NEUROSCIENCE & ARTIFICIAL INTELLIGENCE

MECHANISMS, PERSPECTIVES & CONSEQUENCES

23 · 24 May 2024

Centre Broca Nouvelle-Aquitaine, Bordeaux

Société des Neurosciences

POSTER SESSION

www.neurosciences.asso.fr/JT2024/

1
Do glitches in the OFC neural code explain irrational choices? A neuro-computational approach to value synthesis

Juliette Bénon[1], Jean Daunizeau[1,2]

[1] ICM - Institut du Cerveau, Motivation Brain Behavior Team, Paris, France
[2] Federal Institute of Technology, Zürich, Switzerland

The orbitofrontal cortex (OFC) is known to play a key role in integrating environmental features and signaling value. However, the functional contribution of OFC neurons to value-related computations is still unclear. For example, the OFC may be operating the comparison of options' values for decision making. But another possibility is that OFC neurons are integrating value-relevant features into value estimates, leaving their comparison for downstream decision circuits. This distinction is important, because imperfections in the underlying neural system architectures would yield different forms of irrational behavior.

We trained recurrent neural networks (RNNs) to perform either value construction or value comparison, under the constraint that options' features are sampled progressively in time. For both scenarii, we considered different value coding frames, in terms of options spatial position (left/right), temporal order (first/second) or attentional focus (attended/unattended). We then compared their neural signatures to electrophysiological data recorded by Hunt et al. (2018) in the OFC of two macaque monkeys during a binary decision task.

We found that key OFC representational geometry properties could emerge under both functional scenarii, but only in specific value coding frames. Furthermore, when disrupting the RNNs' feedback connections to explain monkeys' irrational choices, the representational geometry of RNNs becomes even more similar to that of OFC neurons. Interestingly, the ensuing sparsity of neural representations increases, while the amount of energy used by the units decreases. This suggests that some forms of irrational behavior may result from sparsity and energy budget constraints on the OFC neural code.

2
Emergence of left lateralization in encoding models of fMRI responses with larger language models

Laurent Bonnasse-Gahot[1], Christophe Pallier[2,3,4]

[1] EHESS, CAMS, Paris, France
[2] Cognitive Neuroimaging Unit, Neurospin, INSERM-CEA Gif-sur-Yvette, France
[3] INSERM-CEA Gif-sur-Yvette, France
[4] Centre National de la Recherche Scientifique

Over the past decade, studies of naturalistic language processing where participants are scanned while listening to continuous text have flourished. Using word embeddings at first, then large language models, researchers have created encoding models to analyze the brain signals. Presenting these models with the same text as the participants allows to identify brain areas where there is a significant correlation between the fMRI time series and the ones predicted by the models' artificial neurons. One intriguing finding from these studies is that

they have revealed highly symmetric bilateral activation patterns, somewhat at odds with the well-known left lateralization of language processing. Here, we report analyses of an fMRI dataset where we manipulate the complexity of large language models, testing 28 pretrained models from 8 different families, ranging from 124M to 14.2B parameters. First, we observe that the performance of a model in predicting brain responses follows a scaling law, where the fit with brain activity increases linearly with the logarithm of the number of parameters (and the performance of the models on NLP task). Second, we show that a left/right asymmetry gradually appears as model size increases, and that the difference in left/right brain correlation also follows a scaling law. Whereas the smallest models show no asymmetry, larger models fit better and better left hemispheric activations than right hemispheric ones. This finding reconciles computational analyses of brain activity using large language models with the classic observation from aphasic patients showing left hemisphere dominance for language.

