Review of Philosophy and Psychology
https://doi.org/10.1007/s13164-020-00512-7

# The Sense of Effort: a Cost-Benefit Theory of the Phenomenology of Mental Effort

Marcell Székely[1] · John Michael[1,2]

Accepted: 30 September 2020


**Abstract**
In the current paper, we articulate a theory to explain the phenomenology of mental effort. The theory provides a working definition of mental effort, explains in what sense mental effort is a limited resource, and specifies the factors that determine whether or not mental effort is experienced as aversive. The core of our theory is the conjecture that *the sense of effort* is the output of a cost-benefit analysis. This cost-benefit analysis employs heuristics to weigh the current and anticipated costs of mental effort for a particular activity against the anticipated benefits. This provides a basis for spelling out testable predictions to structure future research on the phenomenology of mental effort.

**Keywords** Mental effort · Motivation · Cognitive control · Apathy

## 1 Introduction

The experience of mental effort is a familiar feature of daily life. Driving a car, solving basic math problems, or compiling a grocery list – everyday activities such as these demand sustained attention and self-control to resist distractions and tempting alternatives. As a result, we often experience them as effortful, and sometimes postpone them, neglect to complete them altogether, or perform them with insufficient attention. This can lead to dangerous situations, to financial loss, or to wasted time – as when we wander back and forth among various sections of the supermarket because we have been too lazy to prepare a shopping list to help us navigate efficiently among the aisles.

✉ John Michael
johnmichael.cogsci@gmail.com

Marcell Székely
szekelymarcell@gmail.com

[1] Department of Cognitive Science, Central European University, Budapest, Hungary

[2] University of Stirling, Scotland, UK

Published online: 09 October 2020

Mental effort is not only a salient feature of everyday life: experimental tasks employed throughout the cognitive sciences demand varying degrees of mental effort from participants, i.e. along a continuum from effortless to effortful. On a Stroop task, for example, it feels more effortful to respond correctly to a non-matching stimulus (when the word 'blue' is printed in red font) than to a matching one (when the word 'blue' is printed in blue font), and it also feels particularly effortful to shift back and forth between matching and non-matching trials (MacLeod 1991; Golden and Freshwater 1978). Similarly, copying a statement using our non-dominant hand demands greater cognitive effort than writing with our dominant hand (Petrova 2006). This means that mental effort is a key parameter that must be carefully calibrated even in research that is not directly investigating mental effort.

Given the centrality of mental effort in everyday life and as a key parameter in experimental research in the cognitive sciences, it is no surprise that there has been a wealth of theoretical and empirical research investigating mental effort in recent decades. This has led to the articulation of theoretical models of mental effort (Shenhav et al. 2017; Inzlicht et al. 2018), to new experimental paradigms for investigating mental effort (Apps et al. 2015; Lopez-Gamundi and Wardle 2018), to advances in our understanding of the neural underpinnings of mental effort (Chong et al. 2017), and to new insights into pathologies of motivation and corresponding clinical applications (Le Heron et al. 2018). Nevertheless, a range of fundamental questions remain unanswered, such as: *What is* mental effort? Why does mental effort appear to be a limited resource? What are the factors that determine whether or not mental effort is experienced as aversive?

In the current paper, we articulate a theory to address these basic questions. The core of our theory is the conjecture that *the sense of effort* is the output of a cost-benefit analysis. This cost-benefit analysis employs heuristics to weigh the current and anticipated costs of mental effort for a particular activity against the anticipated benefits.

We begin (Section 2) with a review of current theoretical and empirical research investigating mental effort in general, focusing in particular on the three key questions identified above. Against this background, we present our own theory in Section 3. In Section 4 we formulate a range of novel predictions to be investigated in further research.

## 2 Investigating Mental Effort: Key Concepts and Research Questions

### 2.1 Defining Mental Effort

One central desideratum for a theory of the sense of effort[1] would be to specify a satisfying *definition of mental effort.* As a rough starting point, Inzlicht et al., (2018: 338) characterize mental effort as 'the subjective intensification of mental and/or physical activity in the service of meeting some goal.' While this intuitive characterization provides a rough starting point, it would be highly useful in investigating the functions and mechanisms of mental effort to identify mental effort with some measurable quantity, and to specify a working definition which relates mental effort to other key cognitive constructs while providing a basis for operationalizing mental effort in experimental research. We will now briefly review

[1] We will use the term 'effort' in the remainder of the paper to refer to mental effort. Whether and to what extent our theory may be extended to physical effort is an important question for future research.

two strategies for defining mental effort: one in terms of the function of mental effort, and one in terms of the mechanisms underpinning mental effort.

### 2.1.1 The Functional Strategy

One explanatory strategy is to identify mental effort with a functional role. This is the strategy adopted by Inzlicht et al. (2018): they identify mental effort with '... the process that mediates between how well an organism can potentially perform on some task and how well they actually perform on that task' (338). In more formal terms, Shenhav et al. 2017: 100–101) have proposed the following working definition: 'Effort is what mediates between (a) the characteristics of a target task and the subject's available information-processing capacity and (b) the fidelity of the information-processing operations actually performed, as reflected in task performance.'

