\sigma_0} + t \times \dfrac{1}{\Sigma}} \end{cases} \tag{20}$$

where the change in the value mode $\tilde{\delta}$ is given by:

$$\tilde{\delta}_t = \frac{1}{\dfrac{\Sigma}{\sigma_0} + t} \sum_{t'=1}^{t} \left(y_{t'} - \mu_0 \right) \tag{21}$$

In brief, Equation (21) states that the value mode changes in proportion to prediction errors (i.e. $y_t - \mu_0$), which the ideal observer accumulates as she is sampling more input value signals. The stochasticity of $\tilde{\delta}$ is driven by the random perturbation term in the incoming noisy value signal. Conditioned on the hidden value V, it is easy to show that $E\left[\tilde{\delta} \middle| V\right] \propto V - \mu_0$. This implies that the random walk in Equation (20) actually has a nonzero drift that is proportional to the hidden value. Importantly however, the ideal observer does not know what the hidden value V is. Prior to the decision, her expectation is simply that $E[y] = E[V] = \mu_0$ and therefore $E\left[\tilde{\delta}\right] = 0$. In fact, this holds true at any time *t*: the ideal observer's expectation about the future change in her value belief mode (i.e. $E\left[\mu_{t+1} \middle| y_1, ..., y_t\right] - \mu_t$) is always zero, because her expectation about the next value signal reduces to her current value mode (i.e., $E\left[y_{t+1} \middle| y_1, ..., y_t\right] = \mu_t$). In other words, although the mode changes $\tilde{\delta}$ actually have a nonzero mean (as long as V deviates from the mode of the observer's belief), the ideal observer's expectation about its future realizations is always zero.

Nevertheless, the ideal observer can accurately predict how the precision of her belief will change. Comparing Equations (3) and (20) suggests that, under the ideal observer scenario, type #1 effort efficacy simply reduces to: $\beta = 1/\kappa\Sigma$. This means that type #1 effort efficacy simply increases with the reliability of the input value signal that the ideal observer is sampling.

In addition, although the ideal observer cannot anticipate in what direction the to-be-sampled signal will modify the mode of her posterior belief, she can derive a prediction over the magnitude of the change:

$$E\left[\tilde{\delta}_t^2\right]=t\times\frac{\Sigma+t\sigma_0}{\left(\frac{\Sigma}{\sigma_0}+t\right)^2} \quad (22)$$

where the expectation is derived under the agent's prior belief about the hidden value. Now, Equation 4 defines type #2 effort efficacy in terms of the ratio $E\left[\tilde{\delta}_t^2\right]\Big/\kappa t$ of expected change magnitude over effort investment (where $z=\kappa t$). Note that, under Equation (22), this quantity varies as a function of decision time. Thus, under the ideal observer scenario, type #2 effort efficacy can be approximated as its sample average over all admissible decision times, i.e.: $\gamma\approx 1/T\sum_{t=1}^{T}\left(\Sigma+t\sigma_0\right)\Big/\left(\Sigma/\sigma_0+t\right)^2\kappa$. This is only an approximation of course, since $E\left[\tilde{\delta}_t^2\right]$ eventually tails off as time increases, because noisy value signals that are sampled later in time have a smaller effect on the posterior mode. In other words, would the MCD controller know about the inner workings of the underlying (here: ideal observer) value updating system, it would rely on Equation (22) (as is done in Tajima, Drugowitsch, and Pouget 2016; Tajima et al. 2019b) rather than on Equation (4).

- The attribute-integration case.

Second, let us consider another type of scenario, which essentially proceeds from progressively integrating the value-relevant attributes of alternative options. This typically happens when alternative options can be decomposed into multiple dimensions that may enter in conflict with each other (cf., e.g., risk, delay or effort discounting).

Let $x_1,...,x_k$ be the set of k such value-relevant attributes, the combination of which is specific to each alternative option. Assume that the decision system constructs the value of alternative options according to a weighted sum of attributes, i.e.: $V=\sum_k w_k \times x_k$, where the attribute weights w_k are the same for all alternative options. Assume that each attribute is sampled from a gaussian distribution with mean η_k and variance ς_k, i.e. $p(x_k)=N(\eta_k,\varsigma_k)$. Finally, assume that attributes are available to the decision system one at a time, i.e. decision time steps co-occur with attribute-disclosing events. For the sake of simplicity, we set the decision's temporal horizon to $T=k$, i.e. we focus on the decision to stop integrating value-relevant attributes (and we ignore the additional mental effort that may be required to construct value from known attributes). In what follows, we refer to this scenario as the *attribute integration* model.

In the absence of default preferences, the system holds a prior representation about options' value that is maximally uninformative. This is because, prior to the decision, any combination of value-relevant attributes is admissible, and the system did not disclose the options' attributes yet. The first two moments of the system's prior value representation $p(V)=N(\mu_0,\sigma_0)$ are thus given by:

$$\begin{cases} \mu_0=\sum_{k'=1}^{k} w_{k'} \times \eta_k \\ \sigma_0=\sum_{k'=1}^{k} {w_{k'}}^2 \times \varsigma_k \end{cases} \quad (23)$$

Now, as time unfolds and the decision system discloses the value-relevant attributes, it progressively removes sources of uncertainty about the value of alternative options. In principle, if the system reaches the temporal horizon, then it knows all the attributes and can evaluate the alternative options with infinite precision. However, as long as some attributes are missing, value representations remain uncertain. Let K_t be the set of attribute indices that have been available to the decision system up until time t. At time t, the decision thus holds an updated probabilistic representation of value $p\left(V\middle|x_{K_t}\right)=N\left(\mu_t,\sigma_t\right)$ with the following mean and variance:

$$\begin{cases}\mu_t=\mu_0+\tilde{\delta}_t \\ \sigma_t=\sigma_0-\sum\limits_{k'\in K_t} w_{k'}{}^2\times\varsigma_{k'}\end{cases} \tag{24}$$

where the change in the value mode is simply given by:

$$\tilde{\delta}_t=\sum_{k'\in K_t} w_{k'}\times\left(x_{k'}-\eta_{k'}\right) \tag{25}$$

Note that here, variability in mode changes does not arise from some form of stochasticity or unreliability of input signals, as is the case for the “ideal observer” scenario. Rather, it derives from the arbitrariness of the permutation order with which attributes become available for options’ evaluation. However, should the full set of attributes be eventually disclosed, the estimated value would be $\mu_k=\sum_{k'}^{k} w_{k'}\times x_{k'}$, with full certainty.

