An information-theoretic perspective on the costs of cognition

Alexandre Zénon, Oleg Solopchuk, Giovanni Pezzulo

An information-theoretic perspective on the costs of cognition

Zénon Alexandre*[1,2], Solopchuk Oleg[1,2], Pezzulo Giovanni[3]

[1]Institut de Neuroscience Cognitive et Intégrative d'Aquitaine, Université de Bordeaux, France

[2]Institute of Neuroscience, Université catholique de Louvain, Brussels, Belgium

[3]Institute of Cognitive Sciences and Technologies, National Research Council, Via San Martino della Battaglia 44, 00185, Rome, Italy.

*Corresponding author:

Dr. Alexandre Zénon

E-mail: alexandre.zenon@u-bordeaux.fr

In statistics and machine learning, model accuracy is traded off with complexity, which can be viewed as the amount of information extracted from the data. Here, we discuss how cognitive costs can be expressed in terms of similar information costs, i.e. as a function of the amount of information required to update a person's prior knowledge (or internal model) to effectively solve a task. We then examine the theoretical consequences that ensue from this assumption. This framework naturally explains why some tasks – for example, unfamiliar or dual tasks – are costly and permits to quantify these costs using information-theoretic measures. Finally, we discuss brain implementation of this principle and show that subjective cognitive costs can originate either from local or global capacity limitations on information processing or from increased rate of metabolic alterations. These views shed light on the potential adaptive value of cost-avoidance mechanisms.

Keywords: cognitive effort, information theory, active inference, predictive coding, efficient coding, computational neuroscience

1. Introduction

Demanding cognitive tasks, such as mental arithmetic, are strongly aversive: we tend to avoid partaking in such tasks and they lead to unpleasant subjective feeling of mental exertion (Inzlicht et al., 2015). Various studies have revealed that we take into consideration these cognitive costs when deciding whether or not to engage in a task (Benoit et al., 2017; Kool et al., 2010; Manohar et al., 2015; Schmidt et al., 2012; Westbrook et al., 2013; Westbrook and Braver, 2015). Furthermore, prolonged performance of demanding tasks leads to cognitive fatigue, which is characterized by a subjective dimension – i.e. feeling of exhaustion, impression of worsened ability and decreased willingness to engage in mental activities (Hockey, 2011; van der Linden et al., 2003) – and an objective dimension, with an actual decrease of task performance (Bailey et al., 2007; Tanaka, 2015; van der Linden et al., 2003). However, it is still unclear what is the origin of cognitive costs (i.e., what is costly about cognitive processing?), how to specify them quantitatively, and whether cognitive costs and cognitive fatigue have some adaptive value.

In this article, we address this set of questions from the angle of information theory, by establishing a connection between cognitive costs and computational or information measures, e.g. the amount of information required to update a person's prior knowledge. Throughout the article, we will use the term "cognitive costs" to refer to the percept of effort and the task avoidance associated with cognitive tasks. Conversely, "computational costs" will designate the cost of cognitive tasks from the point of view of computational theory or

artificial intelligence, and "information costs" represent one particular instance in which computational costs are framed in terms of information theory.

2. The information cost of cognitive processes

Recent advances in artificial intelligence and computational neuroscience have led to formalization of cognition as a bounded rationality process (Friston, 2010; Kingma and Welling, 2013; Ortega and Braun, 2013; Tishby et al., 2000; Tkačik and Bialek, 2014). According to this view, rather than aiming systematically at the optimal solution to computational problems, cognitive processes trade off performance with computational costs. Remarkably, the way computational costs are formalized across these different studies is very consistent, despite their different approaches. In fact, whether one starts from an inference problem, in which the evidence for a model is maximized given some data (Genewein et al., 2015; Kingma and Welling, 2013; Tishby et al., 2000), or whether one is more generally attempting to minimize the entropy of future states (Friston, 2010), or whether one takes a decision making perspective, in which expected utility is maximized (Ortega et al., 2015), or even from the point of view of thermodynamics (Ortega and Braun, 2013; Sengupta et al., 2013), computational cost is framed as a measure of divergence between an initial belief (or prior probability distribution over a variable of interest *x,* such as expected reward) and an updated belief (or posterior probability distribution over the same variable *x*) obtained after receiving new data (Donnarumma et al., 2016; Kappen et al., 2012; Maisto et al., 2016, 2015; Polani, 2009; Stoianov et al., 2016; Tishby and Polani, 2011). This measure of difference between probability distributions, called the Kullback-Leibler (KL) divergence, represents the amount of information one needs to collect in order to update the prior to the posterior ($KL(P\|Q)=\sum P\log\frac{P}{Q}$ for probability distributions *P* and *Q*). Here, "amount of information" is

meant in the sense of Shannon's definition of information in terms of *surprisal* or the negative log probability of the data (Shannon, 1948). In other words, expected data provides little information while unexpected data is very informative. If one wants to encode such data without error, the number of binary symbols that will be needed is at least equal to the average surprisal, or *entropy* of the data (entropy is often represented by the letter *H*: $H(P) = -\sum P \log_2 P$).

Whether we are considering an inference problem (identifying latent causes of observations) or a decision making problem (deciding what to do), cognitive activity can be viewed as a process in which input data allows us to refine our previous assumption (about the most probable cause or the best action) to a new, more accurate belief and the cost of this process, from the perspective of information theory, corresponds to the reduction in the entropy that it causes.

The question of the information cost of cognitive control – i.e. the selection of appropriate behaviour in the face of environmental stimulation, on the basis of internal goals (Miller and Cohen, 2001; Pezzulo et al., 2018) - has been addressed for more than 50 years, pioneered, separately, by Hick and Hyman (Hick, 1952; Hyman, 1953). In these early works, information cost was framed in terms of the entropy of the response choice (we will refer henceforth to response choices by the variable name *y*): $H(p(y))$. A typical example is the digit-key association task, in which participants have to press the key that corresponds to the digit they see on the screen. In this task, participants need to update their prior distribution of responses $p_0(y)$ to a posterior distribution $p(y|x)$ (read as the probability of *y* given *x*, where *x* refers to the input data) in which all the probability mass is in one stimulus-response association (see Figure 1, dark grey). If there are 4 possible digits and their probabilities of occurrence are equal, the KL divergence between $p_0(y)$ and $p(y|x)$ is equal to $H(p_0(y))$, which in the present case is -log(1/4) or log(4). Strikingly, what Hick and Hyman showed is that

reaction time is a linear function of $H(p_0(y))$ (Hick, 1952; Hyman, 1953), confirming that it is an accurate measure of information cost in this context and suggesting that the rate of information for this task is constant (i.e. the number of bits processed per unit of time is constant).

However, the uncertainty in the response is clearly not the sole determinant of task complexity. In the task above, if one varies the number of stimuli associated to each button press while keeping the number of responses constant (e.g. 2 stimuli become associated to the same button presses, leading to 8 stimuli for 4 buttons), $H(p_0(y))$ remains constant but reaction times increase, and they do so linearly with $H(p_0(x))$ (Wifall et al., 2016). A similar increase in reaction time was reported in other tasks in which the complexity of the stimulus varies, while the number of response choices remains constant (Fan, 2014; Fan et al., 2008). To explain these findings, it is necessary to extend the aforementioned information theoretic framework, such that information costs depend on both $p(x)$ and $p(y)$.

It is worth noting that x stands for sensory data coming from the outside world and not an internal variable of the agent – hence it cannot be used directly for our formalization. For this, we need to introduce an auxiliary variable x' that would stand for the internal representation of x, which the system uses to choose the action y: $x \to x' \to y$. The addition of this intermediate variable x' affords different levels of compression of the input x (x' can extract more or less of the information that is available in x), which is an important feature for a model of cognition (Grau-Moya and Braun, 2015; Park and Pillow, 2017; Tishby et al., 2000).

Therefore, information cost $\mathcal{C}$ becomes the KL divergence between the prior internal representation of the input $p_0(x')$ and its posterior after observing the outside world $p(x'|x)$, to which we need to add the KL divergence between the prior distribution of responses $p_0(y)$ and their posterior distribution $p(y|x')$:

$$\mathcal{C} = KL(p(x'|x) \| p_0(x')) + KL(p(y|x') \| p_0(y)) = \sum_{x'} p(x'|x) \log \frac{p(x'|x)}{p_0(x')} + \sum_{y} p(y|x') \log \frac{p(y|x')}{p_0(y)}$$

It is worth adding that if one considers the average of these information costs across many trials, and assuming usage of an optimal, marginal prior (see below and Tishby et al., 2000), the total information cost becomes:

$$\mathcal{C} = \sum_{x} p(x) KL(p(x'|x) \| p_{opt}(x')) + \sum_{x'} p(x') KL(p(y|x') \| p_{opt}(y)) = I(x;x') + I(x';y)$$

where *I(x;x')* refers to the *mutual information* between *x* and *x'*, which indicates the reduction in the entropy of *x'* after observing *x* (which is closely related to the notion of epistemic value in active inference; see Friston et al., 2015), while *I(x';y)* represents the reduction in entropy of the response given some internal representation of the input *x'*. We can see from the above formula that this framework implements the property of $\mathcal{C}$ that we wanted: its dependence on both the complexity of the input data representation *x'* and the complexity of the responses *y*. Finally, in order to account for the cognitive costs associated with classic tasks such as the Stroop task for example, we need to make one last addition to our framework, in agreement with earlier proposals by Koechlin and Summerfield (2007). In the Stroop task, subjects must either read a colour word or name the colour of the ink with which the word is written. When both sources of information are incongruent (e.g. the word "blue" is written with red ink), it is more difficult to name the ink colour than to read the word (MacLeod, 1991). This indicates that the stimulus triggers an action (reading the word) that is independent of the task context and that interferes with the response instructed by the task (naming the ink colour) (Cohen et al., 1990). In order to account for this, we need to add a new variable *T* that represents the context of the task. The stimulus then triggers a default, automatic conditional distribution of responses *p(y|x')* (that is thus independent of context), which is finally updated to the final, context-dependent distribution: *p(y|x',T)*. The optimal automatic response

distribution is the marginal of the final response distribution (i.e. the final distribution averaged over all contexts):

$$p(y \mid x') = \sum_T p(y \mid x', T)\, p(T)$$

Our final formula for total information cost with optimal priors becomes (see also Figure 2):

$$\mathcal{C} = \underbrace{I(x;x')}_{\text{perceptual cost}} + \underbrace{I(x';y)}_{\text{cost of automatic process}} + \underbrace{I(T;y \mid x')}_{\text{cost of context-dependent process}} \quad [1]$$

Note that under this formula, $\mathcal{C}$ should not be interpreted as a general measure of the information cost of a task since it depends on the pattern of response y. If y is taken to be equal to the optimal response, then $\mathcal{C}$ can be interpreted as the information cost of performing the task optimally. More generally, there will be a trade-off between $\mathcal{C}$ and performance. This will be discussed in more detail in section 3.5.

3. Predictions of the framework

So far, we have shown that information cost of a cognitive task can be framed as the sum of three terms: the mutual information between inputs and their internal representations; the mutual information between internal representations and automatic responses; and the mutual information between contextual information and automatic responses. Now we explain in more details how to apply this framework in practice and discuss its predictions in terms of expected cost of different types of tasks. This theory predicts that certain kinds of tasks - those that have *many degrees of freedom*, are *unfamiliar*, necessitate to go *against natural*

1 Cost of the context-dependent process:

$$\sum_T p(T) KL(p(y \mid x', T) \,\|\, p(y \mid x')) = \sum_T p(T) \sum_{x'} p(x') \sum_y p(y \mid x', T) \log \frac{p(y \mid x', T)}{p(y \mid x')}$$

$$= \sum_{x'} p(x') \sum_{y,T} p(y, T \mid x') \log \frac{p(y \mid x', T)\, p(T)}{p(y \mid x')\, p(T)} = \sum_{x'} p(x') \sum_{y,T} p(y, T \mid x') \log \frac{p(y, T \mid x')}{p(y \mid x')\, p(T)} = I(T; y \mid x')$$

biases, have *variable statistical structure* or *low signal to noise ratios* - will lead to large information costs.