3
Numerosity representation across changes in object and scene content in convolutional neural networks

Thomas Chapalain[1,2,3], Bertrand Thirion[1,2,3], Evelyn Eger[4,5,6]

[1] Paris-Saclay University, Gif/Yvette, France
[2] CEA
[3] Inria, Gif/Yvette, France
[4] Cognitive Neuroimaging Unit, Neurospin Center Gif/Yvette, France
[5] CEA INSERM
[6] University Paris Saclay

“Number sense”, the ability to rapidly approximate numbers of items without precise counting, is a crucial cognitive capacity of humans and many other animals. Recent studies found that convolutional neural networks (CNNs) exhibit distinguishable responses to visual dot sets varying in quantity, even before training, and proposed such networks as candidate models for the emergence of number sense in biological visual systems. Unlike the highly simplified stimuli used in those studies, real life numerosity perception requires to abstract from visual properties of individual objects and their context of appearance.
Here, we therefore explored how, and with what degree of generalization across non-numerical image properties, numerosity is encoded in population signals at different levels of the hierarchy of CNNs when using original synthetic photo-realistic scenes as stimuli. We found that for several architectures, object categorization trained CNN features allowed for numerosity discrimination across changes in coarse object (animals vs tools) and scene (natural vs artificial) categories in higher convolutional layers, but untrained networks failed at this test. Decoding performance for trained CNNs outperformed that predicted by several low-level image statistics, some of which were previously found to be closely correlated with numerosity in binary dot patterns.
Our results show that object recognition CNNs have highly versatile representations, supporting discriminations beyond those on which they were explicitly trained. They further caution against the use of untrained networks as models for innate numerical abilities and outline the importance of using stimuli with variable visual characteristics to delineate mechanisms of the visual number sense.

4
A connectivity gradient in continuous deep reservoir computing predicts a hierarchy for mixed selectivity in human cortex

Peter Ford Dominey[1,2]

[1]Université de Bourgogne Franche-Comté, Laboratoire CAPS INSERM U1093, DIJON, France
[2]Robot Cognition Laboratory, Marey Institute, Dijon, France

One of the initial motivations for reservoir computing was the effort to understand how recurrent connectivity could explain observations of rich neural activity patterns in the prefrontal cortex of behaving primates. Recurrent connections provide for the projections of inputs into a high dimensional space. In individual reservoir units, this results in activation patterns that display non-linear mixtures of inputs and abstract internal states, referred to as mixed selectivity, which has been identified as a key component of reservoir activity. Interestingly it is also a key element in primate brain activity. An equally prominent characteristic of primate brain activity is a temporal processing gradient, with fast input driven responses in sensory areas, and progressively prolonged time constants in increasingly associative cortical areas. Recent research has explained a temporal integration hierarchy as a function of local connectivity within structured reservoirs. In the current research we test and confirm the hypothesis that this physical hierarchy will also produce a gradient in mixed selectivity in the structured reservoirs. This allows us to predict that the same kind of gradient for mixed selectivity should be observed in the human cortex. Applying the same analysis from the reservoir analysis, we observe the presence of mixed selectivity in human cortex, and evidence for a gradient from lower-level sensorimotor areas, to higher level integrative areas. This research contributes to the characterization of the principals of computation in anatomically structured reservoirs and the human brain.

5
Neuronal features in atypical sensory perception in the Fmr1-/y mouse model of autism

Théo Gauvrit[1], *Ourania Semelidou*[1], *Célien Vandromme*[1], *Adrien Corniere*[1], *Andreas Frick*[1]

[1]Neurocentre Magendie, INSERM U.1215, Mécanismes de la plasticité corticale, Bordeaux, France

Autism spectrum disorder (ASD) is a neurodevelopmental disorder chara–cterized by social impairments and repetitive behaviors. However, the latest ammendment of DSM-5 included altered sensory experience as a core diagnostic criterion, almost universally observed in autistic individuals. In order to explore the neuronal basis of these sensory features we employed the Fmr1-/y mouse model of autism and developed a tactile Go/No-Go task in combination with 2-photon calcium imaging of neocortical neurons. Our results showed hyposensitivity to the tactile stimulus with an increased detection threshold in Fmr1-/y mice, recapitulating observations of human studies with similar stimuli. To investigate the neural underpinnings of this tactile hypo–sensitivity, we explored neocortical activity during the tactile detection task, harnessing machine-learning approaches. Our results showed no difference in the magnitude and duration of the neural responses between the Fmr1-/y group and their WT littermates. However, even though a higher number of neurons was recruited during detected stimuli in WT mice, we did not observe the same responses in Fmr1-/y mice. This last result indicates that the stimulus is not properly encoded in the neocortex of Fmr1-