This strategy is based upon the observation that mental effort is a key determining factor in task performance. There is reason to doubt, however, that the relationship between capacity, mental effort and performance is as straightforward as this definition implies. This is because it is not always the case that increasing mental effort leads to an increase in performance. For example, it does not do so when the optimal procedure for a task *is* the default procedure, as may be the case for some procedures which have been practiced and automated (Logan 1988). Moreover, we can sometimes exert effort in implementing an unpracticed procedure with the intention of performing a task as well as possible – and yet perform worse as a result because we are not yet sufficiently adept with the procedure, or even because we are mistaken about what the optimal procedure is. To illustrate, consider the following example. Beth is asked which of three cities is the largest: Oxford, Edinburgh or Birmingham. Her first impulse is to respond that it must be Oxford, because Oxford is the most famous and therefore has the highest degree of 'availability' to Beth (Nisbett and Wilson 1977; Tversky and Kahneman 1981; Kahneman and Frederick 2002). However, she is able to catch herself before blurting out this answer, and reflects that Oxford is famous for its university, which may not imply that it has a large population. Edinburgh, on the other hand is the capital of Scotland, so it is likely to have a larger population. As it happens, Beth is still wrong: Birmingham has a larger population than either Oxford or Edinburgh, as Beth may have realized if she had considered that Birmingham was once a major industrial center. In this case, Beth inhibited her first impulse and initiated a search of her long-term memory for relevant information. This means that she invested effort in attempting to solve the problem correctly. However, her investment of effort did not improve her performance.

Similarly, Christie and Schrater point out that an increase in cognitive effort can harm performance on tasks in which expert judgment can be out-performed by simple rules based on quantifiable observations. On these tasks, increased deliberation leads people to overweight their own judgements rather than relying on simple formulae or heuristics which would actually lead to an accurate result. As Christie and Schrater explain: 'One simple way to reconcile these results is to assume the existence of both model-based high-cost deliberative neural computations and low-cost model-free experience-based paths (Daw et al. 2005), with the switch to deliberative decision-making resulting in a reduction in performance in cases where model-free solutions are superior.' (Christie and Schrater 2015: 9).

Thus, while the observation that effort typically increases performance provides an important constraint for the construction of a framework for investigating effort, there is

reason to be hesitant about adopting the proposal to define effort as *the factor* mediating between capacity and performance. A more cautious approach is to regard effort as *a factor* which typically mediates between capacity and performance. If effort is not to be defined solely in terms of its function, then, it will be important to introduce further constraints in developing a working definition. Specifically, it will be helpful to attempt to identify the mechanism by which effort typically serves the function of improving performance. In other words, we may distinguish effort from other factors influencing performance by identifying it with a particular mechanism.

### 2.1.2 The Mechanistic Strategy

There is a broad consensus around the idea to equate the exertion of mental effort with the exercise of cognitive control (Shenhav et al. 2017; Inzlicht et al. 2014). Drawing on the familiar analogy between mental and physical effort, the claim is that mental effort regulates the engagement of cognitive control in the same way as physical effort regulates the engagement of muscles. Assuming, for example, that an individual is able to lift a maximum of 100 kg of weight from the ground onto a table, her level of physical effort will determine what proportion of this maximum total weight she actually lifts onto the table. With a low level of effort, she may lift only 20 kg, whereas with a high level of effort she may lift 80 kg. By the same token, mental effort can be conceptualized as mediating between an individual's capacity to perform a cognitive task and her actual performance. For example, while an individual may be capable of completing a series of math problems with 80% accuracy in five minutes, she may not always reach this maximum performance level. Sometimes, if she does not exercise a high level of cognitive control to maintain focus on the task and to ignore distractions and tempting alternatives, she may achieve only 60% in the same amount of time, or she may take a longer time to achieve 80% accuracy. But how does cognitive control affect information processing in the brain and thereby boost performance on cognitively demanding tasks?

Cognitive control, and its effects on performance on cognitive tasks, has been a central topic of research in cognitive psychology over many decades (Ulrich et al. 2015; Schneider and Chein 2003; Anderson 1983; Logan 1978; Shiffrin and Schneider 1977; Stroop 1935). One widely accepted idea that has emerged from this research is that there is an information processing continuum ranging from automatic processes, which can be simultaneously deployed and are experienced as effortless, to control-dependent processes which are serial and experienced as effortful.

Automatic processing is typically fast, efficient and effortless, occurring only after practice in a consistent environment through consistent mapping of stimuli to responses (Moors and De Houwer 2006; Shiffrin and Schneider 1977). For example, in the case of driving a car, the act of putting the key into the ignition launches a learned sequence of actions stored in long-term memory that proceed quickly, efficiently and effortlessly, without the need of cognitive control or monitoring. Such automatic processing works well for routine actions in familiar environments, although its lack of flexibility makes it less functional in complex, dynamic environments.

By contrast, flexibility is the hallmark of controlled processing (Botvinick and Cohen 2014; Shiffrin and Schneider 1977; Atkinson and Shiffrin 1968). It is achieved

through the engagement of executive functions: *inhibition* of dominant or prepotent responses; *updating* working memory contents in response to changing situational demands; and *shifting* flexibly between tasks (Miyake et al. 2000; Miyake and Friedman 2012).

On the other hand, controlled processing is much slower and more effortful than automatic processing, operating serially rather than in parallel, and requiring constant monitoring and the engagement of executive functions. This may be a clue to understanding why people tend to avoid it when possible (Kool et al. 2010; Apps et al. 2015) – as, indeed, it is unsurprising that many organisms tend to avoid physical as well as mental effort (Templer et al. 2018). However, we do not always avoid mental or physical effort, and it is not always experienced as aversive – for example, people willingly spend their Sunday afternoons completing difficult crossword puzzles or exercising at fitness centers. We will return to this issue in Section 3. Before further discussing the aversive nature of some instances of mental effort, however, it will be important to spell out the idea that cognitive control is a limited resource. Different ways of spelling out this idea have been proposed in the literature, which we will now briefly review.

