Università degli Studi di Ferrara

Doctoral course in

MATHEMATICS

Cycle XXXVIII

Coordinator: prof. Lorenzo Brasco

Mathematical models of cognition: from active inference to clinical applications

Scientific Disciplinary Sector: MATH-05/A

Candidate

Dott. Riccardo Maramotti

Supervisors

Prof. Silvia Bonettini
Prof. Giovanna Zamboni
Prof. Giuseppe Pagnoni

Years 2022-2025

Abstract

This thesis presents the application of mathematical methods to the study of cognition and neurodegeneration. The work is organized in three main parts, spanning theoretical foundations, behavioral experiments, and applications to clinical data.

The first part introduces the theory of active inference. Within this framework, the brain is seen as a Bayesian machine that employs probabilistic inference not only for perception, but also to guide action. Action selection is formalized as the minimization of expected free energy, representing a trade-off between seeking out preferred sensations and information gain. Partially Observable Markov Decision Processes (POMDPs) are then presented as formal models of discrete-time active inference.

The second part adopts the active inference approach to model the behavior of 20 healthy adults performing a color–word Stroop task under two levels of intentional cognitive effort. A hierarchical POMDP was inverted on both choice and reaction time data, and competing model structures were compared using Parametric Empirical Bayes. The results suggest that voluntary modulation of effort can be explained by changes in motivation.

The third part includes three mathematical approaches applied to clinical observations. First, the age-specific prevalence of Alzheimer's disease and frontotemporal dementia was estimated using generalized logistic functions. Second, an age-adjusted diagnostic cut-off for neurofilament light chain (NfL) was developed using a two-threshold Support Vector Machine classifier, combined with Bayesian regression to account for confounding variables. Third, the agreement between two laboratory platforms for the measurement of NfL concentrations was evaluated through regression analyses.

Overall, the thesis illustrates how mathematical modeling can be applied to different areas of neuroscience, from abstract theories of cognition to practical challenges related to clinical biomarkers.

Contents

Introduction

Understanding how the human mind functions has always been challenging. However, thanks to advances in computational modeling and neuroimaging, we can now describe cognitive and neural processes using predictive and explanatory mathematical structures. This thesis harbors from that tradition, by demonstrating how mathematics can serve not merely as a tool for data analysis, but also as a framework for studying cognition, behavior, and their clinical alterations.

It is only in recent decades that mathematics has become a central tool for understanding the brain. The first decisive step came from biophysics: in the 1950s, Hodgkin and Huxley described the mechanisms of neuronal excitability through a system of differential equations. However, the true revolution occurred in the 1990s with the advent of functional magnetic resonance imaging (fMRI). This technology generated unprecedented volumes of data, forcing neuroscience to adopt computational methods capable of extracting meaningful information from complex signals. Karl J. Friston was a key figure in this revolution. He was the first to formalize the use of general linear models for fMRI analysis and developed the SPM software, which remains one of the most widely used tools for fMRI data today.

In the 2000s and 2010s, the focus gradually expanded from the brain's physiology to its way of processing information. Bayesian inference, reinforcement learning, and predictive coding began to formalize perception, learning, and decision-making as algorithmic processes. Behavior was no longer just measured; it was also explained as the result of internal computations. Within this context, the active inference framework emerged as a synthesis of these perspectives.

This thesis positions itself as a link between mathematics and empirical neuroscience. It is structured in three main parts: first, a theoretical introduction to active inference; second, an experimental application of this framework to a behavioral study on voluntary effort; and finally, a set of clinical studies extending the inferential logic to biomedical data.

Human cognition operates in a world of uncertainty. Every act of perception, memory, or decision involves hidden causes that must be inferred from noisy sensory data. The active inference framework, presented in Chapter 1, formalizes this idea. It assumes that the brain continuously updates an internal generative model to minimize variational free energy, which quantifies how well sensory inputs are explained by the brain's predictions. By minimizing this free energy, the brain improves its internal model while acting to make its predictions come true. Therefore, perception and action emerge as complementary elements of the same process.