3.1 The costs of tasks that have many degrees of freedom

Tasks that have *many degrees of freedom*, or equivalently, a wide probability distribution of state-action combinations, are expected to be cognitively costly under the proposed information theoretical framework, as they imply low, widely spread prior probabilities - and thus significant information costs to update the priors (see Figure 3). Arithmetic tasks, chess games or creative writing, which are well known for being cognitively demanding (Hess and Polt, 1964; Kellogg, 1987; Marshall, 2002; Westbrook and Braver, 2015b) all assign a small prior probability mass for each possible decision and hence, lead to large divergence with the final posterior obtained when the choice has been made. Intuitively, this would be equivalent to having a very wide response space in Figure 1, with the same, small probability for each possible stimulus-response association. Similarly, tasks that demand a deep contextualisation of the stimulus-response associations (e.g. learning to navigate in a complex maze) will lead to multidimensional prior distributions whose space can inflate very fast, also leading to fast increase in complexity.

Interestingly, the complexity of the environment x has no impact on cognitive costs. Only the complexity of x', its internal representation, will affect information cost C. This can be understood intuitively if one considers, for example, that during the digit-key association task, the digit presented on the screen could be represented as an image of arbitrary size and complexity. Indeed, all x' needs to encode about x is the identity of the digit (i.e. 1 to 4) and what will matter for task difficulty is only the probability of occurrence of that digit (see above). An interesting implication is that – as known since the beginning of artificial intelligence – the way one encodes or represents the task drastically affects the complexity

(or even the possibility) of solving it (Minsky, 1961; Simon, 1956). We will return to the issue of information compression below.

3.2 The costs of novel or unfamiliar tasks

A similar issue arises with *unfamiliar tasks*, that is, tasks in which the statistical structure of the sensory states, state transition probabilities (conditional on the performed actions) or action policies (e.g. sequences of motor actions) are poorly known (see Figure 4). This lack of knowledge of statistical properties of the task leads to non-optimal encoding and large information costs. This is because participants will have to start from uninformative prior distributions $p_0(x')$, $p_0(y)$ and $p_0(y|x')$ that may be far from the true marginal distributions of the task, which they will have to learn across trial repetitions (Genewein et al., 2015; Tishby et al., 2000). In this case the additional cost of starting with the wrong priors can be formalized as:

$$\Delta\mathcal{C} = \underbrace{KL(p_{opt}(x') \,\|\, p_0(x')) + KL(p_{opt}(y) \,\|\, p_0(y)) + KL(p_{opt}(y \mid x') \,\|\, p_0(y \mid x'))}_{\text{cost of imperfect prior belief}} \quad [2]$$

Importantly, here we are assuming that participants have understood the rules of the task and know how to perform it correctly. Therefore, this additional cost of imperfect priors represents the extra information participants need to process because of their poor assumption of task statistics - $p_0(x')$, $p_0(y)$ and $p_0(y|x')$ - not the cost of learning the task rules. This corresponds to the notion of *cross-entropy* in information theory: the number of symbols needed to encode data when using the wrong encoding scheme - i.e. one that is based on the wrong probability distribution - is always larger than entropy (the number of symbols

2 Derivation for the perceptual cost $I(x;x')$:

$$\Delta\mathcal{C} = \sum_x p(x) \sum_{x'} p(x' \mid x) \log \frac{p(x' \mid x)}{p_0(x')} - \sum_x p(x) \sum_{x'} p(x' \mid x) \log \frac{p(x' \mid x)}{p(x')}$$

$$= \sum_x p(x) \sum_{x'} p(x' \mid x) \log \frac{p(x')}{p_0(x')} = \sum_{x'} \sum_x p(x, x') \log \frac{p(x')}{p_0(x')} = KL(p(x') \,\|\, p_0(x'))$$

necessary when the correct distribution is assumed). In behavioural terms, this corresponds to the well-known effect of training on reaction time (Teichner and Krebs, 1974), subjective effort (Mykityshyn et al., 2002), or pupil size (Hyönä et al., 1995; Recarte and Nunes, 2000; Solopchuk et al., 2016), regarded as a reliable index of effort (Beatty and Lucero-Wagoner, 2000; van der Wel and van Steenbergen, 2018). It is also interesting to note that training typically leads to decreased (not increased) brain activation - plausibly, by increased knowledge of task contingencies and the ensuing decrease of the metabolic costs associated with task-related information processing (Solopchuk et al., 2017; Wiestler and Diedrichsen, 2013).

3.3 The costs of counteracting priors or default policies

Following the same reasoning as above, our framework also predicts that tasks that require *counteracting deep priors* or *default policies* would be particularly demanding (see Figure 5). Indeed, if priors are not just uninformative, as assumed in the previous paragraph, but also counter-productive, assuming strong statistical relationships that are no longer true, the KL divergence with the true joint distribution will be even larger.

Similarly to the case of unfamiliar tasks exposed in section 3.2, the extra cost of using the wrong priors can be formalized as:

$$KL(p_{opt}(x') \,\|\, p_0(x')) + KL(p_{opt}(y) \,\|\, p_0(y)) + KL(p_{opt}(y \mid x') \,\|\, p_0(y \mid x'))$$

However, whereas the priors p_0 are easy to estimate in the case of novel tasks, since they are simply non-informative (e.g. uniform distributions), in the present case, they can take many different forms and will have usually to be inferred on the basis of participants' behaviour.

In the example of the Stroop task, the prior on the response issued from the automatic process $p(y|x')$ will be counterproductive in cases of incongruent colour naming task. One plausible

way to model this incorrect prior is by taking the marginal $\sum_T p(y \mid x', T) p_0(T)$ in which the probability of the word-reading context $p_0(T=word\text{-}reading)$ largely dominates the probability of the colour-naming context. This is incorrect in the sense that in the context of the Stroop task, word-reading is no longer more likely than colour-naming. This results in a marginal which favours word-reading responses, leading to extra processing costs in incongruent colour-reading trials (see Figure 5 A). The advantage of this approach is that cost depends only on the task structure $p(x,y)$ and on the assumed prior context probability $p_0(T)$. Here, for simplicity, we considered point estimates for $p_0(T)$, but Dirichlet distributions could also be used, with the advantage of associating a precision to the belief on context probabilities. Very high precisions would be associated with very rigid beliefs on context probabilities, leading to costs that would be relatively insensitive to training.

The present framework can explain why counteracting habits – or default responses to environmental stimuli – is so costly. Habitual behaviour is characterized by fast, automatic processing, low effort and lack of flexibility (Kahneman, 2011; Moors and De Houwer, 2006; Schneider and Chein, 2003). Under the present framework, with overtraining, the encoding of task-specific information follows so closely the statistical task structure that all the task-irrelevant information gets ignored. The ensuing cognitive processing is thus extremely efficient but also crucially dependent on the particular task contingencies that have been learned. Expected stimuli and their associated actions have very large prior probabilities $p_0(x')$, $p_0(y)$, $p_0(y|x')$ and are therefore encoded with minimal cost, while unexpected stimuli or actions have very low prior probabilities and are therefore very costly to encode. This implies that a person following habitual policies has lower costs to engage in familiar tasks but higher costs to engage in novel tasks. This impact of familiarity on cognitive cost could explain why novel environments (e.g. new places, new languages, new people, etc.) are

generally described as being more fatiguing than familiar ones, while natural, familiar environments would have, on the contrary, restoring effects (Kaplan and Berman, 2010).

Priors can also have a deeper meaning from the perspective of the active inference framework (Friston, 2010). Under that view, agents are equipped with hierarchical generative models and perform Bayesian inference; and have the general objective to minimize their free energy or, with some simplifications, their surprise - or the discrepancy between what they expect, based on their beliefs, and what they sense. Importantly, they can minimize their surprise in two ways: by changing their beliefs to make them more similar to what they sense about the world (i.e. perceptual processing) or by changing the world to make it more similar to their prior beliefs (i.e. using actions to fulfil one's own expectations). This duality is possible if one considers that active inference agents are hierarchically organized. While hierarchically lower prior beliefs might faithfully adapt to the external world, hierarchically deeper priors would prescribe what states an agent should achieve by acting (Friston et al., 2012; Pezzulo et al., 2015). These latter, deeper priors hence play the role of goals and motivational factors that are relatively less permeable to learning (i.e., in Bayesian terms, they have very high precision or inverse uncertainty) because they are key to survival. Indeed, a key statement of active inference is that biological agents need to minimize their long-term surprise in order to survive; if one thinks of these deep priors as describing the "good" states in an agent's ecological niche, minimizing surprise means that the agent should attempt to remain always close to these states. One example of deep (and perhaps hard-coded) prior is a homeostatic drive, such as the prior probability of body nutrients being within acceptable physiological range. A discrepancy (prediction error) between such deep prior (e.g., be satiated) and the current interoceptive sensations (e.g., feeling hungry) would not lead to the revision of the prior - since the prior is largely impermeable in virtue of having high precision. Instead, the prediction error would steer a cascade of predictions about the conditions that might restore

body nutrients (e.g., consuming food), which in turn would steer an adaptive policy or action sequence to fulfil these predictions (e.g., open the fridge and take some food). This formulation makes it apparent that a hungry active inference agent would assign a high probability to (predicted) states and policies associated to consuming food. Since the information theoretic perspective on the costs of cognition advanced here assumes that counteracting such high-probability states (or equivalently, pursuing low-probability states) has high information costs, any task that necessitates counteracting deep priors or their ensuing policies should lead to large information and cognitive costs (see Figure 5 B). This perspective may help to explain the effortful nature of self-control (Kool et al., 2010), which consists precisely in going against natural biases, as investigated in the large, but still very controversial literature on ego depletion (Hagger et al., 2016, 2010; Job et al., 2010; Kurzban et al., 2013; Muraven and Baumeister, 2000). The same arguments, based on strong (deep / homeostatic or shallow / habitual) priors can help understand the phenomenology associated to some pathological situations, such as Tourette syndrome and obsessive-compulsive disorders. These and other syndromes have been associated to strong priors that are, however, maladaptive and lead to inappropriate behaviour (Adams et al., 2013; Friston et al., 2014). Patients suffering from these disorders typically describe being able to overcome their (habitual) tics (Delorme et al., 2016) or compulsive behaviour but at the price of tremendous cognitive effort (Kawohl et al., 2009).

3.4 The costs of task switching and dual tasks

Another classic cause of cognitive effort is *task switching*. Changing from a task set to another is associated to a significant cost in performance, usually measured as increased reaction time (Wylie and Allport, 2000). Task switching is also accompanied with a

subjective cognitive effort cost (Apps et al., 2015; Kool et al., 2010) and leads to cognitive fatigue (Borragán et al., 2017). Reaction time costs in task switching have already been modelled by means of information theoretic approaches (Cooper et al., 2015). Here we propose to frame switching costs within our general framework. This can be done by considering that tasks A and B, for example, correspond to two different contexts *T* associated with different probabilities *p(T)*. Following one trial in task A, *p(T=A)* increases and the marginal $p(y|x') = \sum_T p(y|x',T)p(T)$ becomes closer to the correct response probabilities for task A: $p(y|x',T=A)$. However, and for the same reason, when the task switches to B, the context-dependent cost *I(T;y|x')* becomes larger. Therefore, the faster the participant learns the task structure, and the longer she/he is trained on that task, the more expensive the switch cost will be. Practically, this situation can be parametrized by the speed of learning of the context probability *p(T)*. This relationship between *p(T)* and information cost is illustrated in Figure 6, where we simulate an experiment in which participants must switch between two tasks every other trial. Repeated trials are associated with smaller information costs (Figure 6A), because the context distribution *p(T)* was updated in the previous trial in favour of the same context (Figure 6B).

The same arguments may apply to more mundane situations in which one is required to switch continuously between multiple tasks (multi-tasking); or to dual-task situations, in which one has either to maintain a sophisticated internal model where the probability distribution spans the contingencies of both tasks, or to rapidly and repeatedly switch between the tasks to be executed concurrently.