### 2.2 The Costs of Mental Effort

In approaching the question as to why – and in what sense – cognitive control is a limited resource, we will proceed in two steps, one pertaining to the direct costs of cognitive control and one pertaining to indirect costs (i.e. opportunity costs).

#### 2.2.1 Direct Costs

Despite the widespread agreement that controlled processing is limited, there is less consensus as to the source of its limitation. In particular, there is no consensus regarding the answer to the following question: if cognitive control is a limited resource and the sense of effort is an indicator of the expenditure of this resource, what is the *resource*? Current theorizing offers three possible (mutually compatible) explanations: the limit on cognitive control arises either from metabolic, structural or representational constraints.

Metabolic Constraints: Theories of metabolic constraints link mental effort to limited metabolic resources in the brain that deplete with use (Baumeister and Heatherton 1996). For example, an individual who engages in an effortful task (e.g. suppressing thoughts) over an extended period of time would find her capacity to exert effort diminished at a subsequent task that requires self-control (e.g., controlling emotional expressions, delaying gratification). Indeed, initial evidence provided preliminary support for this hypothesis (Muraven and Baumeister 2000; Hagger et al. 2010). However, recent replication attempts have raised considerable doubts concerning the depletion effect (Hagger et al. 2016). Moreover, evidence suggests that the brain's metabolic demands do not change dramatically during task engagement (Kurzban 2010) – indeed, perhaps increasing by only around 1% compared to resting state (Raichle and Mintun 2006). Finally, metabolic theories do not explain the finding that increasing rewards can lead participants to exert more cognitive effort (Camerer and Hogarth 1999; Jimura et al. 2010) and improve

executive function (Krebs et al. 2010) – i.e. if the metabolic resource were truly limited, it should not be possible for participants to improve their performance.

Structural Constraints: Another line of explanation suggests that the capacity limitations of cognitive control are structural, arising from constraints on working memory storage and maintenance on which control-dependent processing depends (Hunt and Lansman 1986; Anderson 1983). The limitation of working memory capacity is well-established: the number of meaningful units (chunks) that can be kept active in working memory ranges between 3 and 5 for adults (Cowan 2001) and its capacity is limited by mutual interference between simultaneously held representations in working memory (Oberauer et al. 2016). This theory does not obviously explain sequential effects – i.e. why exercising control on one task should lead participants to perform worse on a subsequent task (Muraven and Baumeister 2000; Hagger et al. 2010) – although, as noted, recent replication attempts have raised doubts concerning such effects (Hagger et al. 2016). Like theories based on metabolic constraints, this approach fails to provide an explanation of the finding that increasing rewards can lead participants to exert more cognitive effort (Camerer and Hogarth 1999; Jimura et al. 2010) and improve executive function (Krebs et al. 2010).

Representational Constraints: The third line of explanation proposes that capacity limitations on cognitive control arise from the shared use of representation between tasks (Shenhav et al. 2017; Cohen et al. 1990). The core idea is that such shared representations promote efficient learning at the cost of constraining multi-tasking capacity (Musslick et al. 2016). Specifically, shared representations support inference and generalization, but they also give rise to processing interference and conflict during concurrent task execution. To prevent this detrimental crosstalk in the system, the brain limits the number of processes relying on shared representations by engaging cognitive control (Musslick and Cohen 2019) – i.e. by inhibiting some of these processes. Like theories based on metabolic and structural constraints, this approach fails to provide an explanation of the finding that increasing rewards can lead participants to exert more cognitive effort and to improve executive function.

### 2.2.2 Opportunity Costs

A separate question arises from the limited nature of the resource (whatever that resource might be) that is invested. Specifically, the limited nature of the resource requires the brain to prioritize – i.e. the engagement of cognitive control on one task implies that it cannot be devoted to other tasks, meaning that cognitive control always involves opportunity costs (Kurzban et al. 2013). Indeed, insofar as the engagement of cognitive control typically involves the inhibition of default procedures, there is a further source of opportunity costs arising from the failure to perform those default procedures.

Opportunity costs arise in three ways corresponding to the three executive functions: inhibition, updating and shifting (Miyake et al. 2000; Miyake and Friedman 2012). First, *inhibition* creates opportunity costs by stopping fast, parallel and automatic processes aimed at creating benefits for the organism. The loss of these benefits may be registered as an opportunity cost for the organism. For example, in order to boost one's performance on a task that requires one to focus on information presented on a computer screen, such as proofreading

a text, one may need to inhibit the impulse to gaze around the room. This implies that one suspends the default patterns of visual information gathering. Insofar as these patterns are likely to have been shaped by evolution and individual learning to serve the function of acquiring information, it is costly to suspend them. Second, *updating* also creates opportunity costs. Whenever the content of working memory is filled up with the contents of one task, this excludes the contents of other tasks that could generate benefits. Third, *shifting* from one task to another induces a natural extension in the time it takes a person to respond, during which time period the previous task process could have created benefits.