Among the implementations of active inference, the discrete-time Partially Observable Markov Decision Process (POMDP) formulation has proven to be particularly suitable

for modeling behavior. Here, cognition is expressed through a set of probabilistic matrices and vectors representing the likelihood of observations, the state transitions, the prior preferences, the initial beliefs, and the habitual tendencies. POMDP models are computationally tractable, but their relevance to neuroscience depends on their biological interpretability. In the POMDP formulation of active inference, it is hypothesized that confidence about a potential course of action (*policy*) is associated with neuromodulatory, mainly dopaminergic, transmission in basal ganglia-cortical loops, while prior preferences may reflect processes in the limbic system. This interpretability supports testable hypotheses about brain function, rather than simply fitting behavioral data.

The discrete POMDP framework provides the methodological foundation for the experimental work described in Chapter 2, where it is used to model how people adjust their cognitive effort when faced with conflicting demands. This phenomenon was studied using a variant of the Stroop task, a classical neuropsychological paradigm in which participants are shown words that are names of colors (e.g., 'red'), and are required to indicate the color in which each word is written. To name the correct color, participants must inhibit their automatic tendency to read the word. This conflict between habitual and goal-directed responses offers a test of the brain's capacity to modulate internal precision (i.e., confidence) over the available policies.

Within the active inference framework, effort can be understood as the computational cost of redistributing the probability over the *policy space* to favor accuracy over automaticity. To capture this dynamic, a hierarchical POMDP model was employed: a higher level encodes contextual information, providing priors for a lower level that represents single-trial decisions. This hierarchical structure mirrors the functional architecture of the brain, where slower contextual representations constrain faster perceptual and motor processes. With a suitable POMDP model – and under the hypotheses upon which the model is built – it is possible to infer the latent cognitive parameters that best explain participants' behavior. Indeed, using variational Bayesian inversion, these parameters can be estimated in our experimental task from choices and response times, showing that active inference models can capture behavior in a meaningful and interpretable way.

Finally, Chapter 3 covers applications in clinical research. While cognitive experiments offer controlled environments to test theoretical models, clinical datasets present distinctive challenges. They are often small, heterogeneous, constrained by ethical considerations, and difficult to collect. Comorbidities and missing information further complicate analyses. In such contexts, deterministic or purely data-driven approaches are often insufficient. Probabilistic modeling provides a valid alternative: it allows uncertainty to be expressed rather than ignored, incorporates prior knowledge about physiological limits, and regularizes inference to avoid overfitting. In this way, modeling becomes safer, ensuring that conclusions remain interpretable and robust.

Three studies are presented in Chapter 3. First, generalized logistic functions are used to analyze the age-specific prevalence of Alzheimer's disease and frontotemporal dementia. Second, a two-threshold Support Vector Machine classifier is used to develop an age-adjusted diagnostic cutoff for neurofilament light chain concentrations. Third, a non-parametric Passing-Bablok regression is used to evaluate the agreement between two laboratory platforms for biomarker concentrations. Although the domains differ, the guiding principle remains the same: applying advanced mathematical methods to clinical research.

In sum, this thesis treats cognition and clinical measurement as problems of probabilistic inference under uncertainty. The chapters that follow move from foundations to applications: beliefs and precisions in Chapter 1 become testable hypotheses in Chapter 2, and finally are translated into clinical estimates in Chapter 3. The goal is to provide a transparent and reproducible process for moving from assumptions to inferences.

Chapter 1

The Bayesian brain in action: a guide to active inference

The active inference framework has become a powerful and increasingly popular approach for modeling neurocognitive processes, especially for a class of models known as partially-observable Markov decision processes (POMDPs) [1]. This formulation offers a flexible modeling tool that makes it possible for researchers to simulate different kinds of cognitive processes and also predict related neuronal responses using the theory of associated neural processes [2].

An important characteristic of active inference is that it views decision-making as unfolding from the same inferential mechanisms as perception. In contrast to passive perceptual inference, however, decision-making involves choosing actions that are expected to produce preferred sensory outcomes. For instance, an agent inside a dark room can infer that moving closer to a light source will make it easier to see. At the same time, agents also infer actions that reduce uncertainty and promote learning, such as checking different routes on a map to discover the quickest path to a destination. In this way, behavior results from the balance between reward maximization and information gain, a trade-off that has been shown to be consistent with empirical observations about perception and decision making [3].

In this chapter, we will focus specifically on the discrete formulation of active inference, which treats time as a series of discrete steps and represents beliefs and actions as vectors and matrices. However, it is important to note that this scope deliberately excludes other formulations, such as continuous models or hybrid models that combine discrete and continuous states.