/y mice. To further investigate this result we created a model based on the neural responses. We observed that stimulus detection in Fmr1-/y mice could not be predicted like in the WT group, indicating that the differences in the number of recruited neurons in the neocortex might be a sufficient feature to explain tactile hyposensitivity.

6
Stick to your Role! Stability of Personal Values Expressed in Large Language Models

Grgur Kovač[1], Rémy Portelas[2], Masataka Sawayama[3], Peter Ford Dominey[4,5], Pierre-Yves Oudeyer[1]

[1]Flowers Team, INRIA, Bordeaux, France
[2]Ubisoft La Forge, Bordeaux, France
[3]Graduate School of Information Science and Technology, The University of Tokyo, Tokyo, Japan
[4]INSERM UMR1093-CAPS, Université Bourgogne, Dijon, France
[5]Robot Cognition Laboratory, Institute Marey, Dijon, France

The standard way to study Large Language Models (LLMs) through benchmarks or psychology questionnaires is to provide many different queries from similar minimal contexts (e.g. multiple choice questions). However, due to LLM's highly context-dependent nature, conclusions from such minimal-context evaluations may be little informative about the model's behavior in deployment (where it will be exposed to many new contexts). We argue that context-dependence should be studied as another dimension of LLM comparison alongside others such as cognitive abilities, knowledge, or model size. In this paper, we present a case-study about the stability of value expression over different contexts (simulated conversations on different topics), and as measured using a standard psychology questionnaire (PVQ) and behavioral downstream tasks. We consider 19 open-sourced LLMs from five families. Reusing methods from psychology, we study Rank-order stability on the population (interpersonal) level, and Ipsative stability on the individual (intrapersonal) level. We explore two settings: with and without instructing LLMs to simulate particular personalities. We observe similar trends in the stability of models and model families - Mixtral, Mistral and Qwen families being more stable than LLaMa-2 and Phi - over those two settings, two different simulated populations, and even on three downstream behavioral tasks. When instructed to simulate particular personas, LLMs exhibit low Rank-Order stability, and this stability further diminishes with conversation length. This paper highlights the need for future research directions on LLMs that can coherently simulate a diversity of personas, provides a foundational step in that direction.
See the project website at https://sites.google.com/view/llmvaluestability.

7
A Domain-general Strategy for Hidden-state Inference in Humans and Noisy Neural Networks

Jun Seok Lee[1]*, Valentin Wyart*[1]

[1]Ecole Normale Supérieure, DEC LNC2, Inserm U960, Paris, France

Understanding the hidden (latent) states and structures that generate observations of the world is a fundamental aspect of cognition, wherein humans demonstrate exceptional proficiency in the ability to apply similar cognitive strategies across superficially dissimilar contexts sharing the same latent structure. While previous efforts to understand the computational bases of these cognitive strategies through cognitive modeling have largely focused on single contexts, here we take a novel approach which combines two tasks with the same reversal structure. Through cognitive modeling, we first show that humans use the same noisy hidden-state inference strategy across these superficially dissimilar tasks that require reversal learning. Then, using recurrent neural networks (RNNs) featuring either exact or noisy computations trained on the same tasks, we show that noisy RNNs – like humans – benefit from reusing the same la- tent representations for solving the two tasks. Together, our findings underscore the significance of computation noise in constraining the use of mental resources, shed- ding light on its potential functional role in cognition.