Opportunity costs theories provide an elegant explanation of the finding that increasing rewards can lead participants to exert more cognitive effort (Camerer and Hogarth 1999; Jimura et al. 2010) and improve executive function (Krebs et al. 2010) – namely, because increasing the reward value of a task decreases the relative opportunity costs of alternative tasks. On the other hand, research on the so-called depletion effect using sequential task paradigms appears to put pressure on opportunity costs theories. For example, when people engage in a demanding activity at timepoint 1, performance typically decreases on a different task at timepoint 2. For example, relative to participants who were instructed to write freely, participants who were instructed to inhibit the use of common letters while writing at timepoint 1 were less effective at recalling strings of digits in reverse order at timepoint 2 (Schmeichel 2007). This pattern appears to support the hypothesis that the investment of mental effort depletes some resource; it is not obvious how opportunity costs theorists should account for it. However, as noted above, recent replication attempts have raised doubts concerning the depletion effect (Hagger et al. 2016).

One possible response to concerns about depletion effects derives from a version of the opportunity costs theory which also incorporates elements of the aforementioned representational constraints hypothesis. The idea explains how indirect costs arise from the engagement of cognitive control (the engagement of cognitive control on one task implies the inhibition of other task processes). Specifically, Musslick et al. (2018) have proposed that representational constraints on cognitive control reflect an optimal solution to the cognitive stability-flexibility dilemma. When we allocate a high level of cognitive control to one specific task, this has the effect that we are less well configured for other tasks. As a result, the more control we apply to one task, the longer it takes to reconfigure the shared representational pathways to adjust to new task demands, meaning that our exercise of control leads us to sacrifice flexibility for stability. Insofar as we anticipate that we will soon have to switch to another task, then, it would make sense to resist engaging cognitive control on our current task. This anticipated switch cost constitutes the cost of control. Crucially, this hypothesis predicts that the cost of control depends on the stability or flexibility of the environment, that is, on task switch probability.

Responding to similar concerns, Christie and Schrater (2015) have proposed a hybrid theory which combines opportunity costs and metabolic constraints. It is based on the idea that cognitive costs arise from intelligent resource allocation *over time*. They write:

> 'We suggest that an individual's decision of whether or not to incur cognitive costs in a given situation can be fruitfully understood as one of decision making strategy: an agent will only commit limited resources in cases where the payoff is worth it. Unlike 'cost/benefit' models, however, we treat resources as

dynamically utilized and replenished. Much like a marathon runner, an agent attempting to optimize long-term performance may choose to purposefully limit exertion in order to maintain resource reserves for future use. What may appear to be aversion to cognitive effort may in fact be strategic resource allocation'(2015: 2).

This account differs from standard metabolic constraints theories in that it does not attribute sequential effects to the absence of energetic resources following upon performance of cognitively demanding tasks, but, rather, to the strategic husbanding of cognitive resources. This means that it is well-placed to explain why increasing reward increases performance: when the reward value of a task increases, it is worth exerting more effort now and sacrificing the potential rewards to be gained on later tasks.

### 2.3 The Phenomenology of Mental Effort

In view of these various costs of cognitive control, it is no wonder that we often (though not always) experience mental effort as aversive. In particular, the evolutionary rationale is that the aversive experience of mental effort may help us to avoid paying overly high costs, and may also help us to anticipate when increasing costs may lead to a deterioration of performance. This basic idea goes back at least as far as Hull (1943), who formulated the influential 'Law of Least Effort': agents select actions in order to minimize the effort required for reinforcement. Though the law of least effort was initially formulated in relation to physical effort, Kool et al. (2010) have also extended the validity of this assumption to mental effort as well. In their novel behavioral paradigm, participants faced a recurring choice between two alternative lines of action, associated with different levels of cognitive demand. They exhibited a clear bias toward the less demanding option. In a similar vein, it has also been demonstrated that people are willing to accept lesser rewards to avoid mentally effortful actions (Apps et al. 2015).

There is also neuroimaging evidence that demonstrates that the reward network of the brain shows decreasing activity with increasing effort requirements, and higher default mode network activity correlates with higher effort avoidance (Sayali and Badre 2017). Similarly, a mismatch between task difficulty and individual ability is associated with lower levels of intrinsic reward and corresponds to increased activity within the default mode network (Huskey et al. 2018).

Importantly, however, mental effort (and indeed physical effort) is not always experienced as aversive. Indeed, one important open challenge is to identify the circumstances under which mental effort is experienced as aversive and when it is experienced as pleasurable. In the rest of this section, we will briefly review research bearing upon this challenge.

One line of evidence that shows that effort can be experienced as pleasant comes from research on *learned industriousness*. This research provides a framework for understanding when and how *task-extrinsic rewards* can be used to make the exertion of cognitive effort pleasurable rather than aversive. According to the principles of associative learning, if high effort is consistently paired with high reward, the effort itself can become a secondary reinforcer (Eisenberger 1992), and the reinforced high

effort generalizes across behaviours. For example, if a child is repeatedly praised for her exertion of effort after task execution, instead of her performance, then in the future she will tend to take up learning goals – i.e. goals that require the engagement of cognitive control (Dweck and Leggett 1988).

While research on learned industriousness explores the ways in which task-extrinsic rewards can lead to positive experiences of effort, there is also research investigating how task-intrinsic rewards can make the exertion of effort pleasurable rather than aversive. In particular, it is worth considering research on self-determination theory (Deci and Ryan 1985) and flow theory (Getzels and Csikszentmihalyi 1976), both of which illuminate the conditions under which the exertion of cognitive control is experienced as pleasant rather than aversive. Self-determination theory suggests that the intrinsically rewarding nature of self-determined choice can elicit increases in task enjoyment and performance (Lewthwaite et al. 2015; Leotti and Delgado 2011) while flow theory suggests that the state of flow, in which task difficulty is in balance with individual ability and the highest levels of intrinsic reward are reached, results from a network synchronization process between structures within cognitive control and reward networks (Huskey et al. 2018). It is not yet clear, however, why these two factors – self-determination and the experience of flow – make the experience of mental effort pleasant, i.e., whether there is one common underlying mechanism which they engage. It would be desirable for a theory of the sense of effort to explain why these factors have this effect, and also to establish a basis for identifying other factors that would also make the experience of effort pleasant.