Interestingly, our model makes the specific prediction that switching tasks involving common stimuli (e.g. switching between colour-naming and word-reading Stroop task (Wylie and Allport, 2000)), should lead to larger costs than when switching between tasks with different stimuli, in agreement with the literature (Rubin and Meiran, 2005). This is because when

different stimuli are involved, the automatic process will lead to different response probabilities for each stimulus $p(y|x'_A)$ and $p(y|x'_B)$, decreasing interference between the tasks. Moreover, our model also naturally explains task set inertia: the observation that interference from task A persists long after switching to task B (Allport et al., 1994). Indeed, following the switch, $p(T)$ will update in each trial, getting larger and larger for task B, and smaller and smaller for task A, thereby progressively improving performance of task B.

3.5 Trade-off between performance and information rate

Another task feature that is well known to affect cognitive cost is *signal to noise ratio*. When the ratio between signal and noise in sensory data is low, performance decreases, pupil size increases and subjective cognitive effort increases (Manohar et al., 2015; Sarampalis et al., 2009; Zekveld et al., 2014). Since performance and information costs are subjected to a trade-off (Sims, 2016), increasing task performance implies larger information costs. Likewise, when sensory information is immersed in high noise, and one wants to maintain reasonable performance, one needs to raise the encoding precision by decreasing the level of compression of x into x' (the encoding of sensory data) and hence increase information costs by raising $I(x,x')$. Rate distortion theory provides a framework for addressing this type of problem, by describing the relation between minimal information rate (i.e. number of bits used per symbol encoded) and performance (or distortion) in a given information processing system (Shannon, 1959; Sims, 2016). Simply put, in order to minimize information cost, one should compress input data in order to discard information that is irrelevant to the task. For a given task, achieving null distortion (i.e. perfect performance) requires a minimal information rate that is equal to the entropy of the task (i.e. its average surprisal). Beyond this level of compression, relevant information is necessarily discarded and distortion starts to increase, leading to lossy compression. This trade-off is usually implemented as a parameter that

constraints the capacity (or maximal mutual information) of the system (Alemi et al., 2016; Denève et al., 2017; Genewein et al., 2015; Kingma and Welling, 2013; Ortega and Braun, 2013; Tishby et al., 2000): $\underset{p(x'|x),p(y|x'),p(y|x',T)}{\arg\max}\ U-\beta\mathcal{C}$, where U represents some utility measure and β is a (Lagrangian) factor that adjusts the trade-off between costs and performance. In order to apply this framework to behavioural data, this β parameter then has to be fit to observed performance data (Sims, 2016).

This adjustability of information rate evokes the concept of task engagement and the fact that motivational factors and reward incentives can influence task performance by putting high information rate at a premium (Camerer et al., 1999). This mechanism implies some cost-benefit computation to select the optimal trade-off between the cost of information rate and the value associated with performance, akin to many earlier models of effort-based decision making (Chong et al., 2017; Christie and Schrater, 2015; Rigoux and Guigon, 2012; Shenhav et al., 2013; Verguts et al., 2015). In Figure 7, we illustrate how such a trade-off could play out in an example task in which subjects have to report the direction of motion of a random dot kinetogram in two conditions of motion coherence (Zénon and Krauzlis, 2012). Dot motion directions are distributed according to Gaussian distribution (variance=2) and distortion (i.e. error rate) is quantified as the mean-squared error between correct and reported directions (see Figure 7A). Utility is represented as a decreasing function of distortion ($U(D(R))=e^{1-D(R)}$, see Figure 7B) and information cost is a convex function of information rate $\mathcal{C}(R)=R^2$. The particular form of these formulas is chosen arbitrarily for the sole purpose of illustration. The optimal trade-off between distortion and information rate is the one associated with maximal net value (utility minus cost), and depends on the motion coherence (i.e. amount of noise), as shown on the plots of Figure 7 (dashed line, panels C, D).

This schematic example shows how the present framework can help to explain why increasing signal noise (i.e. decreasing motion coherence) leads to increased information cost.

3.6 Summary so far

To summarize, we have offered a unitary perspective that may explain many experimental findings on cognitive effort and fatigue - by appealing to the fact that tasks that are known to be cognitively costly, such as novel tasks or those that require counteracting habitual policies or switching contingencies, all have high information costs associated to encoding or revising probability distributions within the generative models that support task performance. Equipped with this formalization of cognitive costs, we can now turn to their relationship to effort and their potential implementation in the brain.

4. Relevance for subjective effort and task avoidance

As mentioned in the introduction, the cost of cognition manifests itself behaviourally as a subjective percept of effort and as a tendency to avoid demanding tasks. Previous characterizations of the origin of cognitive effort can be classified into two broad categories. First, effort can be framed as a consequence of resource limitations such as depletable metabolic precursors (reviewed in Shenhav et al., 2017). Second, cognitive effort can be described as the phenomenological manifestation of the opportunity cost of engaging limited cognitive resources in demanding cognitive tasks (Kurzban et al., 2013). In the following, we will detail the expected consequences of the implementation of information costs in the brain, as described above. We will show that these consequences can be reconciled with the two aforementioned views on effort, while adding some novel quantitative predictions.

Before that, a methodological caveat is necessary. We have discussed how cognitive computational costs should be evaluated in terms of KL divergence between priors and

posteriors. However, to apply this framework – and measures of information like surprisal, entropy and mutual information – to understand brain cognitive costs, it is necessary to assume that the brain uses an optimal strategy to encode its variables of interest. In such a code, each datapoint is represented with an average number of symbols which is proportional to the negative log of its probability (Shannon, 1948). Here probability is meant in the sense of predicted occurrence, given all the information we have at our disposal. For instance, the word "hatter" is not very common in English sentences in general but a sentence starting with "As mad as a" is much more likely to be continued with the word "hatter". In this latter case, the word "hatter", under optimal encoding strategy, should be encoded with small number of symbols (Lai, 2009). So, in order for the proposed framework to be applicable to brain processes, we need to assume that the brain indeed approaches such an optimal encoding strategy. This assumption is at the core of the efficient coding hypothesis (Collell and Fauquet, 2015; Harremoës and Tishby, 2007; Laughlin, 2001; Sims, 2016; Tkačik and Bialek, 2014; Wei and Stocker, 2015). Efficient coding was initially described as a theory of redundancy reduction, according to which biological systems decorrelate sensory signals to avoid redundancy (Attneave, 1954; Barlow, 1961; Simoncelli and Olshausen, 2001). The theory has been successively extended to include other mechanisms through which neural coding adapts to the statistical structure of its environment (Simoncelli, 2003; Smith and Lewicki, 2006; Tkačik and Bialek, 2014). One crucial aspect of efficient coding is the usage of an adaptive code, in which the cost for encoding each symbol is inversely proportional to its frequency in the environment - given the agent's model of the environment (Collell and Fauquet, 2015; Fairhall et al., 2001).

An impressive body of experimental evidence has been accumulated in favour of this hypothesis (Tkačik and Bialek, 2014). Low-level vision and audition show data filtering properties and neural codes that are closely similar to predictions issued from efficient coding

models (Borst and Theunissen, 1999; Gutnisky and Dragoi, 2008; Laughlin, 2001; Olshausen and Field, 2004; Sharpee et al., 2006; Smith and Lewicki, 2006). Predictability leads to diminished brain activation (Auksztulewicz and Friston, 2016; Bell et al., 2016; Carreiras et al., 2009; Garrido et al., 2013; Lieder et al., 2013; Mars et al., 2008; Meyniel et al., 2016; Overath et al., 2007; Wacongne et al., 2012) and pupil responses (Friedman et al., 1973), but increases reliability of encoding (Kok et al., 2012), in agreement with the idea that predictable stimuli, carrying little information, are encoded more economically. Along the same line, decreased brain activation following training (Chen and Wise, 1995; Solopchuk et al., 2017; Toni et al., 1998; Wiestler and Diedrichsen, 2013) suggests that training decreases metabolic cost by allowing learners to leverage task statistics to optimize brain representations, while increasing the quantity of information being processed. Even though these pieces of evidence do not yet allow us to consider the brain usage of optimally efficient coding as an established fact, we will assume here that the brain indeed uses an efficient code since this assumption remains a pre-condition for our framework to be applicable.

Another important question, if one is to apply the present information theoretic perspective to the brain, concerns the ways different brain areas are taxed by costs. The brain is massively parallel, and while the total information cost of a task can be evaluated, it remains to be determined how the cost affects (and is shared between) specific brain areas and individual neurons within these areas. One possible starting point would be to assume that brain networks supporting perceptual and decision processes would be taxed by the three kinds of costs considered in our proposal; namely, the costs associated to a perceptual stage, in which the sensory input x would be represented through an internal variable x'; the cost of the automatic stage, in which the stimulus leads to a response independently of context; and the costs associated to context-dependent response selection, which would amount to updating $p(y|x')$ to the final response $p(y|x',T)$. While this is certainly a simplification (as perceptual

and decision processes are not segregated in the brain), it would provide a first rough set of hypotheses on the ways different brain areas would be affected by different kinds of costly tasks. However, a more complete view should consider the hierarchical organization of perception-action loops in the brain (Fuster, 1990), and relate different kinds of costs to hierarchical lower and higher stages of cognitive processing.

It is worth noting that hierarchical processing can be viewed either as serial or parallel. Serial processing has the major drawback that total information capacity of the system is equal to the information capacity of its weakest link (Genewein et al., 2015). In contrast, parallel processing consists in architectures in which the outputs from high-level processes provide priors, rather than inputs, to low-level processes (Genewein et al., 2015) and in which the information capacities of the individual parts sum up. Here, while we consider perceptual processing as a serial process (information is first stored in a variable *x'* and is then further processed to provide response *y*) automatic and context-dependent processes follow a parallel architecture, in which the outcome of the automatic process is fed as a prior into the context-dependent one. This parallel hierarchical architecture evokes predictive coding and active inference, which are built on this type of organization (Bastos et al., 2012; Clark, 2013; Friston et al., 2009; Pezzulo et al., 2018; Rao, 2010), or the work of Koechlin and Summerfield, which proposed such a parcellation of cognitive control costs in terms of depth of contextualization (Koechlin and Summerfield, 2007). Naturally, one can leverage the vast neuroimaging literature in order to estimate the decomposition of cognitive tasks into relevant sub-processes but studies using directly information theoretic approaches to study brain activations will be necessary to achieve better specificity (Kriegeskorte and Bandettini, 2007; Wu et al., 2017).

These caveats in mind, we can propose two (non exclusive) approaches to explain why information cost should lead to the phenomenological perception of effort and to task

avoidance. The first approach is to consider that the information capacity of brain areas (i.e. the maximum of the mutual information between the data and its neural representation) is limited. Therefore, increasing the information demand necessarily leads to decreased capacity for other concurrent processes and when a given process utilizes the full capacity, larger information demands translate into longer reaction times. The second approach consists in considering that information costs have direct equivalence in terms of energetic demands. According to this view, the constraint on the system is energetic, rather than informational. This metabolic constraint leads in turn to two potential consequences on subjective cognitive costs in terms of global opportunity costs and metabolic alterations. These different points of view are detailed below (see also Figure 2, blue panel).

4.1 The information capacity perspective on cognitive costs

Neurons have limited information processing capacity (Schneidman et al., 2000), and so do brain areas (Marois and Ivanoff, 2005; Verghese and Pelli, 1992). Therefore, taxing of brain area capacity by a cognitive task will decrease the capacity left for other processes. The functional significance of this decrease depends on the affected brain regions. Here, for brevity, we focus on two brain networks: sensory cortices and the multiple demand system. The sensory cortices encompass large area, with topographic organization, in which processing of different input features or spatial locations leads to activations of different cortical regions (Purves et al., 2001). Declines in capacity in such topographic areas are easy to compensate. The situation may be different in the case of the multiple demand system: an ensemble of brain areas engaged in a large variety of tasks (Fedorenko et al., 2013), including interoceptive processing (Kleckner et al., 2017), i.e. the adaptation of behaviour to fullfil physiological needs, a function essential to survival. Thus, decrease of information capacity within this network may have more adverse behavioural consequences (and presumably more

severe effort and fatigue phenomenology). This bridges the present framework with the opportunity cost view on mental effort (Kurzban et al., 2013), according to which effort must be understood as the cost of forfeiting potentially more valuable courses of action than the current task, due to taxing of limited cognitive resources. This view is also in line with the concept of representational capacity limitation, according to which tasks must compete to access shared processing resources, leading to trade-offs and opportunity costs (Shenhav et al., 2017).