Here again, the decision system cannot anticipate in which direction the future value mode will change, i.e. its expectation over future mode changes always is $E\left[\tilde{\delta}_t\right]=0$ at any point in time (because $E\left[x_k\right]=\eta_k$). Nevertheless, it can derive a prediction over the magnitude of the change, by averaging over all possible permutation orders:

$$E\left[\tilde{\delta}_t^2\right]=\frac{t}{k}\sum_{k'=1}^{k} w_{k'}{}^2\times\varsigma_{k'} \tag{26}$$

Comparing Equations (4) and (26) suggests that type #2 effort efficacy simplifies to: $\gamma=\sigma_0$. This means that type #2 effort efficacy simply scales with the range of attributes’ variation. Note that this prior prediction would need to be updated as time unfolds, because the remaining set of admissible permutations of yet unseen attributes progressively shrinks. This however, only entails a minor modification to Equation (26): practically speaking, a simple truncation of the sum. In any case, deriving the prospective component of oMCD from Equation (26) - or its proper variant - does not entail any approximation.

So how about type #1 effort efficacy? Note that one cannot directly compare Equation (24) to Equation (3), because of the arbitrariness of the order of attribute-disclosing events. However, as for type #2 efficacy, averaging over all possible permutations yields the following expected change in precision: $E\left[1/\sigma_t - 1/\sigma_0\right] = t \times 1/\sigma_0\left(k-t\right)$. Using the same logic as above, this suggests that, in this scenario, type #2 effort efficacy can be approximated as: $\beta \approx 1/\left(k-1\right)\sum_{t=1}^{k-1} 1/\kappa\sigma_0\left(k-t\right)$. Note that we have removed the time horizon from the average over admissible decision times, since it induces a singularity (infinite precision). Of course, would the MCD controller know about the mechanism of value construction (by attribute integration), it would rely on Equation (24) as opposed to Equation (3).

One can see that the definition of type #1 and type #2 effort efficacies actually depends upon the way the decision process changes the value representations. Recall that the above scenarios are just two examples out of many possibilities. In principle, optimal stopping would thus require variants of MCD controllers that are tailored to the underlying decision system. In this context, the MCD architecture that we propose provides an efficient alternative, which generalizes across decision processes and still operate quasi-optimal decision control. The only requirement here, is to calibrate the MCD controller over a few decision trials to learn effort efficacy parameters. Note that such calibration is expected to be very quick (at the limit: only one decision trial), because effort efficacies can be learned on within-trial dynamics (of value representations). This is effectively what we have done here, in an analytical manner, when deriving approximations for the effort efficacy parameters under distinct decision scenarios.

Results

In the previous section of this manuscript, we derived the online, dual prospective/reactive variant of MCD (and disclosed its connection with two exemplar decision systems). We now wish to illustrate its properties, when compared to its prospective variant.

1. What is the performance of oMCD?

At this point, one may ask whether oMCD eventually produces better decisions than prospective MCD, which operates by committing to a predefined temporal threshold. More precisely, under prospective MCD, the decision stops when the expected discounted benefit is maximal, which is evaluated prior to the decision (this corresponds to the red vertical line in Figure 2).

Now, does oMCD yields higher discounted benefits than prospective MCD (on average)?

To answer this question, we resort to Monte-Carlo simulations. In brief, we simulate a particular decision trial in terms of the stochastic dynamics of value representations, according to Equations (3) and (4), using the same decision parameters as for Figure 2. At each time step, oMCD's policy proceeds by comparing the ensuing confidence level to the optimal confidence threshold. When the confidence threshold is reached, we store the response time, which we note τ, as well as the ensuing confidence level and discounted benefit. We proceed similarly for prospective MCD, except that decision times are defined according to Equation (1). We then repeat the procedure to evaluate the average confidence levels, amount of invested resources, and discounted benefits induced by both MCD variants. These are summarized in Figure 3 below. Note: as a reference, we also compare MCD stopping strategies to a so-called "oracle" strategy, which identifies (post-hoc) the apex time, i.e. the time at which the stochastic discounted benefit is maximal. This provides an upper (though unachievable) bound to the expected discounted benefit of any stopping policy.

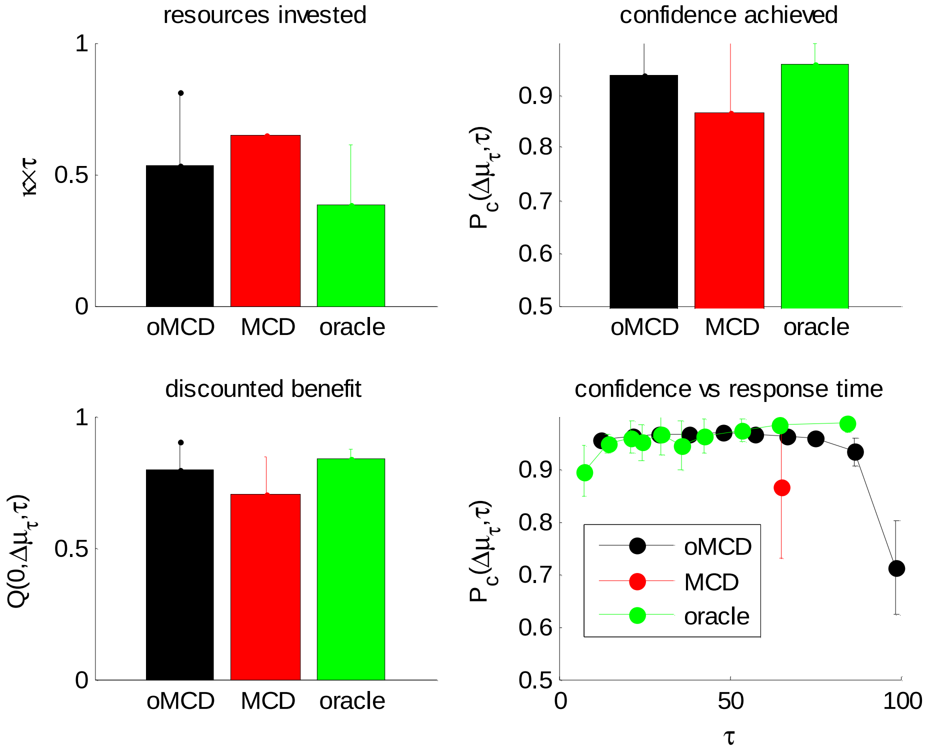

Figure 3: the performance of oMCD's optimal stopping strategy. Upper-left panel: the average amount of resources invested (y-axis) is shown as a function of the thresholding strategy (x-axis, grey: oMCD, red: prospective MCD, green: oracle). **Upper-right panel**: Average confidence level at the time of decision, same format. **Lower-left panel**: The average discounted benefits, same format. **Lower-right panel**: Confidence at response time (y-axis) is plotted against response times deciles (x-axis) for all strategies (oMCD: black, MCD: red, oracle: green).

One can see that prospective MCD tends to invest more resources that oMCD, though this yields lower confidence levels on average. In turn, the ensuing average discounted benefit is lower than that of oMCD (which is closer to that of the oracle). Interestingly, the statistical relationship between response times and confidence actually depends upon the stopping strategy. Under the oracle, this relationship is monotonic and increasing. This is because, everything else being equal, expected confidence increases with decision time - see Equation (6). This does not hold however, under oMCD stopping strategy. In particular, in our example, decisions that take longer eventually yield lower confidence. In fact, this relationship is determined by the shape of the optimal confidence threshold (cf. right panel of Figure 2).