There is also evidence that people vary in how they value the exertion of mental effort. A short form of assessing individual differences in need for cognition (NFC) was developed by Cacioppo et al. (1984), where need for cognition refers to an individual's tendency to engage in and enjoy effortful cognitive endeavours. The NFC scale has proved to be a reliable measure of the value and individual places on the exertion of cognitive effort, e.g. it correctly predicts the amount of money an individual will forego to avoid a cognitive effortful activity (Westbrook et al. 2013) or an individual's extrinsic reward-induced cognitive effort expenditure (Sandra and Otto 2018). While it is well-known that there are individual differences in the degree to which people are willing to invest mental effort and the degree to which they find it aversive, we lack a systematic understanding of the cognitive and motivational causes underlying these differences. Moreover, there has been little research exploring *intra*-individual differences across time and between different contexts. A comprehensive theory of the sense of effort should aim to illuminate the underlying cognitive and motivational causes of these differences.

## 3 A Cost-Benefit Theory of the Sense of Effort

So far, we have explored several key issues for a theory of the sense of effort to address, in particular focusing on the following three questions:

- How should we define mental effort?
- Why does mental effort appear to be a limited resource?

– What are the factors that determine whether or not mental effort is experienced as aversive?

We will now present a theory to address these three questions.

### 3.1 Mental Effort

Our starting point is to conceptualise mental effort as a measure of the extent to which cognitive control inhibits or modifies current default procedures in order to boost performance of an activity. Like the working definition offered by Shenhav et al. (2017), our working definition links effort to the engagement of cognitive control and to enhanced performance. However, for the reasons discussed above (Section 2.1), we do not believe that the relationship between capacity, effort and performance is as straightforward as Shenhav and colleagues imply. Most importantly, our definition does not make actual performance enhancement a defining feature of effort. Instead, it is sufficient that cognitive control be engaged with the *function* of boosting performance. It is important to emphasize that the term 'function' should not be taken to imply a deliberate choice or intention. Instead, we mean cognitive control is engaged because it raises the likely performance level – i.e. if it did not raise the likely performance level, it would not be engaged. It is also important to emphasize that our working definition does not identify mental effort with the engagement of cognitive control full stop; rather, mental effort is a measure of the extent to which cognitive control inhibits and/or modifies default procedures in order to ensure that the procedures that are implemented are specifically tailored to the task context.

To summarize, it is useful to contrast our definition with that offered by Shenhav et al. (2017). They define mental effort as that which mediates between:

(a) 'the characteristics of a target task and the subject's available information-processing capacity'; and
(b) 'the fidelity of the information-processing operations actually performed, as reflected in task performance' (100–101).

In contrast, we replace (b) with:

(b') the flexible adjustment of information-processing to optimise performance of a specific activity.

### 3.2 The Sense of Effort

Building on this, we conceptualise the *sense of effort* (for mental effort) as the output of a process which tracks the expected costs and benefits of effortful mental activity and weighs them against each other. When the expected costs outweigh the expected benefits, an aversive state is generated, the intensity of which is a measure of the anticipated net costs of the current effortful activity. If, on the other hand, the effortful mental activity is expected to increase benefits more than costs, it is experienced as rewarding.

To spell out this proposal, we will first need to explain what costs and benefits enter into the hypothesized cost-benefit analysis. Our working definition is consistent with

many different (mutually compatible) answers to the question as to what the costs of cognitive control are – i.e. with all of the answers considered in section 2.2. We therefore identify the costs of mental effort as the sum of all of the direct costs as well as the indirect (opportunity) costs identified in section 2.2.

What about the benefits? To answer this question, it is useful to carefully consider the function of mental effort: the investment of mental effort enables us to carry out activities that we would not otherwise be able to perform, or to carry them out at a higher performance level than would otherwise be possible, leading to long-term benefits that make the investment of effort worthwhile.

This raises the question as to how we identify those long-term benefits which are to be weighed against the costs of cognitive control. Our answer to this question is that evolution has equipped us to experience pleasure (or any other experience that is rewarding, such as satisfaction, pride, etc) when the exercise of cognitive control serves to improve our performance on tasks which, over long periods of evolution, would have enhanced our inclusive fitness. Of course, the activities which would have enhanced our fitness over long periods of deep evolutionary history do not map perfectly onto the activities which enhance our individual long-term benefits as individuals living in the context of modern societies. This means that the experience of reward when exercising cognitive control is only an imperfect indicator of the benefits to us in the present environment.

Continuing with this line of thought, we may speculate that the sense of effort is likely to be sensitive to cues that are indicative of high anticipated rewards. Ultimately, the association of cues with rewards will have been shaped over the course of evolutionary history, so the rewards in question should typically be rewards which imply fitness gains. In some cases, however, rewards can also become decoupled from evolutionary benefits, so we might also expect the sense of effort to register cues which have been associated with reward through ontogeny – as documented by the research on learned industriousness discussed above. Nevertheless, we may advance the conjecture that mental effort/cognitive control is likely to have yielded high fitness gains in situations in which an agent needs to gather information, to learn, to plan, to navigate or to be vigilant. This means that we should expect cues to these factors to be registered as benefits, and thus to increase the pleasure of effort.