8
Bioinspired head-to-shoulder reference frame transformation for movement-based arm prosthesis control

Bianca Lento[1]*, Vincent Leconte*[1]*, Lucas Bardisbanian*[1]*, Emilie Doat*[1]*, Effie Ségas*[1]*, Aymar de Rugy*[1]

[1]University of Bordeaux, CNRS, INCIA, UMR 5287, Bordeaux, France

Current myoelectric controls of transhumeral prostheses are not satisfactory, and alternatives based on natural arm coordination are being increasingly explored. We recently showed that adding movement goals to shoulder information enabled Artificial Neural Networks (ANNs), trained on natural arm movements, to predict distal joints so well that transhumeral amputees could reach as with their valid arm in Virtual Reality (VR). Yet, this control relies on the position and orientation of the object to reach expressed in a shoulder-centered reference frame, whereas it might only be available in a head-centered reference frame through gaze-guided computer vision. Here, we designed two methods to perform the required transformation from incomplete, orientation-only data from head and shoulder, possibly available in real life with Inertial Measurement Units. The first involved training offline an ANN using a database of natural movements by multiple participants to perform this transformation, while the second was based on a bioinspired space map with online adaptation. Experimental results on twelve intact-limbs participants controlling a prosthesis avatar in VR demonstrated persistent errors with the first method, which could be quickly absorbed with the second method. This second bioinspired and adaptive method effectively spatially encoded the transition from the head to the shoulder associated with different targets in space. It shows promise for managing complex scenarios involving simultaneous

errors and corrections in both position and orientation, such as camera movement or operation in perturbed environments.

9
Functional gradients of mental imagery and perception in human orbitofrontal and occipitotemporal cortex

Jianghao Liu[1,2], Minye Zhan[3], Laurent Cohen[1], Paolo Bartolomeo[1]

[1]ICM - Institut du Cerveau, Hôpital Pitié Salpétrière, Inserm U1127 - CNRS UMR7225 - Sorbonne Université UMR S 1127, Paris, France
[2]Dassault Systèmes, Vélizy-Villacoublay, France
[3]NeuroSpin, Gif-sur-Yvette, France

How do distinct domain-preferring cortical regions contribute to the individual subjective experience of visual mental imagery and visual perception? We systematically examined the role of face- and color-preferring cortical patches using millimeter-scale 7T fMRI in typical imagers and individuals with aphantasia who report no voluntary imagery experience. We identified bilateral face- and color-preferring patches, adjacent to each other and distributed along a posterior-anterior axis from occipitotemporal to orbitofrontal cortex, where the face patches were always located lateral to color patches. From posterior to anterior patches, imagery and perception exhibited similar increasing gradients of domain-selectivity and representation which correlated with subjective ratings of face identities, face shapes, and visual colors, but opposite gradients of activation amplitude and functional connectivity, with orbitofrontal patches leading top-down processes. Importantly, the orbitofrontal patches were activated before the occipito–temporal patches in both perception and imagery. Aphantasic individuals showed decreased imagery representations in orbitofrontal cortex and orbitofrontal-temporal connectivity in both modalities. In addition, the orbitofrontal patch in aphantasia activated earlier in perception but decreased earlier during the imagery maintenance period. Thus, perceptual and imagery processes share similar representations in domain-preferring cortical patches, but exhibit opposite posterior-to-anterior gradients of activity. Altered activity in the orbitofrontal cortex contributes to diminished conscious imagery experience in aphantasia.

10
A robotic model of stress-altered exploration to study curiosity-familiarity unfolding

Elisa Massi[1], Lola Cañamero[1]

[1]CY Cergy-Paris Université, ETIS lab, Neurocybernetics team, Pontoise, France

Exploration is essential in helping animals and humans to survive, allowing them to gather resources, food, and improve their life condition. When exploring, maintaining a balance between curiosity, the willingness to survey new areas, and familiarity, the need to stay in known, safe zones, is crucial.