Note: the relationship between confidence and response time for prospective MCD is trivial, because response time is fixed once decision parameters are set.

Importantly, the performance advantage of oMCD over prospective MCD does not derive from errors in the anticipation of discounted benefits over decision time. This is because, in principle, both MCD variants rely on the exact same information (i.e. decision parameters) to anticipate future discounted benefits. Rather, the performance advantage of oMCD derives from its reactive component, which enables it to realize that future discounted benefits are unlikely to improve. In other terms, one can think of oMCD as attempting to stop in the close neighborhood of the apex of discounted benefits. Recall that an apex is correctly identified if the discounted benefit at response time is higher than all discounted benefits before and after response time. Thus, from the perspective of apex identification, one can think of the decision to “continue until the decision time” as being correct if the discounted benefit at response time was higher than all previous discounted benefits. Reciprocally, the decision to “stop at decision time” was correct if the discounted benefit at response time was higher than all future discounted benefits. Figure 4 below summarizes the apex identification accuracy for oMCD, prospective MCD and oracle strategies, in terms of the rates of correct "continue" and "stop" decisions.

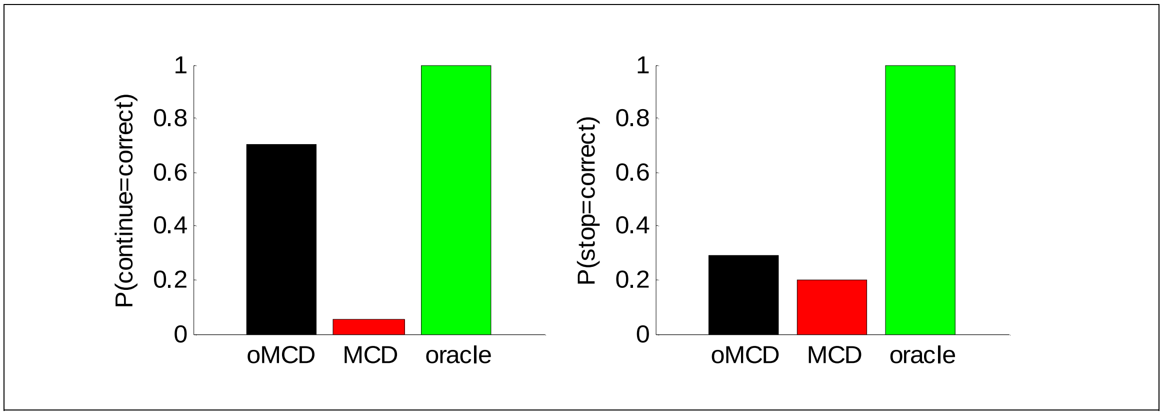

Figure 4: apex identification accuracy. Left panel: The rate of correct "continue until decision time" decisions (y-axis) is shown for all stopping strategies (oMCD: black, MCD: red, oracle: green). **Right panel:** rate of correct "stop at decision time" decisions, same format.

By design, the oracle strategy always accurately identifies the apex time. Clearly, prospective MCD tends to stop the decision too late, since "continue" decisions are almost always wrong (about 95% error rate). In comparison, oMCD is clearly better at identifying the discounted benefit apex. Note that, although oMCD is more often wrong about "stop" decisions (about 70% error rate) than about "continue" decisions (about 30% error rate), it seems that oMCD still tends to sometimes stop too late, when compared to the oracle (cf. upper-left panel of Figure 3). Note that although oMCD's performance advantage over prospective MCD generalizes over most decision parameters, it actually increases with γ, which controls the stochasticity of value representation dynamics. At the limit when $\gamma=0$, oMCD and prospective MCD stopping strategies are identical.

2. How do prospective MCD and oMCD differ?

Both prospective MCD and oMCD models predict decision properties (e.g., response time, confidence, etc), as a function of decision parameters. But how do these predictions compare with each other?

To answer this question, we conducted the following series of Monte Carlo simulations. First, we draw decision parameters (decision importance R, effort unitary cost α, effort cost power ν, effort efficacies β and γ, prior moment of value representations σ_0 and μ_0) at random, following a uniform probability density defined on the [0,8] interval. Second, we determine the resources investment under the prospective variant of MCD, and derive the oMCD optimal confidence threshold. Third, we simulate 500 stochastic dynamics of value mode and identify the corresponding oMCD response times. We then estimate the average resource investment and decision confidence, under oMCD (across the 500 simulations). We repeat this procedure 200 times, for different decision parameter samples. Note that we discard decision parameter samples that yield null resource investment under prospective MCD, because these correspond to decisions without deliberation. Figure 5 below summarizes the results of this Monte-Carlo simulations series.

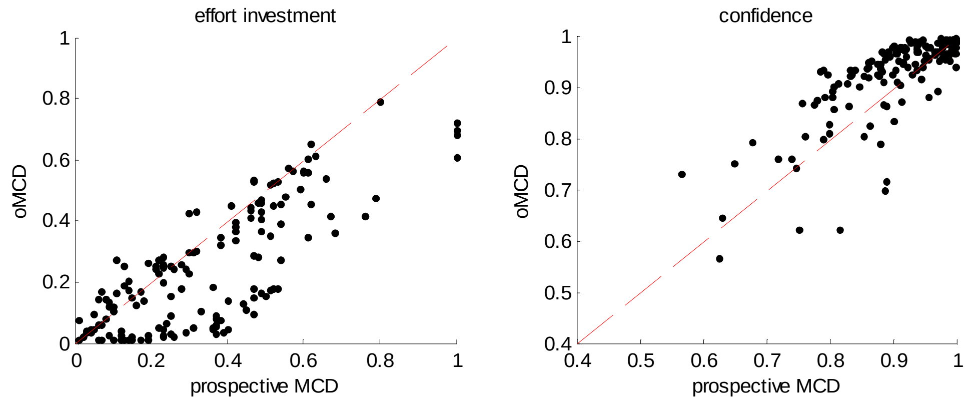

Figure 5: comparison between prospective MCD and oMCD. Left panel: the amount of resources invested under the prospective variant of MCD (x-axis) is plotted against the average amount of resources invested under oMCD (y-axis). Each dot corresponds to a set of decision parameters (200 samples). The red dotted line shows the identity mapping. **Right panel**: decision confidence, same format.

One can see that, on average, resource investments under both MCD variants exhibit a strong correlation (r=0.78). Qualitatively, the same is true for decision confidence (r=0.76). This implies that the impact of decision parameters on resource investments and confidence is very similar under both MCD variants (although oMCD's optimal stopping strategy tends to yield slightly lower resource investments and higher confidence than prospective MCD). This is important, because this means that the known properties of prospective MCD (Lee & Daunizeau, 2021) approximately generalize to oMCD.