Research on fluency effects provides some preliminary motivation for this hypothesis (Ackerman and Zalmanov 2012; Koriat 1997; Thompson et al. 2013; Ackerman and Thompson 2017; Koriat et al. 2006). This research shows that when information processing is quick and smooth, people tend to experience a higher degree of confidence in their learning progress and certainty in their judgments. In the present context, we may interpret this as indicating that smooth information processing provides a cue that the current investment of cognitive resources is tending to yield gains in learning progress or in accurate judgments about the current environment. Of course, this does not yet entail that this current learning progress or these currently accurate judgments are particularly rewarding. However, our proposal generates the prediction that fluency should reduce aversiveness of effort investment when there is a high anticipated reward value of learning progress or of accurate judgments.

On the other hand, we should expect that reductions in these factors should decrease from the reward value of effort. For example, this line of thought suggests that effortful

attempts to learn in unpredictable environments – even if accompanied by a sense of fluency or by rapid progress – should be experienced as aversive. This is because in unpredictable environments, what one learns ceases to be useful once it has been learned (i..e because what has been learned is no longer applicable in the new environment). And indeed, this conjecture is further motivated by results from agent-based modeling reported by Musslick et al. (2019). Specifically, they found that the imposition of limitations upon control impaired performance of any given task, *but reduced the costs associated with task switches*. Because of this, they conclude, the optimal level of control is lower in environments with a higher probability of task switches. Empirical research will be needed in order to probe whether actual human participants conform to this pattern.

## 4 Putting Theory to the Test

It will be important for further research to identify and test unique predictions generated by the theory. To this end, one general strategy would be to identify conditions under which the hypothesized cost-benefit analysis would either generate a more positively valenced experience of effort than one might otherwise expect (i.e., in the absence of the theory), or vice versa. In order to identify such conditions, we should recall that the theory predicts that effort will be experienced as positive whenever the expected rewards outweigh the expected costs, and otherwise as negative. This means that the key to identifying conditions under which the theory leads to unique testable predictions will be to examine the perceived costs and benefits which may modulate the sense of effort.

With this in mind, we can predict that an agent may come to find an effortful task pleasurable if it has been repeatedly paired with a reward. In other words, they may come to find the investment of effort on the task to be pleasurable, not just to be worth enduring for the sake of the reward. If so, then they may come to prefer a difficult version of the task to an easy version, even when the reward for both versions is the same. More generally, the experience of effort should be less aversive and more positive under any conditions which would typically (over the course of evolution) indicate a fitness benefit linked to cognitive control. Thus, we should expect a more positive experience of effort in situations in which an agent would typically benefit from flexible cognition – i.e. from exercising cognitive control – to gather information, to learn, to plan, to navigate or to be vigilant.

In contrast, the exercise of cognitive control should be experienced as more aversive in situations in which it is typically not beneficial to gather information, to learn, to plan, to navigate or to be vigilant – e.g. in highly unpredictable or changing environments. This implies that it should be possible to increase the aversiveness of effort for the same task by manipulating the perceived predictability/constancy of an environment.

In addition to manipulating perceived benefits of cognitive control, a further possibility would be to manipulate the perceived direct or indirect (opportunity) costs (Kurzban et al. 2013). The opportunity costs of an effortful activity could be increased by manipulating the benefits of an alternative activity. One way to do this would be to increase the benefits of default processes which would otherwise be engaged but which

must be halted in order to engage cognitive control. For example, the expected reward arising from the default process of gazing around freely within an environment would likely be higher in an unfamiliar environment than in a familiar environment, and also higher in a changing environment than in a stable environment. We may therefore predict that cognitive control processes which prevent an individual from gazing freely around would be experienced as more aversive in an unfamiliar environment than in a familiar environment, and also more aversive in a dynamically changing environment than in a more stable environment.

## 5 Outlook

We have presented a novel approach to conceptualizing mental effort, and used this as a starting point in spelling out a theory to explain why we sometimes (but not always) experience mental effort as aversive. This theory is based on the notion of a cost-benefit analysis which employs heuristics to weigh the current and anticipated costs of mental effort for a particular activity against the anticipated benefits. The theory not only provides answers to basic questions which have remained unanswered so far in research on mental effort, but also gives rise to novel predictions which may provide an impetus to further research.

We hope that the theory proposed here will provide a starting point for research investigating mechanisms and the phenomenology of mental effort, and thereby shedding valuable new light upon the cognitive and motivational processes underpinning decision-making, learning and the allocation of cognitive resources in healthy individuals as well as in individuals suffering from motivational disorders such as apathy.



## References

Ackerman, R., and H. Zalmanov. 2012. The persistence of the fluency–confidence association in problem solving. *Psychonomic Bulletin & Review* 19: 1187–1192.

Ackerman, R., and V.A. Thompson. 2017. Meta-reasoning: Monitoring and control of thinking and reasoning. *Trends in Cognitive Sciences* 21 (8): 607–617.

Anderson, J.R. 1983. *The architecture of cognition*. Cambridge: Harvard Univ. press.

Apps, M.A., L.L. Grima, S. Manohar, and M. Husain. 2015. The role of cognitive effort in subjective reward devaluation and risky decision-making. *Scientific Reports* 5: 16880.

Atkinson, R.C., and R.M. Shiffrin. 1968. Human memory: A proposed system and its control processes. *Psychology of Learning and Motivation* 2 (4): 89–195.

