# A Meta-Learning Framework for Rationalizing Cognitive Fatigue in Neural Systems

**Yujun Li**[1], **Rodrigo Carrasco-Davis**[2], **Stefano Sarao Mannelli**[2,3], **Younes Strittmatter**[4], **Sebastian Musslick**[4,5]

[1]Yuanpei College, Peking University, Beijing, China,
[2]Gatsby Computational Neuroscience Unit, University College London, London, UK,
[3]Sainsbury Wellcome Centre, University College London, London, UK,
[4]Dept. of Cognitive, Linguistic, and Psychological Sciences, Brown University, Providence, RI, USA,
[5]Institute of Cognitive Science, Osnabrück University, Osnabrück, Germany
Correspondence: liyujun_cassiel@stu.pku.edu.cn, sebastian.musslick@uos.de

## Abstract

The ability to exert cognitive control is central to human brain function, facilitating goal-directed task performance. However, humans exhibit limitations in the duration over which they can exert cognitive control—a phenomenon referred to as cognitive fatigue. This study explores a computational rationale for cognitive fatigue in continual learning scenarios: cognitive fatigue serves to limit the extended performance of one task to avoid the forgetting of previously learned tasks. Our study employs a meta-learning framework, wherein cognitive control is optimally allocated to balance immediate task performance with forgetting of other tasks. We demonstrate that this model replicates common patterns of cognitive fatigue, such as performance degradation over time and sensitivity to reward. Furthermore, we discuss novel predictions, including variations in cognitive fatigue based on task representation overlap. This approach offers a novel perspective on the computational role of cognitive fatigue in neural systems.

**Keywords:** cognitive control; continual learning; meta-learning; rational boundedness

## Introduction

One of the most remarkable features of cognitive control is our inability to exercise it. On the one hand, cognitive control enables us to quickly adapt information processing in response to changing task demands (Cohen, 2017). On the other hand, cognitive control is fundamentally bounded (Posner & Snyder, 1975; Schneider & Shiffrin, 1977), e.g., in the number of tasks we can execute simultaneously, in the amount of control we can allocate to any given task, and critically, in the duration over which we can exert cognitive control—a limitation referred to as cognitive fatigue.

The framework of rational boundedness seeks to explain limitations (bounds) of cognitive control in terms of adaptations to computational problems inherent to neural systems (Musslick & Cohen, 2021; Musslick & Masis, 2023). Yet, we still lack a computational rationalization for why biological neural systems would exhibit cognitive fatigue. In this work, we explore the hypothesis that cognitive fatigue serves to prevent the forgetting of previously acquired tasks in biological neural systems where task processing is inseparable from task learning.

The traditional view of cognitive fatigue conceptualizes cognitive control as a depleting resource, diminishing as control is exerted over time, and inevitably leading to reduced task performance and engagement (e.g., Baumeister & Heatherton, 1996; Christie & Schrater, 2015). Such resource accounts suggest that cognitive fatigue arises proportionally to the intensity and duration of cognitive effort, prompting individuals to conserve energy for exerting cognitive control. Yet, this perspective has been challenged by recent meta-analyses and replication studies (Carter, Kofler, Forster, & McCullough, 2015; Hagger et al., 2016), as well as studies indicating that enhanced task incentives can mitigate the effects of cognitive fatigue (Matthews et al., 2023; Molden et al., 2012).

Alternative hypotheses propose that cognitive fatigue may signal unwanted metabolic accumulations in the brain, such as amyloid-β (Holroyd, 2015) or extracellular glutamate (Wiehler, Branzoli, Adanyeguh, Mochel, & Pessiglione, 2022), which are hypothesized to accumulate with the extended performance of cognitively demanding tasks. However, such substances may merely act as indicators of fatigue rather than its direct cause, akin to how ghrelin signals hunger without being its primary cause. This leaves the question of which function such indicators of cognitive fatigue may serve.

In contrast to resource or metabolic accounts, computational accounts explain cognitive fatigue in terms of opportunity costs. Kurzban, Duckworth, Kable, and Myers (2013) suggests that prolonged engagement in a single task necessitates sacrificing the performance of alternative tasks, with these forfeited opportunities potentially accumulating as cognitive fatigue. This concept of opportunity cost aligns with the reinforcement learning framework proposed by Agrawal, Mattar, Cohen, and Daw (2022), which posits that cognitive fatigue functions to shift focus from current tasks to the internal replay of prior experiences. According to this account, cognitive fatigue serves to balance immediate action against internal replay, which is needed to improve model-based planning. Nonetheless, while this approach provides a computational rationale for cognitive fatigue, it assumes a trade-off between task performance and offline computation rather than explaining it.

In this computational study, we examine a computational account of cognitive fatigue based on the trade-off between



In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.* 

continued task performance and forgetting of previously acquired tasks in neural systems. In biological networks, unlike their artificial counterparts, learning is intrinsically linked to processing; hence, task execution inherently incurs learning. We postulate that cognitive fatigue may act as a safeguard, deterring a neural system from excessively prolonged task performance to protect against the erosion of memory for tasks with shared representations. We operationalize this hypothesis within a meta-learning framework for linear neural networks. This framework allocates control to optimize expected future rewards across a set of tasks (Carrasco-Davis, Masís, & Saxe, 2023; Shenhav, Botvinick, & Cohen, 2013; Musslick, Shenhav, Botvinick, & Cohen, 2015). These expected rewards are quantified by weighing immediate gains from performing the current task against the forgetting of other tasks, applying analytical solutions to predict learning dynamics (Saxe, McClelland, & Ganguli, 2013). We find that our model reproduces basic patterns of cognitive fatigue, and we discuss novel predictions obtained from the meta-learning framework.