So, when a task taxes part of the information capacity of a brain structure, its cost can be expressed in terms of the limitation it imposes on the other processes, dependent on the same structure, which could be run in parallel (Kurzban et al., 2013). However, another important determinant of this opportunity cost is how long the task lasts, or in other words, its associated reaction time. This is especially crucial for tasks whose performance depends on information bottlenecks, i.e. on brain structures whose capacity is fully engaged in the task. The original findings of Hick and Hyman suggested that such capacity limit was reached for tasks as simple as stimulus-response associations (Hick, 1952; Hyman, 1953), since reaction times in their studies was a linear function of information cost. This suggests that in many circumstances, rather than a limit on multi-tasking, information cost must be understood in terms of the opportunity cost of time (Niv et al., 2006; Payne et al., 1996; Zénon et al., 2016), i.e. how long brain resources remain dedicated to the same cognitive activity. In an attempt to formalize these ideas, we propose that, according to the capacity perspective, subjective costs are proportional to the maximum, across all subprocesses, of the ratio between local information costs and local capacity. Thus, subjective cost $\mathcal{F}$ of a cognitive activity could be approximated as $\mathcal{F} \propto \max \frac{C_p}{L_p}, \forall p \in P$, considering that the cognitive activity is composed of an ensemble *P* of subprocesses *p,* that $\mathcal{C}_p$ is the information cost of a specific subprocess and

that L_p is the information capacity of the corresponding brain structure (we use L rather than the standard symbol C to avoid confusion with the cost symbol).

4.2 The metabolic perspective on cognitive costs

Information processing requires energy expenditure (Landauer, 1996; Ortega and Braun, 2013; Sengupta et al., 2013; Still et al., 2012). The relation between informational and energetic costs was central in early theories of efficient coding (Atick, 1992; Attneave, 1954; Barlow, 1961; Borst and Theunissen, 1999; Niven and Laughlin, 2008), which assumed that neural responses, carrying large energetic costs, impose a constraint on brain information processing capacity. Under this framework, the brain attempts to maximize the amount of (mutual) information it processes, given this fixed energetic constraint. More recent approaches have relaxed this principle by considering the constraint to be adjustable (Denève et al., 2017; Ortega and Braun, 2013; Park and Pillow, 2017; Sengupta et al., 2013), opening the door to cost-benefit adjustments of the kind exposed above (see Figure 7), which allow metabolic costs to be adjusted as a function of demands in performance (Genewein et al., 2015; Park and Pillow, 2017; Sims, 2016). So, the metabolic perspective on cognitive costs assumes that subjective effort and task avoidance have a fundamental energetic origin, and that energetic costs E are a function of information costs as defined above: $E = f(\mathcal{C})$. Even though determining function f precisely is difficult, it is possible to approximate it with standard neuroscience techniques. For instance, tackling this question with neuroimaging in the sensory domain could be done by relating the amplitude of the haemodynamic response of brain areas (approximating energy demands) to the mutual information between their activations and their inputs.

If multiple subprocesses p are running in parallel, the energetic cost depends on the sum over the information costs of these tasks. However, we now must add another constraint to this

formula, which is that the total energetic cost across the ensemble of brain processes *B* should be constant: $E_{total} = \sum_{p \in B} E_p = \sum_{p \in B} f(C_p) = k$

This constraint comes from the observation that global cerebral energy consumption does not vary between resting and active conditions (Lennie, 2003; Sokoloff, 2009; Sokoloff et al., 1955). Blood delivers glucose and oxygen to brain in excess of demand, such that in physiological conditions (i.e. in absence of hypoxia or hypoglycaemia), the availability of energetic precursors is not a limiting factor to cognitive activity (Brown and Ransom, 2014). However, total blood delivery to the brain is a constant that cannot be upregulated in response to cognitive demand (Brown and Ransom, 2014). Therefore, the cost of cognitive activity can hardly be explained by the need to curb total energetic consumption. It is noteworthy, however, that despite this lack of change in global energetic demand of the brain, global glucose intake increases in the brain during cognitive activity (Volkow et al., 2008). This utilization of glucose in excess of oxygen consumption is referred to as aerobic glycolysis (Vaishnavi et al., 2010). It is modulated by arousal (Dienel and Cruz, 2016) and while the function of aerobic glycolysis remains debated, it may be linked to cortical plasticity (Goyal et al., 2014) and the replenishment of glutamate and GABA reserves (Hertz and Chen, 2017). This increased glucose demand during active behaviour may appear to justify resource depletion theories of cognitive effort, according to which glucose is the main resource that puts a constraint on cognitive activity (Gailliot and Baumeister, 2007). However, this theory, and the experimental evidence on which it is based, have been put under increased criticism recently (Hagger et al., 2016, 2010; Kurzban et al., 2013; Molden et al., 2012; Shenhav et al., 2017).

Here we propose two alternative views. First, if total brain energetic consumption is constant and each brain region consumes energy in proportion to the amount of information it processes, then we must assume that neural activity tied to the execution of a cognitive task,

irrespective of where it takes place in the brain, should decrease the total capacity available to other processes. This generalizes the point made in the previous section, which considered capacity as a local feature, imposed by limited information capacity of neurons and brain structures. In other words, the brain as a whole has limited capacity, due to its constant energy consumption, and allocating metabolic resources to one process leads to opportunity costs related to the reduced capacity available to other processes: $\mathcal{F} = \frac{E_{task}}{E_{total}} = \frac{\sum_{p \in P} E_p}{E_{total}}$, in which P represents all subprocesses involved in a specific task.

Therefore, increased metabolic costs in specific brain regions would entail decreased demands in other regions, which is in line with the findings that brain activations during cognitive tasks are accompanied by commensurate deactivations in resting-state brain areas (Fox and Raichle, 2007). Conversely, following training on a task A, its information cost should decrease (see section 3.2 above), leading to progressively smaller proportion of the total energetic resources being allocated to this task (Solopchuk et al., 2016). This extra capacity gained over the course of training can, therefore, be gradually allocated to other concurrent tasks, such as planification of future actions, or background interoceptive processes.

A second metabolic point of view on the question of why demanding tasks are avoided is that local information costs lead to progressive local metabolic alterations that accumulate over time. The rate of accumulation of these alterations $\dot{\rho}$ would then be a function of neural activity, itself proportional to information costs: $\mathcal{F} = \sum_{p \in P} \dot{\rho}_p$

Cognitive costs would then be proxies that the brain uses to prevent these local metabolic alterations to occur. This point of view has the advantage of providing a direct link between *effort*, which would become the subjective experience of the rate of accumulation of local metabolic alterations (Tucker and Noakes, 2009), and *fatigue*, which would be the

anticipation (Benoit et al., 2017) or the direct consequence of these alterations (Gergelyfi et al., 2015; Hockey, 1997; van der Linden et al., 2003).

Evidence that prolonged local brain activation leads to functional alterations supports this hypothesis (Mednick et al., 2002), but the exact nature of these alterations can only be speculated at this point. Some have proposed that glycogen reserves could deplete (Christie and Schrater, 2015), or amyloid peptides accumulate over time (Holroyd, 2015). Others have suggested that the accumulation of deviations from metabolic steady-state could lead cortical regions to enter a local sleep mode, characterized by slow-wave synchronization and associated with disturbed processing capacity (Siclari and Tononi, 2017). Confirming the existence and deciphering the nature of these metabolic alterations is an important challenge for future research on effort and fatigue.

4.3 The possible roles of arousal within this framework

Arousal could be defined as a global brain state characterized by the amplitude of synchronized low-frequency oscillations and sensory responsiveness (McGinley et al., 2015) and is controlled by brainstem nuclei and neuromodulators such as noradrenaline and acetylcholine (Reimer et al., 2016). An impressive amount of evidence has shown that arousal correlates with cognitive workload (Beatty and Lucero-Wagoner, 2000; Richter et al., 2016; van der Wel and van Steenbergen, 2018). Interestingly, arousal also responds strongly to prediction errors, or surprise (Ferreira-Santos, 2016; Friedman et al., 1973; Kloosterman et al., 2015; Lavín et al., 2014; O'Reilly et al., 2013; Preuschoff et al., 2011). Taken together, these two sets of observations are in line with the view that the modulation of arousal is ultimately a function of uncertainty (Yu and Dayan, 2003), or equivalently, of information load (entropy is uncertainty that needs to be resolved) – which permits to link nicely arousal to our proposed information-theoretic framework.

The functional role of increased arousal in response to information costs remains unclear. One possibility is that arousal mobilizes the metabolic apparatus orchestrated by astrocytes in response to neural activity (O'Donnell et al., 2012; Paukert et al., 2014), including glycogen reserves (Hertz and Zielke, 2004; O'Donnell et al., 2012), while also restricting the circulation of cerebrospinal fluid, thus limiting the capacity of the brain to eliminate potentially harmful metabolites (Xie et al., 2013). Interestingly, although arousal is a global phenomenon, its effect in cortex is believed to be restricted to active regions (Mather et al., 2016). Increased arousal would thus lead to increased metabolic rate in these active regions, while dampening activation in already less active background structures (Mather et al., 2016).

6. Conclusions and related work

The phenomenology of cognitive cost and cognitive effort - in terms of subjective feeling of exhaustion experienced when performing a cognitive task and its associated task-avoidance - is relatively well known. Yet, several aspects of the problem of the costs of cognition are currently debated, including the specification of what is costly in cognitive processing, how to quantify these costs and what is their adaptive value.

We have defended a view that starts from the idea that, if the brain encodes information in accordance with the principles of efficient coding, then the cognitive costs associated to task execution should be a function of the amount of information required to update priors to posterior beliefs. We have used this framework to explain the cognitive costs associated to different classes of tasks known to be subjectively demanding, and have showed that this informational perspective can provide a unitary perspective on several experimental findings in the literature. Furthermore, we have discussed how information costs could translate into

cognitive effort (i.e., the subjective feeling associated to performing costly tasks). We have described three hypotheses regarding this link. First, subjective effort may arise from the usage of limited, local information capacity, chiefly in the multiple demand system, leading to opportunity costs. Second, subjective effort may be the consequence of the global limit on information capacity, caused by the constant energy consumption of the brain, also leading to opportunity costs, of a slightly different nature. Third, subjective effort could be caused by the accumulation of local metabolic alterations, whose rate is a function of information demand. In that case, effort and cognitive fatigue that ensue from long-term engagement in costly tasks could be considered as adaptive mechanisms that prevent individuals from performing activities that may have adverse consequences in the long run, i.e., activities that imply huge local metabolic demands. Importantly, these three proposals are experimentally testable and could help guiding future research in this domain.

Our perspective is coherent with several recent theories, which proposed that task avoidance stems from the (optimal) choice between potential policies while accounting for their respective cost. In this framework, cognitive effort is assumed to depend on the degree to which the task depends on cognitive control (Shenhav et al., 2017, 2013). The present work is in continuity with these earlier proposals as it relies on optimality principles and describes cognitive effort as a cost, which discounts the expected utility of a given course of actions (Apps et al., 2015; Chong et al., 2017; Manohar et al., 2015; Westbrook and Braver, 2015). However, our proposal departs from these theories as it addresses more directly the question of the *causes* of cognitive cost - and casts them in terms of information principles. In particular, the present framework attributes costs to general informational aspects of cognitive processing, rather than specifically to cognitive control. What determines cognitive cost is the amount of information to be processed and tasks involving cognitive control may

be more costly in general because they require updating of entire task sets or control policies (Bhandari and Badre, 2018), which results in cascade reconfiguration or updating of the whole sensory and motor representations, which entails large information costs. Additionally, cognitive control tasks could be more costly because they tax particularly the multiple demand system (Koechlin and Summerfield, 2007; Wu et al., 2017), where the opportunity cost associated to utilization of limited information capacity may be larger. These (not mutually exclusive) hypotheses remain to be tested in future research.