Baumeister, R.F., and T.F. Heatherton. 1996. Self-regulation failure: An overview. *Psychological Inquiry* 7 (1): 1–15.

Botvinick, M.M., and J.D. Cohen. 2014. The computational and neural basis of cognitive control: Charted territory and new frontiers. *Cognitive Science* 38 (6): 1249–1285.

Cacioppo, J.T., R.E. Petty, and C. Feng Kao. 1984. The efficient assessment of need for cognition. *Journal of Personality Assessment* 48 (3): 306–307.

Camerer, C.F., and R.M. Hogarth. 1999. The effects of financial incentives in experiments: A review and capital-labor-production framework. *Journal of Risk and Uncertainty* 19 (1–3): 7–42.

Chong, T.T., M. Apps, K. Giehl, A. Sillence, L.L. Grima, and M. Husain. 2017. Neurocomputational mechanisms underlying subjective valuation of effort costs. *PLoS Biology* 15 (2): e1002598. https://doi.org/10.1371/journal.pbio.1002598.

Christie, S.T., and P. Schrater. 2015. Cognitive cost as dynamic allocation of energetic resources. *Frontiers in Neuroscience* 9: 289.

Cohen, J.D., K. Dunbar, and J.L. McClelland. 1990. On the control of automatic processes: A parallel distributed processing account of the Stroop effect. *Psychological Review* 97 (3): 332–361.

Cowan, N. 2001. The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences* 24 (1): 87–114.

Daw, N.D., Y. Niv, and P. Dayan. 2005. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience* 8 (12): 1704–1711.

Deci, E.L., and R.M. Ryan. 1985. The general causality orientations scale: Self-determination in personality. *Journal of Research in Personality* 19 (2): 109–134.

Dweck, C.S., and E.L. Leggett. 1988. A social-cognitive approach to motivation and personality. *Psychological Review* 95 (2): 256–273.

Eisenberger, R. 1992. Learned industriousness. *Psychological Review* 99: 248–267.

Getzels, J. W., & Csikszentmihalyi, M. (1976). The creative vision: A longitudinal study of problem finding in art.

Golden, C. J., & Freshwater, S. M. (1978). Stroop color and word test.

Hagger, M.S., N.L. Chatzisarantis, H. Alberts, C.O. Anggono, C. Batailler, A.R. Birt, et al. 2016. A multilab preregistered replication of the ego-depletion effect. *Perspectives on Psychological Science* 11 (4): 546–573.

Hagger, M.S., C. Wood, C. Stiff, and N.L. Chatzisarantis. 2010. Ego depletion and the strength model of self-control: A meta-analysis. *Psychological Bulletin* 136 (4): 495–525.

Hull, C.L. 1943. *Principles of behavior*. Vol. 422. New York: Appleton-century-crofts.

Hunt, E., and M. Lansman. 1986. Unified model of attention and problem solving. *Psychological Review* 93 (4): 446–461.

Huskey, R., B. Craighead, M.B. Miller, and R. Weber. 2018. Does intrinsic reward motivate cognitive control? A naturalistic-fMRI study based on the synchronization theory of flow. *Cognitive, Affective, & Behavioral Neuroscience* 18 (5): 902–924.

Inzlicht, M., B.J. Schmeichel, and C.N. Macrae. 2014. Why self-control seems (but may not be) limited. *Trends in Cognitive Sciences* 18 (3): 127–133.

Inzlicht, M., A. Shenhav, and C.Y. Olivola. 2018. The effort paradox: Effort is both costly and valued. *Trends in Cognitive Sciences* 22 (4): 337–349.

Jimura, K., H.S. Locke, and T.S. Braver. 2010. Prefrontal cortex mediation of cognitive enhancement in rewarding motivational contexts. *Proceedings. National Academy of Sciences. United States of America* 107: 8871–8876. https://doi.org/10.1073/pnas.1002007107.

Kahneman, D., and S. Frederick. 2002. Representativeness revisited: Attribute substitution in intuitive judgment. *Heuristics and biases: The psychology of intuitive judgment* 49: 81.

Kool, W., J.T. McGuire, Z.B. Rosen, and M.M. Botvinick. 2010. Decision making and the avoidance of cognitive demand. *Journal of Experimental Psychology: General* 139 (4): 665–682.

Koriat, A. 1997. Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology. General* 126: 349–370.

Koriat, A., H. Ma'ayan, and R. Nussinson. 2006. The intricate relationships between monitoring and control in metacognition: Lessons for the cause-and-effect relation between subjective experience and behavior. *Journal of Experimental Psychology: General* 135 (1): 36–69.

Krebs, R.M., C.N. Boehler, and M.G. Woldorff. 2010. The influence of reward associations on conflict processing in the stroop task. *Cognition* 117: 341–347. https://doi.org/10.1016/j.cognition.2010.08.018.

Kurzban, R. 2010. Does the brain consume additional glucose during self-control tasks? *Evolutionary Psychology* 8 (2): 147470491000800208.

Kurzban, R., A. Duckworth, J.W. Kable, and J. Myers. 2013. An opportunity cost model of subjective effort and task performance. *Behavioral and Brain Sciences* 36 (6): 661–679.

Le Heron, C., M.A.J. Apps, and M. Husain. 2018. The anatomy of apathy: A neurocognitive framework for amotivated behaviour. *Neuropsychologia* 118: 54–67.

Leotti, L.A., and M.R. Delgado. 2011. The inherent reward of choice. *Psychological Science* 22 (10): 1310–1318.