However, when compared to prospective MCD, oMCD possesses a unique feature: namely, the potentially nontrivial statistical relationship between decision properties (e.g. confidence) and resource investments (as proxied using, e.g., response times), *across trials with identical decision parameters*. Although we did not comment on this until now, this was already exemplified in the lower-right panel of Figure 3.

To make this distinction clearer, we performed another set of simulations aiming at evaluating the impact of decision difficulty. Note that difficult decisions are those decisions where the reliability of value representations improve very slowly. Within the MCD framework, increasing decision difficulty can thus be modelled by decreasing type #1 effort efficacy. We systematically varied β from 1 to 4 (having set all the other decision parameters to 4), and evaluated both response times and confidence (for 500 stochastic dynamics of value representations per difficulty level). Figure 6 below summarizes the simulation results.

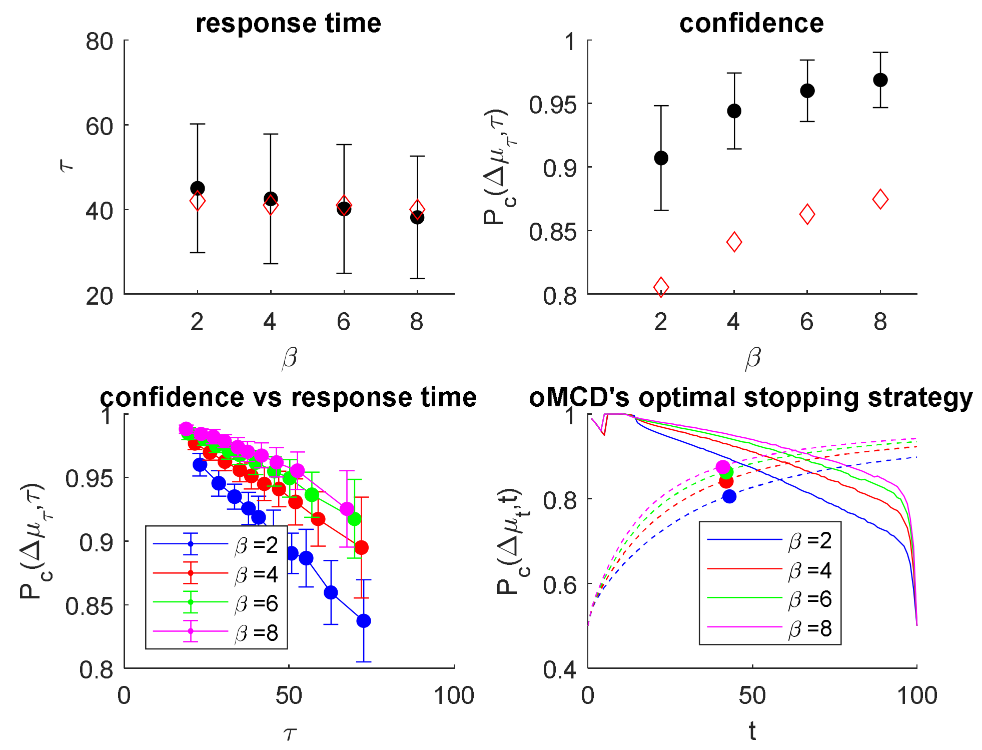

Figure 6: Impact of difficulty level. Upper-left panel: mean effort investment (y-axis, black dots) is plotted as a function of type #1 effort efficacy (x-axis). Errorbars depict standard deviations across trials, and red diamonds show the effort investment under prospective MCD. **Upper-right panel**: decision confidence, same format. **Lower-left panel**: confidence (y-axis) is plotted against response time deciles (x-axis), for each difficulty level (color code: type #1 effort efficacy), under oMCD's optimal policy. **Lower-right panel**: oMCD's confidence thresholds (y-axis, plain lines) are plotted against decision time (x-axis), for each difficulty level (same color code as lower-left panel). Dotted lines show expected confidence (as anticipated by prospective MCD), and dots show the optimal response times, under a prospective MCD strategy.

One can see that the net effect of increasing decision difficulty (or equivalently, decreasing type #1 effort efficacy) is to increase response time and decrease confidence. However, the (slightly concave) shape of the relationship between response time and confidence is preserved across difficulty levels. This derives from the fact that the corresponding dynamics of expected confidence and oMCD's confidence thresholds are qualitatively similar.

Figure 6 also reveals how oMCD's optimal stopping strategy prospectively anticipate the impact of decision difficulty. In brief, the decay rate of oMCD's confidence threshold increases with decision difficulty. However, this is overcompensated by the corresponding decrease in the ascend rate of expected confidence. This eventually determines the way oMCD trades speed against accuracy: difficult decisions are given more deliberation time than easy decisions (here, this is also true for prospective MCD).

Note that the effect of difficulty onto response time, as well as the shape of the relationship between response time and confidence, actually depend upon the decision parameter setting. In other terms, these effects do not generalize to all decision parameter settings. This dependency is exemplified in Figure 7 below, which replicates the above analysis, this time setting all decision parameters to 2 (except type #1 effort efficacy).

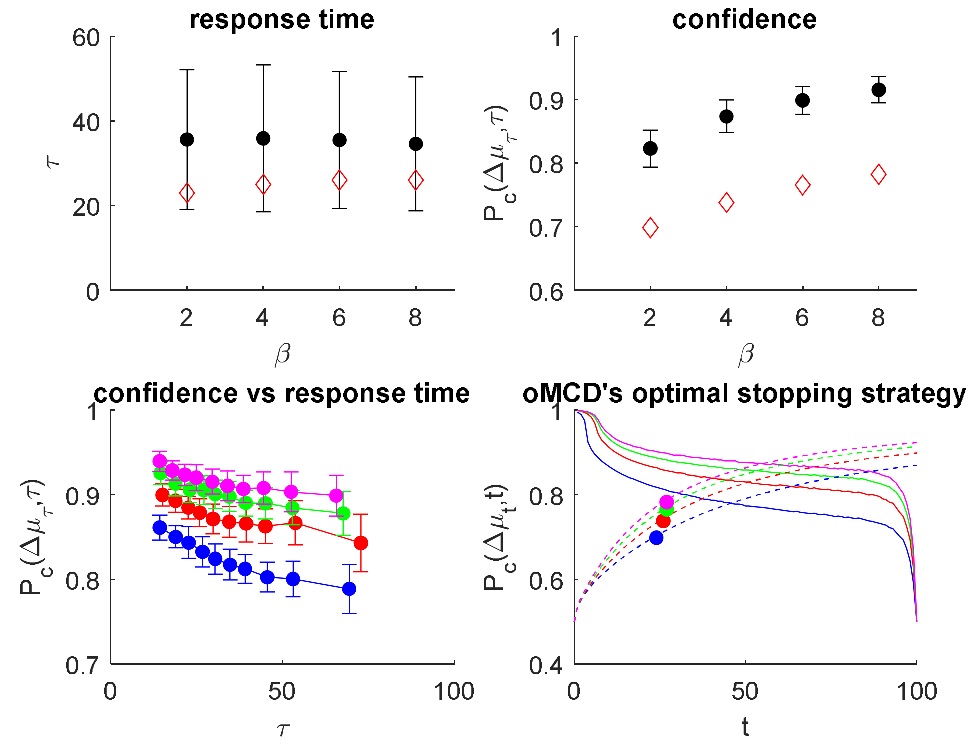

Figure 7: Impact of difficulty level. Same format as Figure 6.