## Task Environment

To simulate cognitive fatigue within a continual learning scenario, we consider a two-layer linear neural network agent performing a sequence of tasks denoted as $\mathbf{T} = \mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_N$ over a time period $T$. Critically, we assess cognitive fatigue once the network has already acquired a set of tasks and is tasked to perform a new (present) task. Accordingly, we divide the task series into two parts:

$$\mathbf{T} = \{\underbrace{\mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_{N-1}}_{\text{previous}}, \underbrace{\mathcal{T}_N}_{\text{present}}\} \tag{1}$$

which are the trajectories of previously acquired tasks $\mathbf{T}_{past} = \{\mathcal{T}_1, \mathcal{T}_2, ..., \mathcal{T}_{N-1}\}$ and the present task $\mathcal{T}_N$. Each task $T_i$ is implemented as a linear input-output mapping given by

$$Y_{\mathcal{T}_i} = W_{\mathcal{T}_i} X_{\mathcal{T}_i} \tag{2}$$

in which $X_{\mathcal{T}_i} \in R^8$ represents an 8-dimensional input vector, $Y_{\mathcal{T}_i} \in R^5$ a 5-dimensional output vector, and $W_{\mathcal{T}_i}$ is the linear mapping implementing the corresponding task.

### Task Mappings

For simulation purposes, we consider a family of tasks described by *teacher networks*. For each task $\mathcal{T}_i$, we determine a random mapping between inputs and labels via a randomly initialized teacher network, mimicking the 2-layer network agent described below (Matiisen, Oliver, Cohen, & Schulman, 2019; Lee, Goldt, & Saxe, 2021). Given the same input data $x$ generated from a normal distribution, each teacher produces a different output $y$, which is determined by the randomly initialized weights of the teacher network. Thus, each task is defined by a different teacher network mapping from $X$ to $Y$.

## Meta-Learning Agent

Our meta-learning agent comprises a two-layer linear neural network model exposed to the task environment described above (Figure 1). Critically, the network implements a mechanism for allocating cognitive control to dynamically adjust its information processing towards the currently relevant task while minimizing forgetting of previously acquired tasks. Below, we describe the network architecture as well as the mechanisms underlying control implementation and allocation within the network.

### Neural Network Architecture

A diagram representing the architecture of the neural network agent is depicted in Figure 1. The network is composed of a two-layers that linearly map the input stimuli $X_i$ into the corresponding responses $Y_i$,

$$Y_i = W_2(t)\, W_1(t)\, X_i. \tag{3}$$

$W_2(t)$ and $W_1(t)$ are the time-dependent weight matrices mapping from the input to the hidden representation and from the hidden representation to the response, respectively. The input, hidden representation, and response have 8, 10, and 5 dimensions, respectively, to accommodate the task environment described above. The network agent is trained using gradient descent on the total mean squared error (MSE) loss $L$ for any given task:

$$\left\langle L_i(\tilde{W}(t), \mathcal{T}_i) \right\rangle = \frac{1}{2}(Y_i - \hat{Y}_i)^2 + \frac{\lambda}{2}\|W\|^2 \tag{4}$$

where $\lambda = 0.001$ is a regularization term.

Our setup assumes linear activation function for the hidden layer for mathematical convenience. Indeed, the linear activation allows for closed-form differential equations of the average learning dynamics, giving explicit access to how continued performance on a given task impacts the forgetting of other tasks. Despite linear activations, the learning dynamics of a multi-layer linear architecture are still non-linear (Saxe et al., 2013), and resembles the trajectory of more complex learning systems (Saxe, McClelland, & Ganguli, 2019; Braun, Dominé, Fitzgerald, & Saxe, 2022). Critically, we can leverage solutions for these learning dynamics to compute the expected amount of forgetting—a factor relevant for optimizing cognitive control in the network.

### Cognitive Control Implementation

Cognitive control serves to bias information processing towards relevant task goals. Following Carrasco-Davis et al. (2023), we implement control signals $G_1$ and $G_2$ as a short-term modulation of the network's weights:

$$\hat{Y}_i = \tilde{W}_2(t)\, \tilde{W}_1(t)\, X_i \tag{5}$$

with

$$\tilde{W}_1(t) = W_1(t) \circ [\mathbb{1} + G_1(t)], \tag{6}$$

$$\tilde{W}_2(t) = W_2(t) \circ [\mathbb{1} + G_2(t)] \tag{7}$$

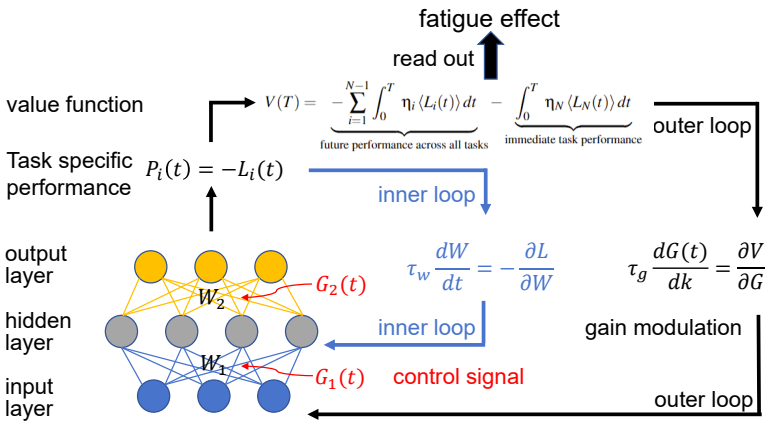


Figure 1: **Illustration of the meta-learning agent.** The agent is characterized by a neural network trained to map stimuli represented in the input layer (blue) via a linear hidden layer (grey) to responses represented in the output layer (yellow). The weights of the network $W_1$, $W_2$ are adjusted to minimize the loss $L_i(t)$ on the task performed at time $t$ (inner loop), which amounts to maximizing performance $P_i(t)$ on that task. The meta-learning agent may deploy cognitive control to minimize the network's loss on the currently performed task $L_i(i = 1, 2, \cdot, N-1)$ while minimizing the loss across all previously acquired tasks $L_N$, maximizing the total value across all tasks and time steps $V(T)$ (outer loop).

where $G_1(t)$ and $G_2(t)$ are control signals matrices, two meta-parameters in the learning framework, and $\mathbb{1}$ is a matrix of 1's with the same dimensions as $G_1$ and $G_2$. The weight modulation of control expressed in Equations 6-7 aligns with the operationalization of cognitive control in connectionist models. In such models, control modulates the sensitivity or responsiveness of task-relevant neurons, acting as a bias of information processing towards relevant tasks (Cohen, Dunbar, & McClelland, 1990; Botvinick, Braver, Barch, Carter, & Cohen, 2001; Musslick et al., 2016; Musslick, Saxe, Novick, Reichman, & Cohen, 2020; Verguts, 2017).

The control signals vectors vary through the entire training period allowing to maximize the overall learning performance (see Carrasco-Davis et al. (2023) for a discussion about how this form of control implements core elements of meta-learning theory). Thus, the control signals vectors have the capacity to simulate various interventions within the learning system. In particular, we consider gain modulation, wherein $G_1(t)$ and $G_2(t)$ have the same dimensions as $W_1$ and $W_2$, and setting $G \equiv 0$ gives back the unmodulated layers representing no intervention (see Equations 6-7).

## Meta-Learning as Control Allocation

We posit that the agent allocates cognitive control across time to maximize overall task performance (Figure 1). In this context, we quantify global task performance in terms of the reward rate, expressed as $r$

$$r = \eta P(t), \tag{8}$$

where $\eta$ is the reward accumulation rate and $P(t)$ denotes task performance intended as negative loss $P(t) = -L(t)$.

We assume that the meta-learning agent seeks to maximize the expected value of cognitive control, comporting with normative models of control allocation (Shenhav et al., 2013; Musslick et al., 2015). In particular, the value of cognitive control is assumed to scale with the total reward accumulated over the entire time period for all given tasks,

$$V = \int_0^T \eta P(t) dt. \tag{9}$$

Critically, we consider that the agent seeks to maximize performance under the assumption that it does not know which tasks are required in the future. As a simple heuristic, we assume that all tasks are equally likely. Thus, the agent may seek to maximize the reward rate on the currently performed tasks while also maximizing the expected performance of all previously acquired tasks. The expected value of control, accumulated across all time steps, can therefore be formulated as follows:

$$V(T) = \underbrace{-\sum_{i=1}^{N-1}\int_0^T \eta_i \langle L_i(t)\rangle dt}_{\text{future performance across all tasks}} - \underbrace{\int_0^T \eta_N \langle L_N(t)\rangle dt}_{\text{immediate task performance}} \tag{10}$$

Here, $L_N(t)$ represents the loss of the current task at time $t$, and $L_i(t)$ is the loss of all previously acquired tasks that the network expects to perform in the future. The coefficient $\eta_i$ denotes the reward for task $\mathcal{T}_i$ and is normalized across tasks.

The objective of the meta-learner is to allocate $G$ so as to maximize $V(T)$. However, identifying the optimal $G$ as a function of time requires computing the loss of the network across all time steps, which in turn requires solving the weight change dynamics in the network as a function of $G$. The dynamics can be analytically evaluated in the gradient flow limit, as demonstrated in earlier work (Saxe et al., 2013,

2019; Braun et al., 2022). Following Carrasco-Davis et al. (2023) in the context of the control mechanism just described, the gradient flow limit of Equation 5 leads to the following ordinary differential equations:

$$\tau_w \frac{dW_1}{dt} = (\tilde{W}_2^T \Sigma_{xy}^T \circ \tilde{G}_1) - (\tilde{W}_2^T \tilde{W}_2 \tilde{W}_1 \Sigma_x) \circ \tilde{G}_1 - \lambda W_1 \quad (11)$$

$$\tau_w \frac{dW_2}{dt} = (\Sigma_{xy}^T \tilde{W}_1^T \circ \tilde{G}_2) - (\tilde{W}_2 \tilde{W}_1 \Sigma_x \tilde{W}_1^T) \circ \tilde{G}_2 - \lambda W_2 \quad (12)$$

where the input correlation matrix $\Sigma_x = \langle X^T X \rangle$ and input-output matrix $\Sigma_{xy} = \langle Y X^T \rangle$ for a single task (where the expectation $\langle \cdot \rangle$ is taken for a given dataset), $\tau_w$ scales the learning rate of the weights, and $\tilde{G}_i = (\mathbb{1} + G_i(t))$.