Koechlin and Summerfield have proposed an information theoretical approach to cognitive control that shares several aspects with our own framework. In their study, the total cost of selecting an action is framed as the sum of two terms (Koechlin and Summerfield, 2007). The first is an automatic, goal-independent association between stimuli and actions whose cost is measured as the mutual information between stimuli and actions: $I(x;y)$. The remaining cost of action selection $H(y)-I(x;y)$ then corresponds to cognitive control and could be decomposed further into several hierarchical processes with different levels of contextualization. Similarly to Koechlin and Summerfield's proposal (and others (Genewein et al., 2015; Pezzulo et al., 2013)), our framework assumes a hierarchical architecture in which the cognitive control process takes the outcome of the automatic process as a prior, making their contributions additive (Genewein et al., 2015). The present paper differs from these earlier treatments in various ways, as it focuses on effort costs; considers the cost of perceptual processing, which is necessary to account for known behavioural results (Fan et al., 2008; Wifall et al., 2016); and explicitly discusses rate-distortion theory. Furthermore, our proposal is more explicit in the adoption of a specific formalism to clarify the different potential sources of cost. Finally, the present paper discusses in detail the application of the framework to different types tasks that are known to be demanding.

Another proposal which relates to ours, addresses cognitive effort from a normative perspective in which apparently maladaptive states (cognitive fatigue) constitute the adaptive response of an optimal controller that has (or feels having) low self-efficacy and limited control over one's own environment, similar to learned helplessness in the animal learning literature (Stephan et al., 2016). Although we have not addressed the long-term consequences of being exposed to complex cognitive tasks, our model would be coherent with this proposal in assuming that prolonged expectations of poor outcomes, or poor control, would crystallize task-avoidance behaviour. In other words, an agent whose local metabolic resources are frequently depleted, could develop an adaptive task aversion, manifested, for example, as chronic fatigue syndrome.

The present framework has also some limitations that will need to be covered in future works. For example, it does not fully explain important features of switching costs, such as the fact that moving from a preferred to a less preferred stimulus-response association is less costly than the other way around (Wylie and Allport, 2000), or the fact that preparation periods, no matter how long, cannot suppress the switching costs (Wylie and Allport, 2000). It is also challenging to explain why switching between tasks that are based on similar stimulus features or similar response modalities is associated with smaller costs than switching between dissimilar tasks (Arrington et al., 2003). Explaining this finding would require the model to account for some hierarchical categorization of perceptual and response representations x' and y (Zhao et al., 2017), that go beyond the scope of the present paper.

Another limitation of the current framework is that it does not fully account for adaptive (positive) aspects of costly cognitive processing; for example, engaging in (costly) information-foraging behaviour during learning. Challenging cognitive activity is not always

experienced as aversive but may even be sought for its own sake (Cacioppo et al., 1984; Inzlicht et al., 2018); and on the contrary, idling or engaging in repetitive, monotonous tasks can be unpleasant (Nakamura and Csikszentmihalyi, 2002). Costly information-seeking is in apparent contradiction with our framework, but it can be explained by the mediating influence of intrinsic motivation and epistemic value (Friston et al., 2015, 2017). Complex tasks need mastery, and intrinsic motivation appears to depend in large part on the feeling of competence and autonomy, i.e. the feeling of being capable of performing a task, despite its difficulty (Ryan and Deci, 2000) - and in the active inference setting, on the necessity to improve one's internal models (Friston et al., 2017). In other words, improving one's internal model to minimize surprise in the future can compensate the short-term costs of investing effort in the present moment. Spontaneous engagement in demanding tasks would therefore depend on the cost-benefit comparison between immediate informational and/or metabolic costs and future needs, mediated by mechanisms of exploration, model learning or intrinsic motivation that are part and parcel of active inference (Friston et al., 2017) – and which complement nicely the information theoretic framework advanced here.

Another limitation of the present work stems from its reliance on a series of assumptions: it is valid only to the extent that those assumptions are true. First and foremost, it relies on the concept of efficient coding, according to which brain's encoding of information makes optimal use of its metabolic resource, such that energetic cost is proportional to the entropy of the encoded information. Even though this concept is based on a large body of evidence, it cannot yet be viewed as an established fact. Second, we assume a certain cognitive architecture, in which perceptual encoding of sensory information occurs first and is then fed into two parallel cognitive processes: one automatic and one context-dependent. We believe that this architecture is the simplest one that can account for all the categories of demanding cognitive tasks reviewed in the paper. Naturally, more complex architectures could be

considered. Finally, we also base our estimation of cognitive cost on the priors used by subjects for perceptual encoding and cognitive control processing. The accuracy of the predictions made by the model will, therefore, depend crucially on the choice of those priors. A possible approach to this problem is to base the choice of the priors on observed data. Indeed, since priors are the only unknown parameters in the model, they can be fitted to participants' behaviour (Sims, 2016).

In sum, we have reviewed an information theoretical perspective on the costs of cognition that links cognitive effort to principles of efficient coding and active inference. This framework systematizes and extends previous treatments that used information theoretic principles to explain cognitive control processes (Koechlin and Summerfield, 2007; Ortega and Braun, 2013; Sengupta et al., 2013). Furthermore, we have discussed how this framework permits to harmonize theories that link cognitive costs to capacity limitations and metabolic principles, while also producing novel empirical predictions.

Acknowledgements

This work was supported by Fonds de la Recherche Scientifique (FNRS–FDP), Fondation Médicale Reine Elisabeth (FMRE), the Fondation Louvain and IdEx Bordeaux.

References

Adams, R.A., Stephan, K.E., Brown, H.R., Frith, C.D., Friston, K.J., 2013. The Computational Anatomy of Psychosis. Front. Psychiatry 4. doi:10.3389/fpsyt.2013.00047

Alemi, A.A., Fischer, I., Dillon, J. V., Murphy, K., 2016. Deep Variational Information Bottleneck 1–19.

Allport, A., Styles, E., Hsieh, S., 1994. Shifting Intentional Set: Exploring the Dynamic Control of Tasks, in: Attention and Performance XV: Conscious and Nonconscious Information Processing. pp. 421–452. doi:10.1126/science.1134404

Apps, M.A.J., Grima, L.L., Manohar, S., Husain, M., 2015. The role of cognitive effort in subjective reward devaluation and risky decision-making. Sci. Rep. 5, 16880. doi:10.1038/srep16880

Arrington, C.M., Altmann, E.M., Carr, T.H., 2003. Tasks of a feather flock together: Similarity effects in task switching. Mem. Cogn. 31, 781–789. doi:10.3758/BF03196116

Atick, J., 1992. Could information theory provide an ecological theory of sensory processing? Netw. Comput. Neural Syst. 3, 213–251. doi:10.1088/0954-898X/3/2/009

Attneave, F., 1954. Some informational aspects of visual perception. Psychol. Rev. 61, 183–193. doi:10.1037/h0054663

Auksztulewicz, R., Friston, K., 2016. Repetition suppression and its contextual determinants in predictive coding. Cortex 80, 125–140. doi:10.1016/j.cortex.2015.11.024

Bailey, A., Channon, S., Beaumont, J.G., 2007. The relationship between subjective fatigue and cognitive fatigue in advanced multiple sclerosis. Mult. Scler. J. 13, 73–80. doi:10.1177/1352458506071162

Barlow, H.B., 1961. Possible principles underlying the transformation of sensory messages. Sens. Commun. 217–234.

Bastos, A.M., Usrey, W.M., Adams, R.A., Mangun, G.R., Fries, P., Friston, K.J., 2012. Canonical Microcircuits for Predictive Coding. Neuron 76, 695–711. doi:10.1016/j.neuron.2012.10.038

Beatty, J., Lucero-Wagoner, B., 2000. The pupillary system, in: Handbook of Psychophysiology (2nd Ed.). pp. 142–162.

Bell, A.H., Summerfield, C., Morin, E.L., Malecek, N.J., Ungerleider, L.G., 2016. Encoding of Stimulus Probability in Macaque Inferior Temporal Cortex. Curr. Biol. 26, 2280–2290. doi:10.1016/j.cub.2016.07.007

Benoit, C.-E., Solopchuk, O., Borragan, G., Carbonnelle, A., Van Durme, S., Zenon, A., 2017. Objective but not subjective fatigue increases cognitive task avoidance. bioRxiv. doi:10.1101/208322

Bhandari, A., Badre, D., 2018. Learning and transfer of working memory gating policies. Cognition. doi:10.1016/j.cognition.2017.12.001

Borragán, G., Slama, H., Bartolomei, M., Peigneux, P., 2017. Cognitive fatigue: A Time-based Resource-sharing account. Cortex 89, 71–84. doi:10.1016/j.cortex.2017.01.023

Borst, A., Theunissen, F.E., 1999. Information theory and neural coding. Nat. Neurosci. 2, 947–57. doi:10.1038/14731

Brown, A.M., Ransom, B.R., 2014. Astrocyte glycogen as an emergency fuel under conditions of glucose deprivation or intense neural activity. Metab. Brain Dis. 30, 233–239. doi:10.1007/s11011-014-9588-2

Cacioppo, J.T., Petty, R.E., Kao, C.F., 1984. The efficient assessment of need for cognition. J. Pers. Assess. 48, 306–307. doi:10.1207/s15327752jpa4803_13

Carreiras, M., Riba, J., Vergara, M., Heldmann, M., Münte, T.F., 2009. Syllable congruency and word frequency effects on brain activation. Hum. Brain Mapp. 30, 3079–3088. doi:10.1002/hbm.20730

Chen, L.L., Wise, S.P., 1995. Neuronal activity in the supplementary eye field during acquisition of conditional oculomotor associations. J Neurophysiol 73, 1101–1121. doi:10.1016/S0896-6273(00)80658-3

Christie, S.T., Schrater, P., 2015. Cognitive cost as dynamic allocation of energetic resources. Front. Neurosci. 9, 1–15. doi:10.3389/fnins.2015.00289

Clark, A., 2013. Whatever next? Predictive brains, situated agents, and the future of cognitive science. Behav. Brain Sci. doi:10.1017/S0140525X12000477

Cohen, J.D., Dunbar, K., McClelland, J.L., 1990. On the control of automatic processes: A parallel distributed processing account of the stroop effect. Psychol. Rev. 97, 332–361. doi:10.1037/0033-295X.97.3.332

Collell, G., Fauquet, J., 2015. Brain activity and cognition: a connection from thermodynamics and information theory. Front. Psychol. 6, 818. doi:10.3389/fpsyg.2015.00818

Cooper, P.S., Garrett, P.M., Rennie, J.L., Karayanidis, F., 2015. Task uncertainty can account for mixing and switch costs in task-switching. PLoS One 10, 1–17. doi:10.1371/journal.pone.0131556

Delorme, C., Salvador, A., Valabrègue, R., Roze, E., Palminteri, S., Vidailhet, M., De Wit, S., Robbins, T., Hartmann, A., Worbe, Y., 2016. Enhanced habit formation in Gilles de la Tourette syndrome. Brain 139, 605–615. doi:10.1093/brain/awv307

Denève, S., Alemi, A., Bourdoukan, R., Dene, S., 2017. The Brain as an Efficient and Robust Adaptive Learner. Neuron 94, 969–977. doi:10.1016/j.neuron.2017.05.016

Dienel, G.A., Cruz, N.F., 2016. Aerobic glycolysis during brain activation: adrenergic regulation and influence of norepinephrine on astrocytic metabolism. J. Neurochem. 14–52. doi:10.1111/jnc.13630

Donnarumma, F., Maisto, D., Pezzulo, G., 2016. Problem Solving as Probabilistic Inference with Subgoaling: Explaining Human Successes and Pitfalls in the Tower of Hanoi. PLoS Comput. Biol. doi:10.1371/journal.pcbi.1004864

Fairhall, A.L., Lewen, G.D., Bialek, W., de Ruyter van Steveninck, R.R., 2001. Efficiency and ambiguity in an adaptive neural code. Nature 412, 787–792. doi:10.1038/35090500

Fan, J., 2014. An information theory account of cognitive control. Front. Hum. Neurosci. 8. doi:10.3389/fnhum.2014.00680