1.1 The Bayesian brain

The *Bayesian brain* hypothesis states that the brain performs probabilistic inference to interpret sensory input and guide behavior [4, 5]. According to this view, perception is an inferential process that combines (top-down) prior information about the most likely causes of sensations with (bottom-up) sensory stimuli [6]. This corresponds to the process of estimating the most probable causes of sensory data, integrating prior beliefs with new evidence.

The main assumption is that the brain is equipped with an internal model of how sensory data are generated – a *generative model*. This model specifies a probabilistic mapping from hidden causes (i.e., environmental states that cannot be observed directly)

to observed data o. Therefore, the agent uses sensory data to infer the most plausible hidden states s that could have generated them. This is formalized by Bayes' theorem:

$$p(s|o) = \frac{p(o|s)p(s)}{p(o)} \tag{1.1}$$

Here, the term $p(s)$ is the *prior*, which represents the agent's beliefs about the state of the world before receiving new data. The term $p(o|s)$ is the *likelihood*, reflecting how probable the observation o would be if the hidden cause was s. The resulting *posterior* $p(s|o)$ is the updated belief about s after observing o. Finally, $p(o)$ is the *model evidence*, which serves as a normalizing constant ensuring that the posterior is a proper probability distribution.

Once all possible states and observations are defined, the generative model is expressed as a joint probability distribution $p(s, o)$, which represents the likelihood of all combination of states and observations. By applying the product rule of probability, this joint distribution can be factorized as:

$$p(s, o) = p(s|o)p(o) = p(o|s)p(s) \tag{1.2}$$

The case of a single set of states and observations results in a two-dimensional distribution, while more complex models lead to higher-dimensional distributions. Although these are harder to visualize and computationally more demanding, they can be conceptually treated in the same way.

1.1.1 Surprise and Bayesian surprise

A computational implementation of the Bayesian brain is the *predictive coding* theory [7], which suggests that the brain continuously generates predictions about incoming sensory input and updates its beliefs by computing prediction errors. These errors are propagated across hierarchical cortical areas to adjust predictions and improve future inference. The process is driven by the minimization of *surprise*, mathematically defined as:

$$-\log p(o) \tag{1.3}$$

This quantity reflects how unlikely a sensory observation o is under the current model of the world (and is therefore sometimes more explicitly written as $-\log p(o|m)$, where m denotes the generative model). Therefore, the notion of evidence $p(o)$ becomes especially meaningful when framed in terms of surprise, as it quantifies how valid the model is as a predictor of observations: the lower the evidence, the less adequate the model, and the greater the resulting surprise. An unexpected stimulus – such as hearing a loud bang in a quiet room – generates high surprise, since $p(o)$ is low. Thus, surprise quantifies how 'wrong' the system was in its prediction, regardless of whether it learns from the event.

However, minimizing surprise only would lead to pathological behavior (e.g., an agent might remain in a dark room to avoid unexpected inputs). To behave adaptively, the agent must also seek information-rich stimuli that improve its internal model. This leads to the concept of *Bayesian surprise* [6], defined as the Kullback-Leibler (KL) divergence between posterior and prior:

$$D_{KL}[p(s|o)\|p(s)] \tag{1.4}$$

where the KL divergence is a (non-symmetric) measure of dissimilarity between two distributions defined as the expected difference of the log of the distributions, with the expectation taken with respect to the first distribution:

$$D_{KL}[q(\theta)\|p(\theta)] \coloneqq \mathbb{E}_{q(\theta)}\left[\log \frac{q(\theta)}{p(\theta)}\right] = \sum_{\theta} q(\theta) \log \frac{q(\theta)}{p(\theta)} \tag{1.5}$$

Bayesian surprise quantifies the amount of belief updating. The difference between surprise and Bayesian surprise can be summarized with the following example. Seeing a cat coming out of the bush in front of the park bench where we are sitting may be surprising, but not necessarily informative (we know cats exist and can be in a park). In contrast, seeing a talking cat would induce high Bayesian surprise, as it forces a drastic revision of prior beliefs about what a cat can do. These two notions of surprise play a central role in the Free Energy Principle, which formalizes how agents balance prediction, belief updating, and action. We will explore this framework in Section 1.3.

1.2 Active Inference: the fusion of action and perception

The predictive coding framework models perception as a process of minimizing prediction errors by updating internal beliefs. Active inference generalizes this principle by recognizing that agents do not only update their beliefs passively, as they also act to fulfill their predictions and preferences [6]. Both processes rely on the same generative model, which not only predicts incoming sensory inputs based on hidden states, but also includes expectations about the consequences of possible actions. In this way, the agent can simulate both how the world generates sensory data and how its own actions can shape future observations.