This result allows us to understand the dynamics of learning as a function of the input and input-output statistics. Notice that the time dynamics of weight learning can be obtained analytically. Furthermore, these equations are linearly additive for multiple tasks, allowing to easily extend the analysis to an arbitrary number of tasks. Finally, observe that, if the dynamics of the control signal is ignored, the model's loss function is solely determined by current labels $Y^N$ and outputs $\hat{Y}^N$, leading to weight updates optimizing only for the current task performance. This can result in catastrophically forgetting previously learnt tasks (Kirkpatrick et al., 2017). In contrast, if control allocation is present across time $G(t)$, the optimization maximizes current task performance while minimizing forgetting on previously acquired tasks, as prescribed by the value function in Equation 10.

### Operationalization of Cognitive Fatigue

Before describing each simulation experiment, we introduce different metrics for cognitive fatigue based on computational and metabolic accounts, operationalized within the meta-learning framework. These metrics will be contrasted in Simulation Experiment 2.

**Computational Metric.** In alignment with the main hypothesis of the paper, we define cognitive fatigue as an alarm signal that indicates a high amount of forgetting on previously acquired tasks induced by performing the current task for an extended period of time. In particular, at a given time $t$, the contribution to the cognitive fatigue rate due to the forgetting of task $i$ is given by

$$f_i(t) = P_i^{best} - P_i(t) = L_i(t) - L_i^{min} \quad (13)$$

where $L_i^{min}$ and $P_i^{best}$ represent the minimum loss and maximum performance, respectively, for task $i$ achieved throughout the entire learning process. Thus, $L_i(t)$ and $P_i(t)$ represent the current loss and the current performance for that task. Under linear activation functions, $L_i(t)$ is computed using the input-output correlation $\Sigma_{x,y}^i$ and the current weights $W_1(t), W_2(t)$ without the control signal term.

The overall fatigue rate is then the sum of the fatigue rates across previously acquired tasks

$$f(t) = \sum_i f_i(t). \quad (14)$$

Finally, we calculate the overall cumulated fatigue $F(T)$ as the integration of the momentary fatigue over time:

$$F(T) = \int_0^T f(t)dt. \quad (15)$$

**Metabolic Metrics.** In addition to the above computational (performance-based) metric for cognitive fatigue, we propose two alternative metrics based on traditional metabolic accounts. The first account assumes that cognitive fatigue rate scales with the *amount* of control allocated at any moment,

$$f_{\text{amount}}(t) = \sum_l \|G_l(t)\|, l = 1,2 \quad (16)$$

where $l$ indexes the weight matrix to which cognitive control is applied at time $t$. This definition reflects the assumption that greater control signals lead to greater depletion of metabolic resources. It also comports with findings that fluctuations of cognitive fatigue can be predicted by the amount of exerted cognitive effort (Matthews et al., 2023).

The second account considers a metabolic metric based on the assumption that cognitive fatigue rate scales with the *change* of control from one timestep to another,

$$f_{\text{change}}(t) = \sum_l \|\frac{dG_l(t)}{dt}\|, l = 1,2 \quad (17)$$

reflecting the assumption that resources only deplete if the control signal needs to be adjusted. Similar to the computational (performance-based) metric, we compute the cumulated fatigue by integrating each metric across all time steps (cf. Equation 15).

## Simulation Experiment 1: Performance-Based Fatigue During Sequential Task Learning

In a first experiment, we examine the computational (performance-based) fatigue metric in the continual learning scenario described above. Specifically, we assess the computational fatigue metric as a function of time spent executing the most recent task in a sequence of four tasks.

**Simulation Procedure.** We simulate the performance of the meta-learning network on four tasks using the teacher network framework. The meta-learning agent is first trained for 200 iterations of three of the four tasks, acquiring a high level of performance. Then, the agent is trained on the fourth task while optimizing the control $G$. In this last phase, we assess computational fatigue while the network performs the fourth task.

**Results.** Figure 2 depicts the total and task-specific momentary fatigue rates in the continual learning scenario. Simulation results indicate that cumulated fatigue generally increases with time on task (Figure 2b). Interestingly, the momentary fatigue rate is greatest during initial task acquisition (Figure 2a,c,e), comporting with observations that the initial learning phase of novel tasks is perceived as effortful and control-demanding (Schneider & Shiffrin, 1977; Shiffrin & Schneider, 1977).