Fan, J., Guise, K.G., Liu, X., Wang, H., 2008. Searching for the majority: Algorithms of voluntary control. PLoS One 3. doi:10.1371/journal.pone.0003522

Fedorenko, E., Duncan, J., Kanwisher, N., 2013. Broad domain generality in focal regions of frontal and parietal cortex. Proc. Natl. Acad. Sci. U. S. A. 110, 16616–21. doi:10.1073/pnas.1315235110

Ferreira-Santos, F., 2016. The role of arousal in predictive coding. Behav. Brain Sci. 39, e207. doi:10.1017/S0140525X15001788

Fox, M.D., Raichle, M.E., 2007. Spontaneous fluctuations in brain activity observed with functional magnetic resonance imaging. Nat. Rev. Neurosci. 8, 700–711. doi:10.1038/nrn2201

Friedman, D., Hakerem, G., Sutton, S., Fleiss, J.L., 1973. Effect of stimulus uncertainty on the pupillary dilation response and the vertex evoked potential. Electroencephalogr. Clin. Neurophysiol. 34, 475–484. doi:10.1016/0013-4694(73)90065-5

Friston, K., 2010. The free-energy principle: a unified brain theory? Nat. Rev. Neurosci. 11, 127–138. doi:10.1038/nrn2787

Friston, K., Kiebel, S., Barlow, H.B., Feynman, R.P., Neal, R.M., Hinton, G.E., Neisser, U., 2009. Predictive coding under the free-energy principle. Philos. Trans. R. Soc. Lond. B. Biol. Sci. 364, 1211–21. doi:10.1098/rstb.2008.0300

Friston, K., Rigoli, F., Ognibene, D., Mathys, C., Fitzgerald, T., Pezzulo, G., 2015. Active inference and epistemic value. Cogn. Neurosci. 6, 187–224. doi:10.1080/17588928.2015.1020053

Friston, K., Thornton, C., Clark, A., 2012. Free-energy minimization and the dark-room problem. Front. Psychol. 3, 130. doi:10.3389/fpsyg.2012.00130

Friston, K.J., Lin, M., Frith, C.D., Pezzulo, G., Hobson, J.A., Ondobaka, S., 2017. Active Inference, Curiosity and Insight. Neural Comput. 1–51. doi:10.1162/neco_a_00999

Friston, K.J., Stephan, K.E., Montague, R., Dolan, R.J., 2014. Computational psychiatry: the brain as a phantastic organ. The Lancet Psychiatry 1, 148–158. doi:10.1016/S2215-0366(14)70275-5

Fuster, J.M., 1990. Prefrontal Cortex and the Bridging of Temporal Gaps in the Perception-Action Cycle. Ann. N. Y. Acad. Sci. 608, 318–336. doi:10.1111/j.1749-6632.1990.tb48901.x

Gailliot, M.T., Baumeister, R.F., 2007. The physiology of willpower: Linking blood glucose to self-control. Personal. Soc. Psychol. Rev. 11, 303–327. doi:10.1177/1088868307303030

Garrido, M.I., Sahani, M., Dolan, R.J., 2013. Outlier Responses Reflect Sensitivity to Statistical Structure in the Human Brain. PLoS Comput. Biol. 9. doi:10.1371/journal.pcbi.1002999

Genewein, T., Leibfried, F., Grau-Moya, J., Braun, D.A., 2015. Bounded Rationality, Abstraction, and Hierarchical Decision-Making: An Information-Theoretic Optimality Principle. Front. Robot. AI 2. doi:10.3389/frobt.2015.00027

Gergelyfi, M., Jacob, B., Olivier, E., Zénon, A., 2015. Dissociation between mental fatigue and motivational state during prolonged mental activity. Front. Behav. Neurosci. 9. doi:10.3389/fnbeh.2015.00176

Goyal, M.S., Hawrylycz, M., Miller, J.A., Snyder, A.Z., Raichle, M.E., 2014. Aerobic glycolysis in the human brain is associated with development and neotenous gene expression. Cell Metab. 19, 49–57. doi:10.1016/j.cmet.2013.11.020

Grau-Moya, J., Braun, D. ~A., 2015. Adaptive information-theoretic bounded rational decision-making with parametric priors. ArXiv e-prints.

Gutnisky, D.A., Dragoi, V., 2008. Adaptive coding of visual information in neural populations. Nature 452, 220–224. doi:10.1038/nature06563

Hagger, M.S., Chatzisarantis, N.L.D., Alberts, H., Anggono, C.O., Batailler, C., Birt, A.R., Brand, R., Brandt, M.J., Brewer, G., Bruyneel, S., Calvillo, D.P., Campbell, W.K., Cannon, P.R., Carlucci, M., Carruth, N.P., Cheung, T., Crowell, A., De Ridder, D.T.D., Dewitte, S., Elson, M., Evans, J.R., Fay, B.A., Fennis, B.M., Finley, A., Francis, Z., Heise, E., Hoemann, H., Inzlicht, M., Koole, S.L., Koppel, L., Kroese, F., Lange, F., Lau, K., Lynch, B.P., Martijn, C., Merckelbach, H., Mills, N. V., Michirev, A., Miyake, A., Mosser, A.E., Muise, M., Muller, D., Muzi, M., Nalis, D., Nurwanti, R., Otgaar, H., Philipp, M.C., Primoceri, P., Rentzsch, K., Ringos, L., Schlinkert, C., Schmeichel, B.J., Schoch, S.F., Schrama, M., Schütz, A., Stamos, A., Tinghög, G., Ullrich, J., vanDellen, M., Wimbarti, S., Wolff, W., Yusainy, C., Zerhouni, O., Zwienenberg, M., 2016. A Multilab Preregistered Replication of the Ego-Depletion Effect. Perspect. Psychol. Sci. 11, 546–573. doi:10.1177/1745691616652873

Hagger, M.S., Wood, C., Stiff, C., Chatzisarantis, N.L.D., 2010. Ego depletion and the strength model of self-control: a meta-analysis. Psychol. Bull. 136, 495–525. doi:10.1037/a0019486

Harremoës, P., Tishby, N., 2007. The information bottleneck revisited or how to choose a good distortion measure, in: IEEE International Symposium on Information Theory - Proceedings. pp. 566–570. doi:10.1109/ISIT.2007.4557285

Hertz, L., Chen, Y., 2017. Integration between Glycolysis and Glutamate-Glutamine Cycle Flux May Explain Preferential Glycolytic Increase during Brain Activation, Requiring Glutamate. Front. Integr. Neurosci. 11, 18. doi:10.3389/fnint.2017.00018

Hertz, L., Zielke, H.R., 2004. Astrocytic control of glutamatergic activity: Astrocytes as stars of the show. Trends Neurosci. 27, 735–743. doi:10.1016/j.tins.2004.10.008

Hick, W.E., 1952. On the rate of gain of information. Q. J. Exp. Psychol. 4, 11–26. doi:10.1080/17470215208416600

Hockey, G.R., 1997. Compensatory control in the regulation of human performance under stress and high workload; a cognitive-energetical framework. Biol. Psychol. 45, 73–93.

Hockey, G.R.J., 2011. A Motivational Control Theory of Cognitive Fatigue. Cogn. fatigue Multidiscip. Perspect. Curr. Res. Futur. Appl. 167–187. doi:10.1037/12343-008

Holroyd, C.B., 2015. The waste disposal problem of effortful control, in: Motivation and Cognitive Control. pp. 235–260. doi:10.4324/9781315656878

Hyman, R., 1953. Stimulus information as a determinant of reaction time. J. Exp. Psychol. 45, 188–196. doi:10.1037/h0056940

Hyönä, J., Tommola, J., Alaja, A.-M., 1995. Pupil Dilation as a Measure of Processing Load in Simultaneous Interpretation and Other Language Tasks. Q. J. Exp. Psychol. Sect. A 48, 598–612. doi:10.1080/14640749508401407

Inzlicht, M., Bartholow, B.D., Hirsh, J.B., 2015. Emotional foundations of cognitive control. Trends Cogn. Sci. doi:10.1016/j.tics.2015.01.004

Inzlicht, M., Shenhav, A., Olivola, C.Y., 2018. The Effort Paradox: Effort Is Both Costly and Valued. Trends Cogn. Sci. doi:10.1016/j.tics.2018.01.007

Job, V., Dweck, C.S., Walton, G.M., 2010. Ego depletion-is it all in your head? Implicit theories about willpower affect self-regulation. Psychol. Sci. 21, 1686–1693. doi:10.1177/0956797610384745

Kahneman, D., 2011. Thinking , Fast and Slow (Abstract), Book. doi:10.1007/s13398-014-0173-7.2

Kaplan, S., Berman, M.G., 2010. Directed Attention as a Common Resource for Executive Functioning and Self-Regulation. Perspect. Psychol. Sci. 5, 43–57. doi:10.1177/1745691609356784

Kappen, H.J., Gómez, V., Opper, M., 2012. Optimal control as a graphical model inference problem. Mach. Learn. doi:10.1007/s10994-012-5278-7

Kawohl, W., Bruhl, A., Krowatschek, G., Ketteler, D., Herwig, U., 2009. Functional magnetic resonance imaging of tics and tic suppression in Gilles de la Tourette syndrome. World J Biol Psychiatry 10, 567–570. doi:10.1080/15622970802118356

Kingma, D.P., Welling, M., 2013. Auto-Encoding Variational Bayes 1–14. doi:10.1051/0004-6361/201527329

Kleckner, I.R., Zhang, J., Touroutoglou, A., Chanes, L., Xia, C., Simmons, W.K., Quigley, K.S., Dickerson, B.C., Feldman Barrett, L., 2017. Evidence for a large-scale brain system supporting allostasis and interoception in humans. Nat. Hum. Behav. 1, 0069. doi:10.1038/s41562-017-0069

Kloosterman, N.A., Meindertsma, T., van Loon, A.M., Lamme, V.A.F., Bonneh, Y.S., Donner, T.H., 2015. Pupil size tracks perceptual content and surprise. Eur. J. Neurosci. 41, 1068–1078. doi:10.1111/ejn.12859

Koechlin, E., Summerfield, C., 2007. An information theoretical approach to prefrontal executive function. Trends Cogn. Sci. 11, 229–235. doi:10.1016/j.tics.2007.04.005

Kok, P., Jehee, J.F.M., de Lange, F.P., 2012. Less Is More: Expectation Sharpens Representations in the Primary Visual Cortex. Neuron 75, 265–270. doi:10.1016/j.neuron.2012.04.034

Kool, W., McGuire, J.T., Rosen, Z.B., Botvinick, M.M., 2010. Decision making and the avoidance of cognitive demand. J. Exp. Psychol. Gen. 139, 665–682.