One can see that, although the relationship between response time and confidence is now convex, decision difficulty has qualitatively similar effects on oMCD's behavior than before. However, prospective MCD and oMCD now slightly differ. More precisely, under prospective MCD, increasing decision difficulty now *decreases* response time. This is not the case for oMCD, which preserves its speed-accuracy tradeoff properties.

In summary, the shape of the relationship between confidence and response time is mostly determined by the dynamics of oMCD’s optimal confidence threshold, which itself depends upon decision parameters. Therefore, we systematically investigated the impact of each decision parameter on oMCD’s confidence threshold dynamics (when setting all the other ones to 1). This is summarized in Figure 8 below.

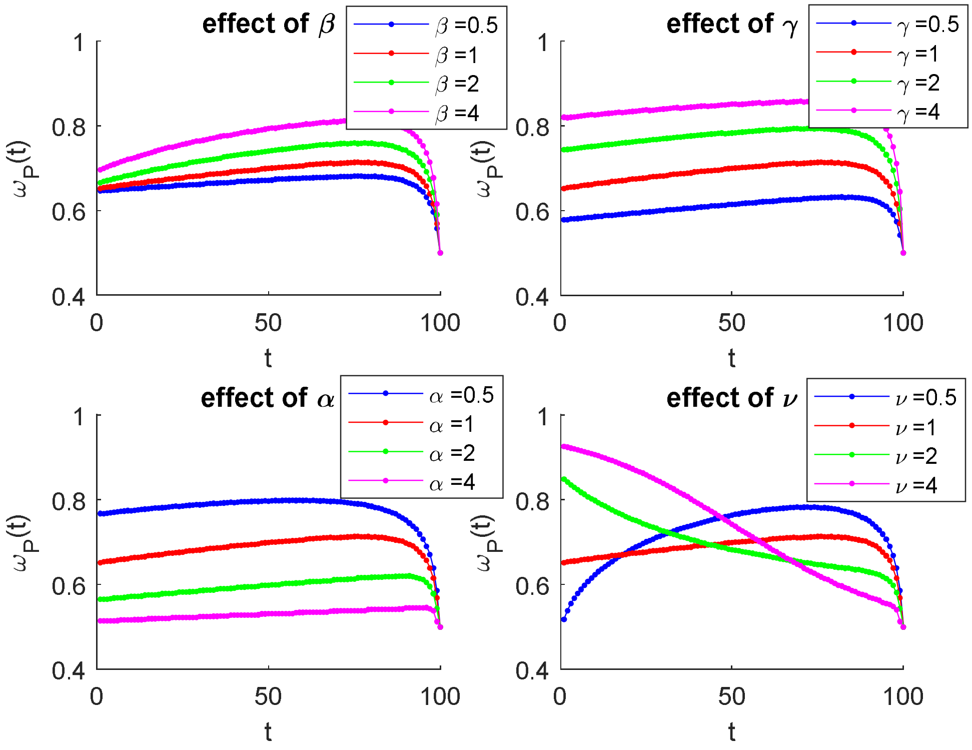

Figure 8: Impact of decision parameters on oMCD's optimal confidence threshold dynamics. Upper-left panel: Effect of type #1 effort efficacy. Optimal confidence threshold (y-axis, black dots) is plotted against decision time (x-axis), for different β levels (color code). **Upper-right panel**: Effect of type #2 effort efficacy, same format. **Lower-left panel**: Effect of unitary effort cost, same format. **Lower-right panel**: Effect of cost power, same format.

The net effect of increasing effort efficacy (either type #1 or type #2) is to increase the absolute confidence threshold. In other terms, the demand for confidence increases with effort efficacy. In contrast, the demand for confidence decreases with unitary effort cost. Note that the effect of increasing decision importance is exactly the same as that of decreasing unitary effort cost (not shown). Importantly, the shape of the confidence threshold dynamics is approximately invariant to changes in effort efficacy or unitary effort cost (here, the threshold globally increases over most admissible decision times, until it begins to fall when approaching the time horizon).

The only parameter that eventually changes the qualitative dynamics of oMCD's optimal confidence threshold is the cost power (cf. lower-left panel in Fig. 8). In brief, increasing the cost power tends to decrease the initial slope of oMCD's confidence threshold dynamics.

Here, the latter eventually falls below zero (i.e. the confidence threshold decreases with decision time) when the effort cost becomes superlinear (ν>1). Note that this happens without changing the dynamics of expected confidence (as when changing the unitary effort cost α). In other terms, the shape of the relationship between decision time and confidence is, for the most part, independent from the inner workings of the underlying decision system.

In conclusion, although oMCD relies on the same parameters than prospective MCD, it can produce a wider range of confidence/RT relationships (in particular, the latter can be non-monotonic, monotonically concave or monotonically convex).

3. Does MCD reproduce established empirical results?

As we highlighted before, MCD is compatible with most standard decision processes. However, the inner workings of value representation updates determine the choice that is made. This is important, since some of the decision features may depend upon, e.g., whether the system eventually arrives at a choice that is consistent with the comparison of options' values or not (in case stochastic perturbations alter the decision rule). Inspecting these kinds of effects thus requires committing to further assumptions regarding decision processes. In what follows, we perform Monte-Carlo simulations under the two above decision process scenarios: namely, the *ideal observer* model and the *attribute integration* model.

Let us first consider the ideal observer model. First, we simulate 1000 stochastic dynamics of Bayesian value belief updates according to Equation (20), having set the parameters of the ideal observer model as follows: $R=1$, $\alpha=0.1$, $\nu=2$, $\sigma_0=10$, $\mu_0=0$, $\Sigma=100$, and sampling a hidden value signal V (per trial) under the ideal observer's prior belief. Second, we identify the oMCD-optimal confidence threshold dynamics, having set the effort efficacy parameters to their analytical approximation, as given in Equation (22) and related derivations. We then store the ensuing decision times and their associated decision confidence, as well as the choice of the ideal observer (as given by the comparison of value modes at decision time). Figure 9 below summarizes the results of this Monte-Carlo simulations series.