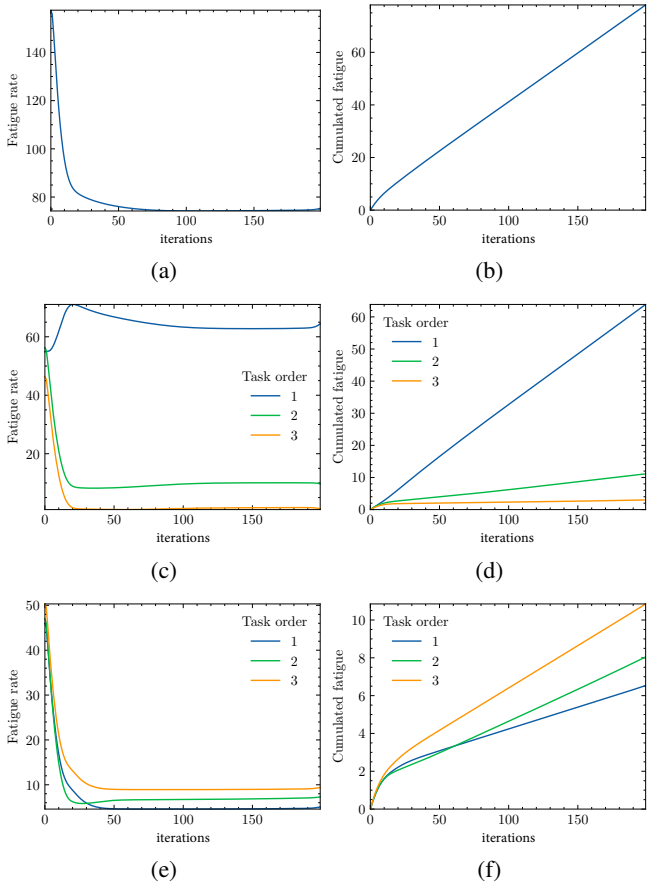


Figure 2: **Computational metric of fatigue in a continual learning scenario.** Each plot depicts the overall (a-b) or task-specific (c-f) fatigue as a function of iterations performing a novel (fourth) task after pre-training the network on three other tasks. The left column depicts momentary fatigue rates, while the right column depicts cumulated fatigue. "Task order" refers to the position of a task within a sequence, where higher numbers represent tasks that the network performed more recently. Panels (a) and (b) depict the overall fatigue rate and overall cumulated fatigue. The fatigue rate is highest at the beginning of the novel task and plateaus just below a value of 80, leading to a linear increase in the cumulated fatigue. Panels (c-f) showcase the fatigue signals elicited by the losses for the three previously acquired tasks individually in two different simulation runs (c-d and e-f). For example, in Panel (c), Task 1, the furthest away in the sequence, exhibits the highest fatigue rate. Panel (d) demonstrates a swift rise in accumulated fatigue shortly after transitioning to Task 4, followed by a steadier, more gradual increase.

## Simulation Experiment 2: Impact of Reward on Computational and Metabolic Fatigue

Cognitive fatigue is known to be mediated by incentives: participants can overcome cognitive fatigue if offered financial incentives to perform the task (Molden et al., 2012). However, it is essential to note that excessively high rewards can lead to adverse effects, known as 'choking under pressure' (Mesagno & Beckmann, 2017). The latter is characterized by a significant psychological burden, leading to impaired performance and control allocation. In Experiment 2, we examine the effects of incentives on both computational and metabolic metrics of cognitive fatigue.

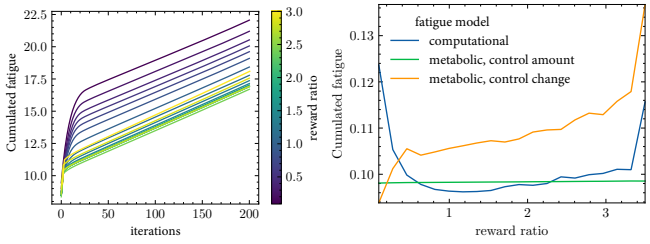


Figure 3: **Computational cumulated cognitive fatigue as a function of reward ratio.** Each color-coded line in (a) represents a different reward ratio, with ratios above 1 reflecting greater reward provided for the currently performed task relative to the previously acquired tasks. The different colored lines in (b) represent the three different metrics of cognitive fatigue: based on the amount of performance decline (computational metric; Equation 13), based on the amount of control (metabolic metric 1, Equation 16), and based on the rate of control change (metabolic metric 2; Equation 17).

**Simulation Procedure.** The network structure and other simulation parameters are the same as in Simulation Experiment 1, except for the task environment. Here, we simulate a task environment with only two tasks and pre-train the network on one of the tasks. In addition, we manipulate the reward ratio $\frac{\eta_i}{\eta_{i-1}}$ between the currently performed task $i$ and the previously performed task $i-1$. A ratio above 1 reflects a greater reward provided to the currently performed task, and a ratio below 1 reflects a greater reward provided to the previously performed task. Based on this setup, we investigate the influence of different reward ratios on cognitive fatigue and task performance.

**Results.** Figure 3a depicts the effect of different reward ratios on computational cumulated fatigue across time. Simulation results indicate that the cumulated fatigue experienced during the currently performed task reduces with greater relative reward provided for the currently performed task, comporting with observations that task incentives can mitigate cognitive fatigue (Molden et al., 2012).

Figure 3b contrasts the computational (performance-based) metric for cumulated fatigue against the two metabolic metrics previously introduced. We find that the metabolic metric based on the control amount is not impacted by the reward ratio. Conversely, the metabolic metric

based on control change increases steadily as the currently performed task receives more reward. Thus, both metrics are inconsistent with the effect of incentive on cognitive fatigue, at least for low reward ratios. The computational (performance) based metric of fatigue resembles an inverted U-shape. When the reward rate of the new task is below the critical value, higher rewards correspond to lower fatigue rates and cumulated fatigue. Curiously, when the task reward surpasses the critical value, greater rewards allocated to performing the current task can intensify cognitive fatigue, mirroring a rise in mental stress, as observed in the 'choking under pressure' phenomenon (Mesagno & Beckmann, 2017).

## General Discussion

The phenomenon of cognitive fatigue, manifesting in decreasing performance and increasing disengagement over time spent on a task, is one of the core limitations of our ability to exert cognitive control. Despite its prevalence, a computational explanation for the emergence of cognitive fatigue in neural systems has been elusive. In this study, we explore the hypothesis that cognitive fatigue serves the computational role of deterring a neural system from excessive engagement in a current task to prevent forgetting of previously acquired tasks. We formalized our hypothesis within a meta-learning framework for continual learning. Specifically, we exposed a neural network agent to a series of tasks and employed meta-learning to optimize control allocation to the currently performed task in order to maximize the performance across all tasks.