Kriegeskorte, N., Bandettini, P., 2007. Analyzing for information, not activation, to exploit high-resolution fMRI. Neuroimage 38, 649–662. doi:10.1016/j.neuroimage.2007.02.022

Kurzban, R., Duckworth, A., Kable, J.W., Myers, J., 2013. An opportunity cost model of subjective effort and task performance. Behav. Brain Sci. 36, 661–679. doi:10.1017/S0140525X12003196

Lai, Y., 2009. Introduction to Arithmetic Coding Theory and Practice. Imid 2009 1069–1072. doi:10.1147/rd.282.0135

Landauer, R., 1996. Minimal Energy Requirements in Communication. Science (80-.). 272, 1914–1918. doi:10.1126/science.272.5270.1914

Laughlin, S.B., 2001. Energy as a constraint on the coding and processing of sensory information. Curr. Opin. Neurobiol. 11, 475–480. doi:10.1016/S0959-4388(00)00237-3

Lavín, C., San Martín, R., Rosales Jubal, E., 2014. Pupil dilation signals uncertainty and surprise in a learning gambling task. Front. Behav. Neurosci. 7. doi:10.3389/fnbeh.2013.00218

Lennie, P., 2003. The cost of cortical computation. Curr. Biol. 13, 493–497. doi:10.1016/S0960-9822(03)00135-0

Lieder, F., Daunizeau, J., Garrido, M.I., Friston, K.J., Stephan, K.E., 2013. Modelling Trial-by-Trial Changes in the Mismatch Negativity. PLoS Comput. Biol. 9. doi:10.1371/journal.pcbi.1002911

MacLeod, C.M., 1991. Half a century of research on the Stroop effect: an integrative review. Psychol. Bull. 109, 163–203. doi:Doi 10.1037//0033-2909.109.2.163

Maisto, D., Donnarumma, F., Pezzulo, G., 2016. Nonparametric problem-space clustering: Learning efficient codes for cognitive control tasks. Entropy. doi:10.3390/e18020061

Maisto, D., Donnarumma, F., Pezzulo, G., 2015. Divide et impera: Subgoaling reduces the complexity of probabilistic inference and problem solving. J. R. Soc. Interface. doi:10.1098/rsif.2014.1335

Manohar, S.G., Chong, T.T.J., Apps, M.A.J., Batla, A., Stamelou, M., Jarman, P.R., Bhatia, K.P., Husain, M., 2015. Reward Pays the Cost of Noise Reduction in Motor and Cognitive Control. Curr. Biol. 25, 1707–1716. doi:10.1016/j.cub.2015.05.038

Marois, R., Ivanoff, J., 2005. Capacity limits of information processing in the brain. Trends Cogn. Sci. 9, 296–305. doi:10.1016/j.tics.2005.04.010

Mars, R.B., Debener, S., Gladwin, T.E., Harrison, L.M., Haggard, P., Rothwell, J.C., Bestmann, S., 2008. Trial-by-Trial Fluctuations in the Event-Related Electroencephalogram Reflect Dynamic Changes in the Degree of Surprise. J. Neurosci. 28, 12539–12545. doi:10.1523/JNEUROSCI.2925-08.2008

Mather, M., Clewett, D., Sakaki, M., Harley, C.W., 2016. Norepinephrine ignites local hotspots of neuronal excitation: How arousal amplifies selectivity in perception and memory. Behav. Brain Sci. 39, e200. doi:10.1017/S0140525X15000667

Mednick, S.C., Nakayama, K., Cantero, J.L., Atienza, M., Levin, A.A., Pathak, N., Stickgold, R., 2002. The restorative effect of naps on perceptual deterioration. Nat. Neurosci. 5, 677. doi:10.1038/nn864

Meyniel, F., Maheu, M., Dehaene, S., 2016. Human Inferences about Sequences: A Minimal Transition Probability Model. PLoS Comput. Biol. 12. doi:10.1371/journal.pcbi.1005260

Miller, E.K., Cohen, J.D., 2001. An integrative theory of prefrontal cortex function. Annu. Rev. Neurosci. 24, 167–202. doi:10.1146/annurev.neuro.24.1.167

Minsky, M., 1961. Steps toward Artificial Intelligence. Proc. IRE 49, 8–30. doi:10.1109/JRPROC.1961.287775

Molden, D.C., Hui, C.M., Scholer, A.A., Meier, B.P., Noreen, E.E., D'Agostino, P.R., Martin, V., 2012. Motivational Versus Metabolic Effects of Carbohydrates on Self-Control. Psychol. Sci. 23, 1137–1144. doi:10.1177/0956797612439069

Moors, A., De Houwer, J., 2006. Automaticity: A Theoretical and Conceptual Analysis. Psychol. Bull. 132, 297–326. doi:10.1037/0033-2909.132.2.297

Muraven, M., Baumeister, R.F., 2000. Self-Regulation and Depletion of Limited Resources : Does Self-Control Resemble a Muscle ? Psychol. Bull. 126, 247–259. doi:10.1037//0033-2909.126.2.247

Mykityshyn, A.L., Fisk, A.D., Rogers, W. a, 2002. Learning to use a home medical device: Mediating age-related differences with training. Hum. Factors 44, 354–364. doi:10.1518/0018720024497727

Nakamura, J., Csikszentmihalyi, M., 2002. The Concept of Flow Optimal Experience and Its Role in Development. Handb. Posit. Psychol. 89–105. doi:10.1007/978-94-017-9088-8_16

Niv, Y., Daw, N.D., Joel, D., Dayan, P., 2006. Tonic dopamine: opportunity costs and the control of response vigor. Psychopharmacology (Berl). 191, 507–520.

Niven, J.E., Laughlin, S.B., 2008. Energy limitation as a selective pressure on the evolution of sensory systems. J. Exp. Biol. 211, 1792–1804. doi:10.1242/jeb.017574

O'Donnell, J., Zeppenfeld, D., McConnell, E., Pena, S., Nedergaard, M., 2012. Norepinephrine: A neuromodulator that boosts the function of multiple cell types to optimize CNS performance. Neurochem. Res. 37, 2496–2512. doi:10.1007/s11064-012-0818-x

O'Reilly, J.X., Schüffelgen, U., Cuell, S.F., Behrens, T.E.J., Mars, R.B., Rushworth, M.F.S., Schuffelgen, U., Cuell, S.F., Behrens, T.E.J., Mars, R.B., Rushworth, M.F.S., 2013. Dissociable effects of surprise and model update in parietal and anterior cingulate cortex. Proc. Natl. Acad. Sci. U. S. A. 110, E3660–E3669. doi:10.1073/pnas.1305373110

Olshausen, B.A., Field, D.J., 2004. Sparse coding of sensory inputs. Curr. Opin. Neurobiol. doi:10.1016/j.conb.2004.07.007

Ortega, P.A., Braun, D.A., 2013. Thermodynamics as a theory of decision-making with information-processing costs Subject Areas : Author for correspondence : Proc. R. Soc. A Math. Phys. Eng. Sci. 469, 20120683–20120683. doi:10.1098/rspa.2012.0683

Ortega, P.A., Braun, D.A., Dyer, J., Kim, K.-E., Tishby, N., 2015. Information-Theoretic Bounded Rationality. doi:10.3390/e16084662

Overath, T., Cusack, R., Kumar, S., Von Kriegstein, K., Warren, J.D., Grube, M., Carlyon, R.P., Griffiths, T.D., 2007. An information theoretic characterisation of auditory encoding. PLoS Biol. 5, 2723–2732. doi:10.1371/journal.pbio.0050288

Park, I.M., Pillow, J.W., 2017. Bayesian Efficient Coding. bioRxiv. doi:10.1101/178418

Paukert, M., Agarwal, A., Cha, J., Doze, V.A., Kang, J.U., Bergles, D.E., 2014. Norepinephrine controls astroglial responsiveness to local circuit activity. Neuron 82, 1263–1270. doi:10.1016/j.neuron.2014.04.038

Payne, J.W., Bettman, J.R., Luce, M.F., 1996. When time is money: Decision behavior under opportunity-cost time pressure. Organ. Behav. Hum. Decis. Process. 66, 131–152. doi:10.1006/obhd.1996.0044

Pezzulo, G., Rigoli, F., Chersi, F., 2013. The mixed instrumental controller: Using value of information to combine habitual choice and mental simulation. Front. Psychol. 4. doi:10.3389/fpsyg.2013.00092

Pezzulo, G., Rigoli, F., Friston, K., 2015. Active Inference, homeostatic regulation and adaptive behavioural control. Prog. Neurobiol. doi:10.1016/j.pneurobio.2015.09.001

Pezzulo, G., Rigoli, F., Friston, K.J., 2018. Hierarchical Active Inference: A Theory of Motivated Control. Trends Cogn. Sci. xx, 1–13. doi:10.1016/j.tics.2018.01.009

Polani, D., 2009. Information: Currency of life? HFSP J. doi:10.2976/1.3171566

Preuschoff, K., 't Hart, B.M., Einhäuser, W., 2011. Pupil dilation signals surprise: Evidence for noradrenaline's role in decision making. Front. Neurosci. 5, 1–12. doi:10.3389/fnins.2011.00115

Purves, D., Augustine, G., Fitzpatrick, D., Katz, L., LaMantia, A.-S., McNamara, J., Williams, M., 2001. Neuroscience. 2nd edition [WWW Document]. Sunderl. Sinauer Assoc. 2001.

Rao, R.P.N., 2010. Decision Making Under Uncertainty: A Neural Model Based on Partially Observable Markov Decision Processes. Front. Comput. Neurosci. 4.

Recarte, M.A., Nunes, L.M., 2000. Effects of verbal and spatial-imagery tasks on eye fixations while driving. J. Exp. Psychol. Appl. 6, 31–43. doi:10.1037/1076-898X.6.1.31

Richter, M., Gendolla, G.H.E., Wright, R.A., 2016. Three Decades of Research on Motivational Intensity Theory: What We Have Learned About Effort and What We Still Don't Know, Advances in Motivation Science. doi:10.1016/bs.adms.2016.02.001

Rubin, O., Meiran, N., 2005. On the origins of the task mixing cost in the cuing task-switching paradigm. J. Exp. Psychol. Learn. Mem. Cogn. 31, 1477–1491. doi:10.1037/0278-7393.31.6.1477

Ryan, R.M., Deci, E.L., 2000. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. Am. Psychol. 55, 68–78. doi:10.1037/0003-066X.55.1.68

Schmidt, L., Lebreton, M., Cléry-Melin, M.-L., Daunizeau, J., Pessiglione, M., 2012. Neural Mechanisms Underlying Motivation of Mental Versus Physical Effort. PLoS Biol. 10, e1001266.

Schneider, W., Chein, J.M., 2003. Controlled & automatic processing: Behavior, theory, and biological mechanisms. Cogn. Sci. doi:10.1016/S0364-0213(03)00011-9

Schneidman, E., Segev, I., Tishby, N., 2000. Information capacity and robustness of stochastic neuron models, in: Advances in Neural Information Processing Systems 12. pp. 178–184.

Sengupta, B., Stemmler, M.B., Friston, K.J., 2013. Information and Efficiency in the Nervous System-A Synthesis. PLoS Comput. Biol. 9. doi:10.1371/journal.pcbi.1003157

Shannon, C.E., 1959. Coding Theorems for a Discrete Source With a Fidelity Criterion. Inst. Radio Eng. Int. Conv. Rec. Vol 7 142–163. doi:10.1234/12345678

Shannon, C.E., 1948. A mathematical theory of communication. Bell Syst. Tech. J. 27, 379–423. doi:10.1145/584091.584093

Sharpee, T.O., Sugihara, H., Kurgansky, A. V., Rebrik, S.P., Stryker, M.P., Miller, K.D., 2006. Adaptive filtering enhances information transmission in visual cortex. Nature 439, 936–942. doi:10.1038/nature04519

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. Annu. Rev. Neurosci. 40, 99–124. doi:10.1146/annurev-neuro-072116-031526

Siclari, F., Tononi, G., 2017. Local aspects of sleep and wakefulness. Curr. Opin. Neurobiol. 44, 222–227. doi:10.1016/j.conb.2017.05.008

Simon, H.A., 1956. Rational choice and the structure of the environment. Psychol. Rev. 63, 129–138. doi:10.1037/h0042769

Simoncelli, E.P., 2003. Vision and the statistics of the visual environment. Curr. Opin. Neurobiol. doi:10.1016/S0959-4388(03)00047-3

Simoncelli, E.P., Olshausen, B.A., 2001. Natural image statistics and neural representation. Annu Rev Neurosci 24, 1193–1216. doi:10.1146/annurev.neuro.24.1.1193

Sims, C.R., 2016. Rate-distortion theory and human perception. Cognition 152, 181–198. doi:10.1016/j.cognition.2016.03.020

Smith, E.C., Lewicki, M.S., 2006. Efficient auditory coding. Nature 439, 978–982. doi:10.1038/nature04485

Sokoloff, L., 2009. Local cerebral energy metabolism: its relationships to local functional activity and blood flow. Cereb. Vasc. smooth muscle its Control.