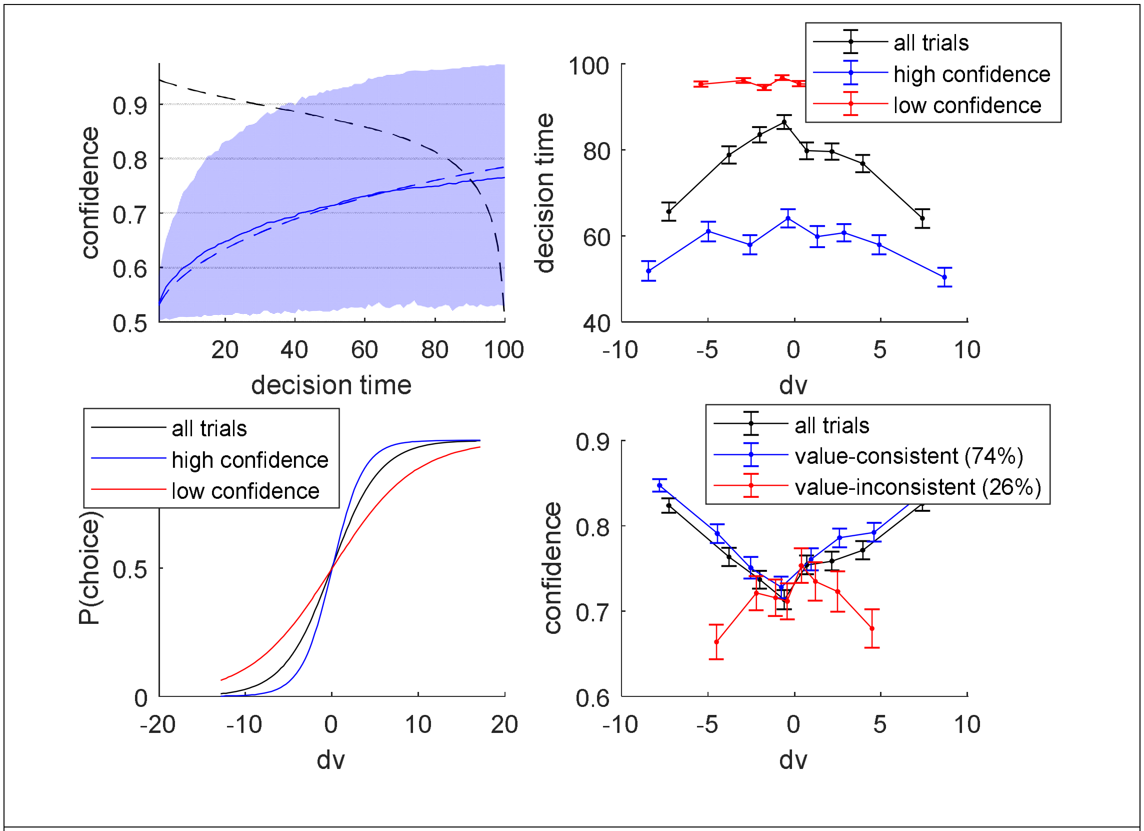

Figure 9: The ideal observer decision process. Upper-left panel: The blue line and shaded area depict the mean and standard deviation of the observer's confidence (under the 500 Monte-Carlo simulations), respectively. The blue dotted line shows the expected confidence under the corresponding MCD approximation, and the black dotted line shows the oMCD-optimal confidence threshold. **Upper-right panel**: Decision time (y-axis) is plotted against the difference in hidden option values (x-axis), for all trials (black), high-confidence trials (blue) and low-confidence trials (red), respectively. **Lower-left panel**: The probability of choosing the first option (y-axis) is plotted against the difference in hidden option values (x-axis), for all trials (black), high-confidence trials (blue) and low-confidence trials (red), respectively. **Lower-right panel**: Choice confidence (y-axis) is plotted against the difference in hidden option values (x-axis), for all trials (black), value-consistent trials (blue) and value-inconsistent trials (red), respectively.

First, one can see that the MCD approximation of within-trial choice confidence dynamics is relatively accurate (upper-left panel), and smoothly trades errors at early and late decision times. Second, on average, decision time decreases with the absolute difference in hidden option values (cf. black line in upper-right panel). This is a standard result in empirical studies of value-based decision making (De Martino et al., 2012; Milosavljevic et al., 2010; Rangel et al., 2008). Third, we reproduce most known relationships between decision times, confidence and choice consistency that are reported in the existing literature (De Martino et al., 2012; Lee & Daunizeau, 2021). In particular, above and beyond the effect of decision value, decision time decreases when choice confidence increases (cf. blue and red lines in upper-right panel). This, in fact, derives from shape of the oMCD confidence threshold dynamics (cf. Figure 8). Also, the consistency of choice with value is higher for high-confidence choices than for low-confidence choices (lower-left panel). This observation derives from performing a logistic regression of choice against hidden value, when splitting trials according to whether they yield a high or a low confidence (De Martino et al., 2012). Finally, on average, choice confidence decreases with the absolute difference in hidden option values (cf. black line in lower-right panel). Note that the oMCD framework also predicts that confidence is higher for choices that are consistent with the comparison of hidden values than for inconsistent choices (cf. red and blue lines in lower-right panel). This suggests that MCD possesses some level of metacognitive sensitivity (Fleming & Lau, 2014), i.e. it reports lower confidence when making a decision that is at odds with the hidden (unknown) value. In addition, when focusing on choices that are inconsistent with the comparison of hidden values, the impact of value difference on confidence reverses, i.e. choice confidence actually *decreases* with the absolute difference in hidden values. This extends known results in the context of perceptual decision making (Kepecs et al., 2008). We note that the latter results actually depend upon the effort cost parameters. In particular, metacognitive sensitivity tends to decrease in parameter regimes where the dynamics of oMCD confidence thresholds stop the decisions very early (e.g. low cost power and high unitary effort cost).

Let us now consider the attribute integration model. First, we simulate 1000 stochastic dynamics of value construction dynamics by attribute integration according to Equation (24), having set the model parameters to yield a similar rate of value-consistent choices than for the above ideal observer case ($R=1$, $\alpha=2$, $v=3$, $k=20$, $\eta_k=1$, $\varsigma_k=10$), and sampling a permutation order for attribute-disclosing events at random for each trial. Second, we identify the oMCD-optimal confidence threshold dynamics, having set the effort efficacy parameters to their analytical approximation, as given in Equation (26) and related derivations. We then store the ensuing decision times and their associated decision confidence, as well as the choice of the attribute-integration decision system (as given by the comparison of value modes at decision time). Figure 10 below summarizes the results of this Monte-Carlo simulations series.