The computational problem addressed in this study—to balance continued performance on a task against retention of previously learned tasks—stems from a computational challenge inherent to biological neural systems. Unlike artificial neural networks, where learning and processing can occur independently, biological systems intertwine task execution with learning. This can lead to a computational dilemma where the execution of one task implies the forgetting of other tasks. Thus, conceptually, our work adds to the framework of rational boundedness (Musslick & Masis, 2023; Musslick & Cohen, 2021), aiming to cast seemingly irrational cognitive bounds, such as cognitive fatigue, in terms of a rational response to computational dilemmas inherent to neural processing systems.

Our findings demonstrate that a computational operationalization of cognitive fatigue, based on the amount of forgetting induced by continued task performance, can replicate common patterns of cognitive fatigue, such as a progressive increase in fatigue with time spent on task. Notably, the meta-learning model shows a pronounced increase in fatigue during the initial stages of task learning, aligning with classic theories of cognitive control that identify early task learning as particularly control-demanding (Posner & Snyder, 1975; Schneider & Shiffrin, 1977). Additionally, in low-reward regimes, our computational measure of fatigue accounts for the observed reduction in perceived fatigue with increased task reward. Intriguingly, we also discovered that excessively high rewards for the currently performed tasks can paradoxically enhance cognitive fatigue, echoing the 'choking under pressure' phenomenon (Mesagno & Beckmann, 2017). Moreover, our analysis suggests that metabolic accounts of cognitive fatigue can only partially explain such incentive-related effects, at least when operationalized within the scope of our meta-learning framework.

The computational exploration of cognitive fatigue in this study may offer novel biobehavioral approaches to predicting cognitive fatigue. For example, the present framework, rooted in learning dynamics of neural systems, introduces hypotheses about how structural similarities between tasks—shown to impact the sharing of learned representations between tasks (Musslick et al., 2017; Lee, Mannelli, Clopath, Goldt, & Saxe, 2022)—contribute to cognitive fatigue. Prior research has posited an inverted U-shaped relationship between the degree of representation sharing and catastrophic interference (Lee et al., 2022), with both minimal and extensive representation sharing reducing the risk of catastrophic forgetting, while moderate representation sharing increases this risk. Within the context of our study's framework, we postulate that cognitive fatigue is most pronounced when the task currently being performed shares a moderate amount of representation with previously acquired tasks. This interplay between task representation sharing and cognitive fatigue presents a promising avenue for future empirical evaluation of the present framework.

While the computational model of fatigue introduced here offers initial insights, it necessitates further computational and empirical investigation. Specifically, it requires validation under different task conditions, such as in the Gaussian Mixture framework (Lesieur et al., 2016) or considering structured input data in the teacher framework (Goldt, Mézard, Krzakala, & Zdeborová, 2020; Mannelli, Gerace, Rostamzadeh, & Saglietti, 2022), allowing to investigate the effect of input similarity together with task similarity. Another future direction considers a comparative analysis with other models of fatigue, including resource-based accounts (Matthews et al., 2023) and computation-based accounts (Agrawal et al., 2022), which have effectively explained other aspects of cognitive fatigue. Such a comparison should encompass a wide array of experimental conditions, including variations in task duration, rewards, and task similarity, paving the way for a deeper understanding of computational and neural mechanisms underlying cognitive fatigue.

## Acknowledgements

The authors thank Andrew Saxe for valuable discussions. S.M. was supported by Schmidt Science Fellows, in partnership with the Rhodes Trust, as well as the Carney BRAINSTORM program at Brown University. R.C.-D. and S.S.M. are supported by the Gatsby Charitable Foundation

(GAT3850).

## References


Agrawal, M., Mattar, M. G., Cohen, J. D., & Daw, N. D. (2022). The temporal dynamics of opportunity costs: A normative account of cognitive fatigue and boredom. *Psychological review*, *129*(3), 564.

Ariely, D., Gneezy, U., Loewenstein, G., & Mazar, N. (2009). Large stakes and big mistakes. *The Review of Economic Studies*, *76*(2), 451–469.

Aston-Jones, G., & Cohen, J. D. (2005). An integrative theory of locus coeruleus-norepinephrine function: adaptive gain and optimal performance. *Annu. Rev. Neurosci.*, *28*, 403–450.

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

Beilock, S. L., Kulp, C. A., Holt, L. E., & Carr, T. H. (2004). More on the fragility of performance: choking under pressure in mathematical problem solving. *Journal of Experimental Psychology: General*, *133*(4), 584.

Bijleveld, E., van Breukelen, F. N., de Segovia Vicente, D., & Schutter, D. J. (2023). Mapping the dose–response relationship between monetary reward and cognitive performance. *Motivation Science*.

Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S., & Cohen, J. D. (2001). Conflict monitoring and cognitive control. *Psychological review*, *108*(3), 624.

Braun, L., Dominé, C., Fitzgerald, J., & Saxe, A. (2022). Exact learning dynamics of deep linear networks with prior knowledge. *Advances in Neural Information Processing Systems*, *35*, 6615–6629.

Carrasco-Davis, R., Masís, J., & Saxe, A. M. (2023). Meta-learning strategies through value maximization in neural networks. *arXiv preprint arXiv:2310.19919*.