Sokoloff, L., Mangold, R., Wechsler, R.L., Kenney, C., Kety, S.S., 1955. The effect of mental arithmetic on cerebral circulation and metabolism. J. Clin. Invest. 34, 1101–1108. doi:10.1172/JCI103159

Solopchuk, O., Alamia, A., Dricot, L., Duque, J., Zénon, A., 2017. cTBS disruption of the supplementary motor area perturbs cortical sequence representation but not behavioural performance. Neuroimage 163, 34–40. doi:10.1016/j.neuroimage.2017.09.013

Solopchuk, O., Alamia, A., Olivier, E., Zénon, A., 2016. Chunking improves symbolic sequence processing and relies on working memory gating mechanisms 23, 108–112. doi:10.1101/lm.041277.115

Stephan, K.E., Manjaly, Z.M., Mathys, C.D., Weber, L.A.E., Paliwal, S., Gard, T., Tittgemeyer, M., Fleming, S.M., Haker, H., Seth, A.K., Petzschner, F.H., 2016. Allostatic Self-efficacy: A Metacognitive Theory of Dyshomeostasis-Induced Fatigue and Depression. Front. Hum. Neurosci. 10. doi:10.3389/fnhum.2016.00550

Still, S., Sivak, D.A., Bell, A.J., Crooks, G.E., 2012. Thermodynamics of prediction. Phys. Rev. Lett. 109, 1–5. doi:10.1103/PhysRevLett.109.120604

Stoianov, I., Genovesio, A., Pezzulo, G., 2016. Prefrontal goal codes emerge as latent states in probabilistic value learning. J. Cogn. Neurosci. doi:10.1162/jocn_a_00886

Tanaka, M., 2015. Effects of Mental Fatigue on Brain Activity and Cognitive Performance: A Magnetoencephalography Study. Anat. Physiol. s4. doi:10.4172/2161-0940.S4-002

Teichner, W.H., Krebs, M.J., 1974. Laws of visual choice reaction time. Psychol. Rev. 81, 75–98. doi:10.1037/h0035867

Tishby, N., Pereira, F.C., Bialek, W., 2000. The information bottleneck method 1–11. doi:10.1108/eb040537

Tishby, N., Polani, D., 2011. Information Theory of Decisions and Actions, in: Perception-Action Cycle. doi:10.1007/978-1-4419-1452-1_19

Tkačik, G., Bialek, W., 2014. Information processing in living systems 1–21. doi:10.1146/annurev-conmatphys-031214-014803

Toni, I., Krams, M., Turner, R., Passingham, R.E., 1998. The time course of changes during motor sequence learning: a whole-brain fMRI study. Neuroimage 8, 50–61. doi:10.1006/nimg.1998.0349

Tucker, R., Noakes, T.D., 2009. The physiological regulation of pacing strategy during exercise: A critical review. Br. J. Sports Med. doi:10.1136/bjsm.2009.057562

Vaishnavi, S.N., Vlassenko, A.G., Rundle, M.M., Snyder, A.Z., Mintun, M.A., Raichle, M.E., 2010. Regional aerobic glycolysis in the human brain. Proc. Natl. Acad. Sci. 107, 17757–17762. doi:10.1073/pnas.1010459107

van der Linden, D., Frese, M., Meijman, T.F., 2003. Mental fatigue and the control of cognitive processes: effects on perseveration and planning. Acta Psychol. (Amst). 113, 45–65.

van der Wel, P., van Steenbergen, H., 2018. Pupil dilation as an index of effort in cognitive control tasks: A review. Psychon. Bull. Rev. 1–11. doi:10.3758/s13423-018-1432-y

Verghese, P., Pelli, D.G., 1992. The information capacity of visual attention. Vision Res. 32, 983–995.

Volkow, N.D., Fowler, J.S., Wang, G.J., Telang, F., Logan, J., Wong, C., Ma, J., Pradhan, K., Benveniste, H., Swanson, J.M., 2008. Methylphenidate decreased the amount of glucose

needed by the brain to perform a cognitive task. PLoS One 3. doi:10.1371/journal.pone.0002017

Wacongne, C., Changeux, J.-P., Dehaene, S., 2012. A Neuronal Model of Predictive Coding Accounting for the Mismatch Negativity. J. Neurosci. 32, 3665–3678. doi:10.1523/JNEUROSCI.5003-11.2012

Wei, X.-X., Stocker, A.A., 2015. A Bayesian observer model constrained by efficient coding can explain "anti-Bayesian" percepts. Nat. Neurosci. 18, 1509–1517. doi:10.1038/nn.4105

Westbrook, A., Braver, T.T.S., 2015. Cognitive effort: A neuroeconomic approach. Cogn. Affect. Behav. Neurosci. 15, 395–415. doi:10.3758/s13415-015-0334-y

Westbrook, A., Kester, D., Braver, T.S., 2013. What Is the Subjective Cost of Cognitive Effort? Load, Trait, and Aging Effects Revealed by Economic Preference. PLoS One 8, 1–8. doi:10.1371/journal.pone.0068210

Wiestler, T., Diedrichsen, J., 2013. Skill learning strengthens cortical representations of motor sequences. Elife 2013. doi:10.7554/eLife.00801

Wifall, T., Hazeltine, E., Toby Mordkoff, J., 2016. The roles of stimulus and response uncertainty in forced-choice performance: an amendment to Hick/Hyman Law. Psychol. Res. 80, 555–565. doi:10.1007/s00426-015-0675-8

Wu, T., Dufford, A.J., Egan, L.J., Mackie, M.-A., Chen, C., Yuan, C., Chen, C., Li, X., Liu, X., Hof, P.R., Fan, J., 2017. Hick–Hyman Law is Mediated by the Cognitive Control Network in the Brain. Cereb. Cortex 1–16. doi:10.1093/cercor/bhx127

Wylie, G., Allport, A., 2000. Task switching and the measurement of "switch costs." Psychol. Res. 63, 212–233. doi:10.1007/s004269900003

Xie, L., Kang, H., Xu, Q., Chen, M.J., Liao, Y., Thiyagarajan, M., O'Donnell, J., Christensen, D.J., Nicholson, C., Iliff, J.J., Takano, T., Deane, R., Nedergaard, M., Xie, L., Kang, H., Xu, Q., Chen, M.J., Liao, Y., Thiyagarajan, M., Donnell, J.O., Christensen, D.J., Nicholson, C., Iliff, J.J., Takano, T., Deane, R., Nedergaard, M., 2013. Sleep Drives Metabolite Clearance from the Adult Brain. Science (80-.). 373, 373–377. doi:10.1126/science.1241224

Yu, A.J., Dayan, P., 2003. Expected and unexpected uncertainty: ACh and NE in the neocortex. Adv. neural Inf. Process. … 15, 157–164. doi:citeulike-article-id:496920

Zénon, A., Devesse, S., Olivier, E., 2016. Dopamine manipulation affects response vigor independently of opportunity cost. J. Neurosci. 36. doi:10.1523/JNEUROSCI.4467-15.2016

Zénon, A., Krauzlis, R.J.R.J., 2012. Attention deficits without cortical neuronal deficits. Nature 489, 434–7. doi:10.1038/nature11497

Zhao, S., Song, J., Ermon, S., 2017. Learning Hierarchical Features from Deep Generative Models, in: Proceedings of the 34th International Conference on Machine Learning. pp. 4091–4099.

Figure legends

Figure 1. Probability distribution of responses in a simple digit-key association task, before (light gray) and after (dark gray) seeing the to-be-pressed digit on the screen.

Figure 2. Summary of the present theoretical framework. The upper panel, shaded in green, illustrates the mathematical formulation of information costs *C*. The lower panel, shaded in blue, shows the three proposals that we make regarding how these information costs *C* could translate into subjective effort *F*.

Figure 3. Schematic illustration of the probability distribution of different chess moves. The large size of the state-action space leads to small prior probabilities for all options. The information cost associated with the selection of the final choice is shown as a thick blue line, which represents the KL divergence between the prior and the posterior probability distribution of the chess moves. Log probabilities are shown on the y-axis for consistency with the mathematical definition of the KL divergence, such that for the chosen action, the sum of its log prior probability $log(p_0)$ and KL equals 0 (since KL equals to $-log(p_0)$, given that $p(y|x',T)$ for the chosen action is equal to one).

Figure 4. Example of the difference in information costs between learned and novel task contingencies. The example is inspired by the (Chalk et al., 2010) study, where subjects learn to perform a motion discrimination task in which motion direction is distributed non-uniformly. After learning (in blue), the cost of performing the task when presented with a frequent motion direction is smaller than before learning (in orange). Conventions are similar to Figure 3.

Figure 5. A. Simulation of Stroop task. The different sources of information cost are shown for word-reading (WR) and colour-naming (CN) Stroop trials. Units of information cost are in nats, since they were computed with natural logs. The inset shows the assumed probability of the two task contexts. The larger probability of word-reading task leads to biased automatic output *p(y|x')*, which explains the difference in the information cost between the two conditions. The present simulation assumes three possible colours. The different values are computed as follows (here x' is assumed to be equal to x, representing the different stimulus configurations): $p_0(T) = \{0.9, 0.1\}$, while $p_{opt}(T) = \{1, 0\}$ for the word-reading task and {0, 1} for the colour reading task.

$p(x') = \sum_{T,y} p(x', y \mid T) p_{opt}(T)$ and $p(y) = \sum_{T,x'} p(x', y \mid T) p_{opt}(T)$

$p_{opt}(y \mid x') = \frac{\sum_T p(x', y \mid T) p_{opt}(T)}{p(x')}$ while $p_0(y \mid x') = \frac{\sum_T p(x', y \mid T) p_0(T)}{p(x')}$

$I(x; x') = H(p(x')) - H(p(x' \mid x)) = H(p(x'))$, since *H(x'|x)* is assumed to be zero

$$I(x'; y) = \sum_{x', y} p_{opt}(y \mid x') p(x') \log \frac{p_{opt}(y \mid x') p(x')}{p(x') p(y)}$$

$$p(T, y \mid x') = \frac{p(x', y, T)}{p(x')}$$

$$I(c; y \mid x') = \sum_{x', y, T} p(x', y, T) \log \frac{p(T, y \mid x')}{p_{opt}(T) p_{opt}(y \mid x')}$$

Perceptual cost = *I(x;x')*; Automatic cost = *I(x';y)*;

Cognitive control cost = $I(c; y \mid x) + KL(p_{opt}(y \mid x') \,\|\, p_0(y \mid x'))$

B. Informal illustration of the marshmallow test, which is an example of a task requiring to counteract default policies and deep priors (Mischel, 2014). Children are told that they will

receive 2 marshmallows if they don't eat the one in front of them. The prior (and/or default policy) associated to consumption of high carbohydrate food is skewed in favour of immediate consumption. This results in KL divergence between prior and posterior being larger for longer delays. These larger information costs associated with delayed consumption may explain, informally, why restraining from eating the marshmallow requires self-control and effort.

Figure 6. Simulation of switching task. Two tasks, A and B, alternate every other trials. Both tasks involve the same stimuli but different stimulus-response associations. A. Total information cost is shown for switch (green) and repeat trials (red). The difference between these two types of trials is entirely explained by the change in *p(T)*, shown in panel B.
B. The change in the modelled *p(T=task A)* is shown as a function of trials. The experiment begins with uniform task (or context) probability distribution *p(T)*. In each trial *p(T)* is updated with a learning rate α of 0.3:

$p_{t+1}(\text{current task}) = p_t(\text{current task}) + \alpha(1 - p_t(\text{current task}))$, and

$p_{t+1}(\text{other task}) = p_t(\text{other task}) + \alpha(0 - p_t(\text{other task}))$. Changes of *p(T)* following switch and repeat trials are shown in green and red, respectively.

Figure 7. Adjustment of information rate as a function of noise levels. A. Schematic relationship between information rate and distortion in an example motion discrimination task in two conditions of motion coherence. Given that distortion is quantified as the mean-squared error and that the signal follows Gaussian distribution with variance $\sigma 2$, distortion can be evaluated as: $D(R) = \frac{\sigma^2}{2^{2R}}$. B. Utility *U* associated to distortion as a function of information rate in the 2 motion coherence conditions (arbitrary concave function:

$U(D(R)) = e^{1-D(R)}$). C. Cost function of information rate. This cost function is independent of motion coherence, leading to overlapping curves. An arbitrary convex function was chosen: $C(R) = R^2$. The dashed lines indicate the cost corresponding to the optimal information rates in both conditions. D. Net value, corresponding to the value associated to distortion levels, to which the cost of information rate is subtracted. The optimal trade-off between distortion and information rate is indicated with the dashed line.