Many studies have investigated the dynamics of exploration in novel environments, but how much of this balance can be influenced by the emotional state of the animal is still a question to debate. One of the major forces that impact the emotional state is stress, induced by the

interaction between the animal, other animals and the environment, both perceived through the animal's senses. We propose a robotic framework to study global exploration dynamics under stress condition in which the robot's default exploratory behaviour is altered by its emotional state. In this experiment, robots explore the environment by doing circular motions with obstacle avoidance. Stress factors include the invasion of the robot's peripersonal space by its peers and by obstacles, perceived via visual and touch sensors. In the proposed model, this affects the emotional state by changing the levels of cortisol and oxytocin, which in turn alter the behaviour in terms of speed, rotation angle and size of the peripersonal space. The proposed framework is not explicitly designed to model concepts of curiosity and familiarity, our hypothesis is that these emerge nonetheless from the influence of the emotional state in the behaviour of the robot.

11
Information theoretic study of the neural geometry induced by category learning

Jean-Pierre Nadal[1,2], Laurent Bonnasse-Gahot[2]

[1] Laboratoire de Physique de l'ENS (LPENS), CNRS UMR 8023, Ecole Normale Supérieure, Paris, France
[2] Centre d'Analyse et de Mathématique Sociales (CAMS), CNRS UMR 8557, EHESS, Paris, France

Categorization is an important topic both for biological and artificial neural networks. Here, we take an information theoretic approach to assess the efficiency of the representations induced by category learning. We show that one can decompose the relevant Bayesian cost into two components, one for the coding part and one for the decoding part. Minimizing the coding cost implies maximizing the mutual information between the set of categories and the neural activities. We analytically show that this mutual information can be written as the sum of two terms that can be interpreted as (i) finding an appropriate representation space, and, (ii) building a representation with the appropriate metrics, based on the neural Fisher information on this space. One main consequence is that category learning induces an expansion of neural space near decision boundaries. Finally, we provide numerical illustrations that show how Fisher information of the coding neural population aligns with the boundaries between categories.

12
Does exposure to emotional stimuli induce lasting functional brain changes? A machine-learning predictive approach.

Ernesto Sanz-Arigita[1], Gwladys Lere[1], Yannick Daviaux[2], Pierre Philip[2], Gwenaelle Catheline[1], Ellemarije Altena[1]

[1]SWAN team, INCIA - Université de Bordeaux, CNRS UMR 5287, Bordeaux, France
[2]UMR 6033 SANPSY CHU Pellegrin, Bordeaux, France

BACKGROUND: Based on prior work showing insomnia-related increased brain functional connectivity to emotional stimuli (Sanz-Arigita et al., 2021), the present study addresses whether transitory changes induced during exposure to positive emotional stimuli do translate to lasting changes in the brain functional connectome after exposure. ML-models trained on brain connectivity patterns before and during the emotional stimuli were used to predict post-stimulation connectivity patterns for the insomnia and control groups.

METHODS: We designed a functional magnetic resonance imaging (fMRI) single-session protocol comprising 3 sequential brain activity acquisitions: a basal measure (resting state condition RS#1), followed up by the emotional stimuli (FUN) (standardized, short humorous films), and a final post-exposure measure (RS#2). The protocol was applied to patients with insomnia and age-matched controls (n = 20 / 20). Functional connectivity matrices corresponding to the three fMRI acquisitions were computed. Machine-learning algorithms were trained on brain connectivity patterns a) prior to the stimulation (RS#1), and b) during the exposure to emotional stimuli (FUN) to predict post-stimulation brain connectivity patterns of insomnia patients and controls (RS#2).

CONCLUSION: Machine learning algorithms were able to predict brain connectivity patterns after exposure to humorous stimuli, based on connectivity patterns generated either before or during exposure. Logistic regression outperformed the rest of ML-models. This result supports our hypothesis that connectivity changes induced by exposure to emotional stimuli further generate lasting post-exposure brain changes which differ between insomniacs and control subjects.