Carter, E. C., Kofler, L. M., Forster, D. E., & McCullough, M. E. (2015). A series of meta-analytic tests of the depletion effect: Self-control does not seem to rely on a limited resource. *Journal of Experimental Psychology: General*, *144*(4), 796.

Castrillon, G., Epp, S., Bose, A., Fraticelli, L., Hechler, A., Belenya, R., … others (2023). An energy costly architecture of neuromodulators for human brain evolution and cognition. *Science Advances*, *9*(50), eadi7632.

Chalnick, A., & Billman, D. (1988). Unsupervised learning of correlational structure. In *Proceedings of the tenth annual conference of the cognitive science society* (pp. 510–516). Hillsdale, NJ: Lawrence Erlbaum Associates.

Christie, S. T., & Schrater, P. (2015). Cognitive cost as dynamic allocation of energetic resources. *Frontiers in neuroscience*, *9*, 289.

Cohen, J. D. (2017). Cognitive control: core constructs and current considerations. *The Wiley handbook of cognitive control*, 1–28.

Cohen, J. D., Dunbar, K., & McClelland, J. L. (1990). On the control of automatic processes: a parallel distributed processing account of the stroop effect. *Psychological review*, *97*(3), 332.

David, L., Vassena, E., & Bijleveld, E. (2022). The aversiveness of mental effort: A meta-analysis.

Feigenbaum, E. A. (1963). The simulation of verbal learning behavior. In E. A. Feigenbaum & J. Feldman (Eds.), *Computers and thought.* New York: McGraw-Hill.

Ferguson, K. A., & Cardin, J. A. (2020). Mechanisms underlying gain modulation in the cortex. *Nature Reviews Neuroscience*, *21*(2), 80–92.

Finn, C., Abbeel, P., & Levine, S. (2017). Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning* (pp. 1126–1135).

Ford, C. E., Wright, R. A., & Haythornthwaite, J. (1985). Task performance and magnitude of goal valence. *Journal of Research in Personality*, *19*(3), 253–260.

Frömer, R., Lin, H., Dean Wolf, C., Inzlicht, M., & Shenhav, A. (2021). Expectations of reward and efficacy guide cognitive control allocation. *Nature communications*, *12*(1), 1030.

Garbers, Y., & Konradt, U. (2014). The effect of financial incentives on performance: A quantitative review of individual and team-based financial incentives. *Journal of occupational and organizational psychology*, *87*(1), 102–137.

Goldt, S., Advani, M., Saxe, A. M., Krzakala, F., & Zdeborová, L. (2019). Dynamics of stochastic gradient descent for two-layer neural networks in the teacher-student setup. *Advances in neural information processing systems*, *32*.

Goldt, S., Mézard, M., Krzakala, F., & Zdeborová, L. (2020). Modeling the influence of data structure on learning in neural networks: The hidden manifold model. *Physical Review X*, *10*(4), 041044.

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

Hill, J. A. C. (1983). A computational model of language acquisition in the two-year old. *Cognition and Brain Theory*, *6*, 287–317.

Holroyd, C. (2015). The waste disposal problem of effortful control. In *Motivation and cognitive control* (pp. 247–272). Routledge.

Holroyd, C. (2023). The controllosphere: The neural origin of cognitive effort.

Hospedales, T., Antoniou, A., Micaelli, P., & Storkey, A. (2021). Meta-learning in neural networks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, *44*(9), 5149–5169.

Ishii, A., Tanaka, M., & Watanabe, Y. (2014). Neural mechanisms of mental fatigue. *Reviews in the Neurosciences*, *25*(4), 469–479.

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., ... others (2017). Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, *114*(13), 3521–3526.

Klein-Flügge, M. C., Kennerley, S. W., Friston, K., & Bestmann, S. (2016). Neural signatures of value comparison in human cingulate cortex during decisions requiring an effort-reward trade-off. *Journal of Neuroscience*, *36*(39), 10002–10015.

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

Lee, S., Goldt, S., & Saxe, A. (2021). Continual learning in the teacher-student setup: Impact of task similarity. In *International conference on machine learning* (pp. 6109–6119).

Lee, S., Mannelli, S. S., Clopath, C., Goldt, S., & Saxe, A. (2022). Maslow's hammer for catastrophic forgetting: Node re-use vs node activation. *arXiv preprint arXiv:2205.09029*.

Lesieur, T., De Bacco, C., Banks, J., Krzakala, F., Moore, C., & Zdeborová, L. (2016). Phase transitions and optimal algorithms in high-dimensional gaussian mixture clustering. In *2016 54th annual allerton conference on communication, control, and computing (allerton)* (pp. 601–608).

Lewis, C. (1978). *Production system models of practice effects*. Doctoral dissertation, Department of Psychology, University of Michigan, Ann Arbor.

Lieder, F., Shenhav, A., Musslick, S., & Griffiths, T. L. (2018). Rational metareasoning and the plasticity of cognitive control. *PLoS computational biology*, *14*(4), e1006043.

Liljeholm, M., & O'Doherty, J. P. (2012). Anything you can do, you can do better: neural substrates of incentive-based performance enhancement. *PLoS biology*, *10*(2), e1001272.

Mannelli, S. S., Gerace, F., Rostamzadeh, N., & Saglietti, L. (2022). Unfair geometries: exactly solvable data model with fairness implications. *arXiv preprint arXiv:2205.15935*.