Lewthwaite, R., S. Chiviacowsky, R. Drews, and G. Wulf. 2015. Choose to move: The motivational impact of autonomy support on motor learning. *Psychonomic Bulletin & Review* 22 (5): 1383–1388.

Logan, G.D. 1978. Attention in character-classification tasks: Evidence for the automaticity of component stages. *Journal of Experimental Psychology: General* 107 (1): 32–63.

Logan, G.D. 1988. Toward an instance theory of automatization. *Psychological Review* 95 (4): 492–527.

Lopez-Gamundi, P., and M.C. Wardle. 2018. The cognitive effort expenditure for rewards task (C-EEfRT): A novel measure of willingness to expend cognitive effort. *Psychological Assessment* 30 (9): 1237–1248.

MacLeod, C.M. 1991. Half a century of research on the Stroop effect: An integrative review. *Psychological Bulletin* 109 (2): 163–203.

Miyake, A., and N.P. Friedman. 2012. The nature and organization of individual differences in executive functions: Four general conclusions. *Current Directions in Psychological Science* 21 (1): 8–14.

Miyake, A., N.P. Friedman, M.J. Emerson, A.H. Witzki, A. Howerter, and T.D. Wager. 2000. The unity and diversity of executive functions and their contributions to complex "frontal lobe" tasks: A latent variable analysis. *Cognitive Psychology* 41 (1): 49–100.

Moors, A., and J. De Houwer. 2006. Automaticity: A theoretical and conceptual analysis. *Psychological Bulletin* 132 (2): 297–326.

Muraven, M., & Baumeister, R. F. (2000). Self-regulation and depletion of limited resources: Does self-control resemble a muscle?. Psychological bulletin, 126(2), 247. ISO 690.

Musslick, S., & Cohen, J. D. (2019). A Mechanistic Account of Constraints on Control-Dependent Processing: Shared Representation, Conflict and Persistence. In CogSci (pp. 849–855).

Musslick, S., Cohen, J. D., & Shenhav, A. (2019). Decomposing individual differences in cognitive control: A model-based approach. In CogSci (pp. 2427-2433).

Musslick, S., Dey, B., Özcimder, K., Patwary, M. M. A., Willke, T. L., & Cohen, J. D. (2016). Controlled vs. Automatic Processing: A Graph-Theoretic Approach to the Analysis of Serial vs. Parallel Processing in Neural Network Architectures. In CogSci.

Musslick, S., Jang, S. J., Shvartsman, M., Shenhav, A., & Cohen, J. D. (2018). Constraints associated with cognitive control and the stability-flexibility dilemma. In CogSci.

Nisbett, R.E., and T.D. Wilson. 1977. The halo effect: Evidence for unconscious alteration of judgments. *Journal of Personality and Social Psychology* 35 (4): 250–256.

Oberauer, K., S. Farrell, C. Jarrold, and S. Lewandowsky. 2016. What limits working memory capacity? *Psychological Bulletin* 142 (7): 758–799.

Petrova, P. K. (2006). Fluency effects: New domains and consequences for persuasion (Vol. 67, no. 11).

Raichle, M.E., and M.A. Mintun. 2006. Brain work and brain imaging. *Annual Review of Neuroscience* 29: 449–476.

Sandra, D.A., and A.R. Otto. 2018. Cognitive capacity limitations and need for cognition differentially predict reward-induced cognitive effort expenditure. *Cognition* 172: 101–106.

Sayali, C., & Badre, D. (2017). Neural systems of cognitive demand avoidance. BioRxiv, 211375.

Schmeichel, B.J. 2007. Attention control, memory updating, and emotion regulation temporarily reduce the capacity for executive control. *Journal of Experimental Psychology: General* 136 (2): 241–255.

Schneider, W., and J.M. Chein. 2003. Controlled & automatic processing: Behavior, theory, and biological mechanisms. *Cognitive Science* 27 (3): 525–559.

Shenhav, A., S. Musslick, F. Lieder, W. Kool, T.L. Griffiths, J.D. Cohen, and M.M. Botvinick. 2017. Toward a rational and mechanistic account of mental effort. *Annual Review of Neuroscience* 40: 99–124.

Shiffrin, R.M., and W. Schneider. 1977. Controlled and automatic human information processing: II. Perceptual learning, automatic attending and a general theory. *Psychological review* 84 (2): 127.

Stroop, J.R. 1935. Studies of interference in serial verbal reactions. *Journal of Experimental Psychology* 18 (6): 643–662.

Templer, V.L., E.K. Brown, and R.R. Hampton. 2018. Rhesus monkeys metacognitively monitor memories of the order of events. *Scientific Reports* 8 (1): 11541.

Thompson, V.A., J.A.P. Turner, G. Pennycook, L.J. Ball, H. Brack, Y. Ophir, and R. Ackerman. 2013. The role of answer fluency and perceptual fluency as metacognitive cues for initiating analytic thinking. *Cognition* 128 (2): 237–251.

Tversky, A., & Kahneman, D. (1981). *Judgments of and by representativeness* (no. TR-3). STANFORD UNIV CA DEPT OF PSYCHOLOGY.

Ulrich, R., H. Schröter, H. Leuthold, and T. Birngruber. 2015. Automatic and controlled stimulus processing in conflict tasks: Superimposed diffusion processes and delta functions. *Cognitive Psychology* 78: 148–174.

Westbrook, A., D. Kester, and T.S. Braver. 2013. What is the subjective cost of cognitive effort? Load, trait, and aging effects revealed by economic preference. *PLoS One* 8 (7): e68210.