REFERENCES
Sanz-Arigita et al., 2021. Brain reactivity to humorous films is affected by insomnia. Sleep 44, zsab081.
https://doi.org/10.1093/sleep/zsab081

13
Conceptual generalization of learned pain modulation is mediated by semantic networks

Dylan SUTTERLIN[1,2], Tor WAGER[3], Leonie KOBAN[4]

[1]CNRS - UMR 5292 - CRNL, Bron, France
[2]University of Montreal, Montreal QC, Canada
[3]Department of Psychological and Brain Sciences, Dartmouth College, Hanover, NH, USA
[4]Lyon Neuroscience Research Center (CRNL), CNRS, INSERM, Université Claude Bernard Lyon 1, Bron, France

The ability to learn associations in our environment and apply them to new situations (generalize) is a fundamental aspect of human cognition. Generalization often relies on perceptual features but can also involve conceptual relationships between learned and novel cues. In the context of pain, previous studies have shown that learned associations, such as repeated pairings between visual stimuli and varying levels of painful stimulation influence pain expectations and experience. However, the brain mechanisms underlying the conceptual generalization of learned pain associations remain unidentified. This project assesses the brain mechanisms underlying the generalization of learned expectations on pain reports. We hypothesize that brain regions involved in conceptual and semantic processing and in affective learning should mediate this effect. During functional magnetic resonance imaging (fMRI), 36 healthy adults first performed a pain learning task, where one vehicle and one animal (counterbalanced across participants) were associated with high or low pain stimulations, respectively. Subsequently, in the generalization task, novel but conceptually related images

(drawing, words or new pictures of other vehicles and animals) were presented before receiving medium-intensity painful stimulation and participants rated their pain in each trial. Multi-level whole-brain mediation analysis revealed that brain regions implicated in semantic processing (hippocampus, temporal pole, posterior cingulate cortex), value encoding (orbitofrontal cortex) and threat (amygdala) mediated the effects of novel generalization cues on pain ratings. These findings suggest that brain regions involved in affective learning and conceptual processing interact to shape pain experience, with implications for understanding the contextual modulation of pain in clinical settings.

14
Proposal of a "neuroethics by design" for the convergence of neuroscience, computer science and engineering through the study of neurotechnologies

Laure Tabouy[1,2]

[1]University of Paris-Saclay, CESP INSERM U1018, Team ethics and epistemology, Villejuif, France
[2]Agence de la biomédecine, 1 Av. du Stade de France, Saint-Denis, France

Neurotechnologies are devices used to study the structure and function of the brain. The convergence of neuroscience, computer science and engineering, and their growing sophistication on the global market, are making them more miniaturized, more efficient and more powerful. This accentuates the porous boundaries between medical and non-medical, civil and military uses, as well as the diversity of objectives, uses and investments. They bring hope, but their capacity to influence or manipulate is worrying. Considering ethics as constructive criticism aimed at protecting our future from our present actions, and based on an interdisciplinary literature review, in neuroscience, ethics, law, digital technology, science fiction and philosophy, my PhD project is to rethink neuroethics in light of the convergence of neuroscience, computer science and engineering, developing a field-based "neuroethics by design", calibrated and applied to these neurotechnologies. In order to provide the best possible response, I am carrying out a reflexive study, based on an interdisciplinary literature review, which I am supplementing with investigative work in the field by carrying out qualitative studies with researchers and entrepreneurs, a flow chart and risk mapping in collaboration with a company and a law firm. The aim was for this by-design proposal to be an accompaniment encompassing all existing recom–mendations and codes, and laws on neurotechnologies and AI/digital, in order to articulate them together. Based on this reflective and field work, prospective scenarios will be sketched out for a neuroethics by design for responsible neurotechnologies that are strategically, scientifically and operationally relevant.