Matiisen, T., Oliver, A., Cohen, T., & Schulman, J. (2019). Teacher–student curriculum learning. *IEEE transactions on neural networks and learning systems*, *31*(9), 3732–3740.

Matlock, T. (2001). *How real is fictive motion?* Doctoral dissertation, Psychology Department, University of California, Santa Cruz.

Matthews, J., Pisauro, M. A., Jurgelis, M., Mueller, T., Vassena, E., Chong, T. T.-J., & Apps, M. A. (2023). Computational mechanisms underlying the dynamics of physical and cognitive fatigue. *Cognition*, *240*, 105603.

Mesagno, C., & Beckmann, J. (2017). Choking under pressure: Theoretical models and interventions. *Current opinion in psychology*, *16*, 170-175.

Mobbs, D., Hassabis, D., Seymour, B., Marchant, J. L., Weiskopf, N., Dolan, R. J., & Frith, C. D. (2009). Choking on the money: reward-based performance decrements are associated with midbrain activity. *Psychological science*, *20*(8), 955–962.

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. *Psychological science*, *23*(10), 1137–1144.

Musslick, S., & Cohen, J. D. (2021). Rationalizing constraints on the capacity for cognitive control. *Trends in Cognitive Sciences*, *25*(9), 757–775.

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 *Proceedings of the 38th Annual Meeting of the Cognitive Science Society* (p. 1547–1552).

Musslick, S., & Masis, J. A. (2023). Pushing the bounds of bounded optimality and rationality. *Cognitive Science*, *47*(4), e13259.

Musslick, S., Saxe, A., Novick, A., Reichman, D., & Cohen, J. D. (2020). On the rational boundedness of cognitive control: Shared versus separated representations. *PsyArXiv preprint: https://doi.org/10.31234/osf.io/jkhdf*. doi: 10 .31234/osf.io/jkhdf

Musslick, S., Saxe, A., Özcimder, K., Dey, B., Henselman, G., & Cohen, J. D. (2017). Multitasking capability versus learning efficiency in neural network architectures..

Musslick, S., Shenhav, A., Botvinick, M. M., & Cohen, J. D. (2015). A computational model of control allocation based on the expected value of control. In *Reinforcement Learning and Decision Making Conference 2015.*

Newell, A., & Simon, H. A. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall.

Ohlsson, S., & Langley, P. (1985). *Identifying solution paths in cognitive diagnosis* (Tech. Rep. No. CMU-RI-TR-85-2). Pittsburgh, PA: Carnegie Mellon University, The Robotics Institute.

Poskanzer, C., & Aly, M. (2023). Switching between external and internal attention in hippocampal networks. *Journal of Neuroscience*, *43*(38), 6538–6552.

Posner, M. I., & Snyder, C. (1975). *Attention and cognitive control. information processing and cognition: The loyola symposium.* Hillsdale NJ: Erlbaum.

Ritz, H., Leng, X., & Shenhav, A. (2022). Cognitive control as a multivariate optimization problem. *Journal of Cognitive Neuroscience*, *34*(4), 569–591.

Saxe, A. M., McClelland, J. L., & Ganguli, S. (2013). Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. *arXiv preprint arXiv:1312.6120*.

Saxe, A. M., McClelland, J. L., & Ganguli, S. (2019). A mathematical theory of semantic development in deep neural networks. *Proceedings of the National Academy of*

*Sciences*, *116*(23), 11537–11546.

Schneider, W., & Shiffrin, R. M. (1977). Controlled and automatic human information processing: I. detection, search, and attention. *Psychological review*, *84*(1), 1.

Schnotz, W., & Bannert, M. (2003). Construction and interference in learning from multiple representation. *Learning and instruction*, *13*(2), 141–156.

Shenhav, A., Botvinick, M. M., & Cohen, J. D. (2013). The expected value of control: an integrative theory of anterior cingulate cortex function. *Neuron*, *79*(2), 217–240.

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

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

Shrager, J., & Langley, P. (Eds.). (1990). *Computational models of scientific discovery and theory formation*. San Mateo, CA: Morgan Kaufmann.

Tervo, D. G. R., Kuleshova, E., Manakov, M., Proskurin, M., Karlsson, M., Lustig, A., … Karpova, A. Y. (2021). The anterior cingulate cortex directs exploration of alternative strategies. *Neuron*, *109*(11), 1876–1887.

Verguts, T. (2017). Binding by random bursts: A computational model of cognitive control. *Journal of Cognitive Neuroscience*, *29*(6), 1103–1118.

Vilalta, R., & Drissi, Y. (2002). A perspective view and survey of meta-learning. *Artificial intelligence review*, *18*, 77–95.

Wang, J. X. (2021). Meta-learning in natural and artificial intelligence. *Current Opinion in Behavioral Sciences*, *38*, 90–95.

Westgate, E. C., & Steidle, B. (2020). Lost by definition: Why boredom matters for psychology and society. *Social and Personality Psychology Compass*, *14*(11), e12562.

Wiehler, A., Branzoli, F., Adanyeguh, I., Mochel, F., & Pessiglione, M. (2022). A neuro-metabolic account of why daylong cognitive work alters the control of economic decisions. *Current Biology*, *32*(16), 3564–3575.

Zedelius, C. M., Veling, H., Custers, R., Bijleveld, E., Chiew, K. S., & Aarts, H. (2014). A new perspective on human reward research: How consciously and unconsciously perceived reward information influences performance. *Cognitive, Affective, & Behavioral Neuroscience*, *14*, 493–508.