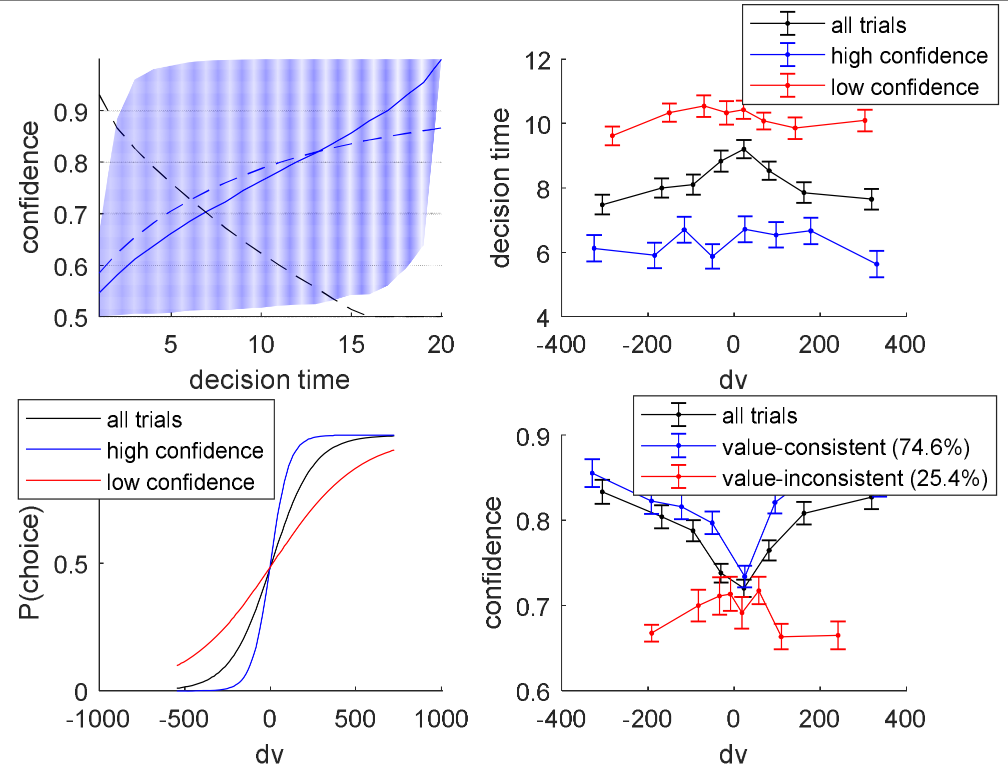

Figure 9: The attribute-integration decision process. Same format as Figure 9.

In brief, one can see that we qualitatively reproduce the above relationships between decision time, confidence and choice consistency. This is important, since this means that these relationships tend to generalize across different decision processes. However, this equivalence is only qualitative, and does not always hold. For example, reducing the unitary effort cost eventually renders the oMCD confidence threshold dynamics concave. Under the attribute-integration scenario, this reverses the impact of the difference in option values onto confidence for value-inconsistent choices back again. This does not seem to happen under the ideal observer scenario.

Discussion

In this work, we have presented the online/reactive metacognitive control of decisions or oMCD. We have compared it to its prospective MCD variant, and highlighted its main properties, when coupled with different underlying decision systems (in particular: the ideal observer and the attribute integration cases).

One of the main assumptions behind MCD is that mental effort investments are regulated by a unique controller that needs to operate under agnostic assumptions about the inner workings of the underlying decision system. Here, we have shown that the confidence gains of very different decision processes can be approximated using the same computational architecture, which relies on calibrating simple effort efficacy parameters. We did not, however, investigate the possibility that effort costs may be qualitatively different for different decision processes.

Recall that the notion of (mental) effort cost was central to the early definition of automatic versus controlled processing, with the former described as easy and effortless, and the latter as effortful (Schneider & Shiffrin, 1977). In line with this idea, we think of processing value-relevant information as involving the investment of (limited) cognitive resources, which may eventually override default/automatic preferences. Now, three main assumptions have been laid out for explaining the cost of controlled processes. As we will see, these suggest distinct neurocognitive mechanisms by which effort may operate.

First, mental effort may exhaust biological (e.g., metabolic) energy (Baumeister et al., 1998) that results from the additional complexity of controlled processes. Although attractive, this assumption is difficult to reconcile with the observation that some automatic/effortless processes (e.g., face recognition) arguably involve more complex computations than some controlled processes (e.g., one-digit arithmetic). Second, engaging control resources on a given task necessarily means foregoing the benefits of engaging these resources on other tasks. This induces an *opportunity cost* (Kurzban et al., 2013) that increases with the time spent on the task. However, this assumption falls short of an explanation for why effort cost increases with effort intensity, irrespective of effort duration (Blain et al., 2016). Third, effort signals may proxy detrimental interferences that would result from the multiple and conflictual loading of shared neural resources. Such conflict-induced signal would then trigger cognitive control, which disengages bottlenecks of brain information processing pathways from non-instrumental multitasking demands (Botvinick et al., 2001). Here effort regulation is essentially a reactive process, whereby cognitive control is engaged until target task demands are met. In this scenario, effort sensations *feel like a cost*, because they co-occur with situations in which cognitive control strives against high resistance (cf. difficult decisions). Under the framework of MCD, a possibility is that decision making requires the active maintenance of multiple value representations that tend to interfere with each other (e.g., because they involve the same neural population). In this case, cognitive control may alter the associated neural code with the aim of dampening these interferences. We will test these ideas using artificial neural network models of MCD in forthcoming publications.

Finally, we note that the architecture of oMCD model lends itself nicely to other kinds of decision processes: in particular, perceptual or evidence-based decisions. In this context, decision confidence can be defined (somewhat more straightforwardly) as the subjective probability of being correct (Pouget et al., 2016). Nevertheless, as long as (i) decision confidence can be described using some variant of Equation (2), and (ii) the mean and variance of the relevant perceived quantity follow Equations (3) and (4), the overall MCD architecture will provide an optimal solution to the resource allocation problem. Note that when describing perceptual detection or discrimination processes using some form of *ideal observer* scenario, those conditions are trivially satisfied (Daunizeau et al., 2010; Drugowitsch et al., 2012). This would also hold for perceptual categorization processes, which may rather resemble attribute integration scenarios (Summerfield & Tsetsos, 2012).
In fact, oMCD's potential generalization ability derives from its agnostic stance regarding the nature of information processing that takes place in the underlying decision system. This is

also why oMCD can in principle be extended to describe the metacognitive control of other kinds of cognitive processes (e.g., reasoning or memory encoding/retrieval). In this context, an interesting avenue is to consider the impact of metacognitive adaptation onto the regulation of mental effort. Note that, because MCD only relies upon confidence monitoring to control effort, it requires a systematic calibration (in terms of, e.g., effort efficacy estimation) to guaranty the quasi-optimality of resource allocation. As we highlighted before, we expect such calibration to converge very quickly (e.g., over a few training trials). This is because effort efficacies can be learned from within-trial confidence dynamics. Nevertheless, whether and how this specific kind of metacognitive adaptation eventually modifies mental effort regulation is virtually unknown.

References

Baumeister, R. F., Bratslavsky, E., Muraven, M., & Tice, D. M. (1998). Ego depletion: Is the active self a limited resource? *Journal of Personality and Social Psychology*, *74*(5), 12521265.