15
Neuro3D — Neural networks for the representation of 3D environments

Paul Uteza[1,2], Hervé Rouault[2,3]

[1] INMED, Inserm U1249, Faculté des sciences de Luminy, Marseille, France
[2] CENTURI, Turing Centre for Living Systems, Marseille, France
[3] Centre de Physique Théorique - UMR 7332
Aix-Marseille Université, Marseille, France

The mammalian visual system is able to perceive three-dimensional objects from their two-dimensional projections on the retina. This process, involving deep areas of the visual system, is complex and not fully understood. However, psychophysical experiments such as a seminal study by Shepard and Metzler have provided insightful characterizations of this process. This experiment measured the reaction times of subjects asked to determine the relative orientation of 3D objects presented on a screen.

In this study, we investigated how recurrent neural networks (RNNs) can be used to model mental rotation. We trained RNNs to generate images of objects with specified rotations. Our results showed that RNNs can accurately model the human ability to mentally rotate objects. Additionally, we analyzed the neural dynamics of the RNNs during the rotation process. We found that the number of computation steps required to rotate an object increases as the rotation angle increases. Furthermore, we observe the emergence of intermediary representations encoding linearly increasing rotation angles. These results account for the main observations of the Shepard-Mezler experiment.

Our findings suggest that RNNs can be used to model the neural dynamics of mental rotation. This could lead to new insights into the cognitive processes underlying this ability. Additionally, our work could be used to develop new artificial intelligence (AI) systems that can perform mental rotation tasks.

16
A Statistical Physics Approach of Exploitation-Exploration Dilemma in Human Adaptive Behavior.

Constantin Vaillant Tenzer[1,2], Etienne Koechlin[1,2,3]

[1]Ecole Normale Supérieure – DEC LNC2, Inserm U960, Paris, France
[2]Université Paris Cité, Paris, France
[3]Sorbonne Université, Paris, France

Humans exhibit impressive adaptive behavior in diverse environments.

However, a unifying theory that accounts for the fundamental principles underlying human adaptive behavior is lacking.

We propose a statistical physics theory describing adaptive behavior based on the following postulate: actions are chosen to maximize knowledge acquisition within internal world models of the environment under energetic resource constraints.

We derived a unique choice distribution this contraint optimization problem. Since some parameters have no analytical solutions, we computed simple and very close approximations.

We applied this model to multi-armed bandits and studied the dynamics of behavior over time.

We are launching an online behavioral experiment - a multi-armed bandit task - to test several predictions of our model.

17
Fatigue behavior as a constrained resource allocation policy

Morgan Verdeil[1], Jean Daunizeau[1]

[1]ICM - Institut du Cerveau, Hôpital Pitié Salpétrière, Inserm U1127 - CNRS UMR7225 - Sorbonne Université UMR S 1127, Paris, France

Mental fatigue alters behavior in diverse and sometimes counter-intuitive ways. Our working hypothesis is that this results from neurobiological constraints imposed on the brain's resource allocation policy.

We start with the premise that allocating (e.g., neuroenergetic) resources to any mental task increases cognitive performance, which is instrumental for obtaining reward. We then consider two possible kinds of neurobiological constraints. On the one hand, the brain's reservoir of resources may be bounded, eventually thwarting cognitive performance if the limit is reached. On the other hand, investing resources may produce neurometabolic byproducts, whose accumulation may eventually hinder neurocognitive efficiency. Under those constraints, how should the brain allocate its resources?

We use optimal control theory to derive the optimal resource allocation policy under each type of constraint. This provides a computational definition of fatigue, in terms of proximity with the (asymptotic) state of maximal constraint impact. Second, we show that these constraints induce a resource allocation cost which scales with expected future foregone reward. Third, we show that this cost increases with fatigue. This enables us to reproduce most established empirical behavioral/cognitive results on healthy mental fatigue.

Crucially, optimal resource allocation policies rely on accurate representations of the brain's mental efficacy, of its fatigue dynamics and of future task contingencies. We show how biased representations (e.g., underconfidence or optimistic reward expectations) yield suboptimal patterns of resource allocation, which are reminiscent of pathological fatigue. Future work will build on this framework to assess suboptimal neurocognitive resource management in persons at risk for burnout syndrome.