Bellman, R. (1957). *Dynamic Programming* (Princeton University Press). https://press.princeton.edu/books/paperback/9780691146683/dynamic-programming

Blain, B., Hollard, G., & Pessiglione, M. (2016). Neural mechanisms underlying the impact of daylong cognitive work on economic decisions. *Proceedings of the National Academy of Sciences*, *113*(25), 69676972. https://doi.org/10.1073/pnas.1520527113

Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S., & Cohen, J. D. (2001). Conflict monitoring and cognitive control. *Psychological Review*, *108*(3), 624652.

Daunizeau, J. (2017). Semi-analytical approximations to statistical moments of sigmoid and softmax mappings of normal variables. *arXiv:1703.00091 [q-bio, stat]*. http://arxiv.org/abs/1703.00091

Daunizeau, J., den Ouden, H. E. M., Pessiglione, M., Kiebel, S. J., Stephan, K. E., & Friston, K. J. (2010). Observing the observer (I): Meta-bayesian models of learning and decision-making. *PloS one*, *5*(12), e15554. https://doi.org/10.1371/journal.pone.0015554

De Martino, B., Fleming, S. M., Garrett, N., & Dolan, R. J. (2012). Confidence in value-based choice. *Nature Neuroscience*, *16*(1), 105110. https://doi.org/10.1038/nn.3279

Drugowitsch, J., Moreno-Bote, R., Churchland, A. K., Shadlen, M. N., & Pouget, A. (2012). The Cost of Accumulating Evidence in Perceptual Decision Making. *Journal of Neuroscience*, *32*(11), 36123628. https://doi.org/10.1523/JNEUROSCI.4010-11.2012

Feinberg, E. A., & Shwartz, A. (2012). *Handbook of Markov Decision Processes: Methods and Applications*. Springer Science & Business Media.

Fleming, S. M., & Lau, H. C. (2014). How to measure metacognition. *Frontiers in Human Neuroscience*, *8*. https://doi.org/10.3389/fnhum.2014.00443

Griffiths, T. L., Lieder, F., & Goodman, N. D. (2015). Rational Use of Cognitive Resources: Levels of Analysis Between the Computational and the Algorithmic. *Topics in Cognitive Science*, *7*(2), 217–229.

Kahneman, D. (2011). *Thinking, Fast and Slow*. Macmillan.

Kepecs, A., Uchida, N., Zariwala, H. A., & Mainen, Z. F. (2008). Neural correlates, computation and behavioural impact of decision confidence. *Nature*, *455*(7210), 227231. https://doi.org/10.1038/nature07200

Kurzban, R., Duckworth, A., Kable, J. W., & Myers, J. (2013). An opportunity cost model of subjective effort and task performance. *Behavioral and Brain Sciences*, *36*(06), 661679. https://doi.org/10.1017/S0140525X12003196

Lebreton, M., Abitbol, R., Daunizeau, J., & Pessiglione, M. (2015). Automatic integration of confidence in the brain valuation signal. *Nature Neuroscience*, *18*(8), 11591167. https://doi.org/10.1038/nn.4064

Lebreton, M., Jorge, S., Michel, V., Thirion, B., & Pessiglione, M. (2009). An Automatic Valuation System in the Human Brain: Evidence from Functional Neuroimaging. *Neuron*, *64*(3), 431439. https://doi.org/10.1016/j.neuron.2009.09.040

Lee, D. G., & Daunizeau, J. (2021). Trading mental effort for confidence in the metacognitive control of value-based decision-making. *eLife*, *10*, e63282. https://doi.org/10.7554/eLife.63282

Lieder, F., Shenhav, A., Musslick, S., & Griffiths, T. L. (2018). Rational metareasoning and the plasticity of cognitive control. *PLOS Computational Biology*, *14*(4), e1006043. https://doi.org/10.1371/journal.pcbi.1006043

Lopez-Persem, A., Domenech, P., & Pessiglione, M. (2016). How prior preferences determine decision-making frames and biases in the human brain. *ELife*, *5*, e20317. https://doi.org/10.7554/eLife.20317

Milosavljevic, M., Malmaud, J., Huth, A., Koch, C., & Rangel, A. (2010). The drift diffusion model can account for value-based choice response times under high and low time pressure. *Judgment and Decision Making*, *5*, 437–449.

Musslick, S., Shenhav, A., Botvinick, M., & D Cohen, J. (2015, juin 7). *A Computational Model of Control Allocation based on the Expected Value of Control*. Conference: Reinforcement Learning and Decision Making 2015.

Papadimitriou, C. H., & Tsitsiklis, J. N. (1987). The Complexity of Markov Decision Processes. *Mathematics of Operations Research*, *12*(3), 441–450. https://doi.org/10.1287/moor.12.3.441

Pouget, A., Drugowitsch, J., & Kepecs, A. (2016). Confidence and certainty : Distinct probabilistic quantities for different goals. *Nature Neuroscience*, *19*(3), 366–374. https://doi.org/10.1038/nn.4240

Rangel, A., Camerer, C., & Montague, P. R. (2008). A framework for studying the neurobiology of value-based decision making. *Nature Reviews Neuroscience*, *9*(7), 545–556. https://doi.org/10.1038/nrn2357

Schneider, W., & Shiffrin, R. M. (1977). Controlled and automatic human information processing : I. Detection, search, and attention. *Psychological Review*, *84*(1), 1–66. https://doi.org/10.1037/0033-295X.84.1.1

Shenhav, A., Botvinick, M. M., & Cohen, J. D. (2013). The Expected Value of Control : An Integrative Theory of Anterior Cingulate Cortex Function. *Neuron*, *79*(2), 217–240. https://doi.org/10.1016/j.neuron.2013.07.007

Shenhav, A., Musslick, S., Lieder, F., Kool, W., Griffiths, T. L., Cohen, J. D., & Botvinick, M. M. (2017). Toward a Rational and Mechanistic Account of Mental Effort. *Annual Review of Neuroscience*, *40*, 99–124. https://doi.org/10.1146/annurev-neuro-072116-031526

Summerfield, C., & Tsetsos, K. (2012). Building Bridges between Perceptual and Economic Decision-Making: Neural and Computational Mechanisms. *Frontiers in Neuroscience*, *6*. https://www.frontiersin.org/articles/10.3389/fnins.2012.00070

Tajima, S., Drugowitsch, J., Patel, N., & Pouget, A. (2019a). Optimal policy for multi-alternative decisions. *Nature Neuroscience*, *22*(9), 15031511. https://doi.org/10.1038/s41593-019-0453-9

Tajima, S., Drugowitsch, J., Patel, N., & Pouget, A. (2019b). Optimal policy for multi-alternative decisions. *Nature Neuroscience*, *22*(9), 15031511. https://doi.org/10.1038/s41593-019-0453-9

Tajima, S., Drugowitsch, J., & Pouget, A. (2016). Optimal policy for value-based decision-making. *Nature Communications*, *7*, 12400. https://doi.org/10.1038/ncomms12400