Perception and action are thereby unified under a common principle: minimization of surprise. Perception does it by updating beliefs so that they become more consistent with incoming data; action, by changing the external world so that future sensory inputs conform more closely to prior expectations. This dual route to surprise minimization solves the so-called *'dark-room problem'*: an agent that only suppresses sensory surprise would indeed prefer a static environment with fully predictable sensory signals; however, a proper active agent will behave so as to increase the probability of attaining its preferred sensations (according to its own phenotype[1]) and of gathering the sensory data it needs to decrease its uncertainty about the world [8, 9].

1.2.1 Generative model and generative process

An important point to consider now is the difference between *generative model* and *generative process* (see Figure 1.1). A generative model, as discussed above, is constituted by beliefs about the world, and these beliefs may be inaccurate. In other words, the explanations about how observations are produced do not need to be an exact description of how they are actually generated. On the other hand, the generative process refers to

[1]For the human phenotype, an extended period of sensory deprivation – such as that engendered by a dark room – is hardly a preferred condition.

what is actually happening in the real world. It describes the 'ground truth' behind the real causes of sensory input [3]. Figure 1.1 represents this canonical architecture. The external process occupies a hidden state denoted by s^*, which is not directly observable by the agent but gives rise to sensory observations o. The agent uses these observations to build or refine its representation s of the external causes of the sensory data, by updating the parameters of the generative model. This distinction is crucial: while s^* is related to the generative process, s belongs to the internal generative model of the agent.

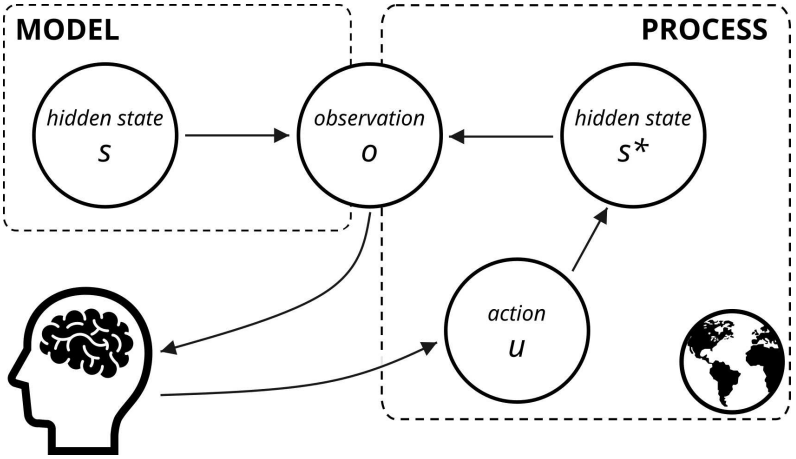

Figure 1.1. Generative process and generative model. Adapted from Parr, Pezzulo, and Friston [6, Figure 2.2].

At each point in time, the agent infers probabilistically the hidden causes of its sensory input, and selects an action u intended to reduce the expected surprise of future observations. This action (on the world) alters the external process, leading to a new observation, and the cycle continues. Mathematically, the integration of action and perception is achieved by including *policies* (π) in the generative model, which can be written as:

$$p(s, o, \pi) = p(o|s, \pi)p(s|\pi)p(\pi) = p(o|s)p(s|\pi)p(\pi) \tag{1.6}$$

where we have dropped the dependency on policies in the likelihood term, under the assumption that likelihood mapping does not depend on policies.

Each policy corresponds to a possible sequence of actions that the agent could perform. To select among them, the agent needs a mechanism to assign a higher value to some policies compared to others. In the active inference scheme, as we will see later on, these values are inferred on the basis of the likelihood of a policy to secure the preferred sensations and harvest useful information; they are then converted into a probability distribution over policies, which is finally sampled by the agent to decide its imminent course of action. This connects the agent back to the generative process, since the selected policy changes the true state of the world through action.

In this framework, *learning* corresponds to a slower optimization that adjusts the parameters of the model itself so that its long-term predictions track more faithfully the statistical structure of the environment [6]. Neurobiologically, this has been linked to experience-dependent plasticity and synaptic consolidation, furnishing an elegant hierarchy of nested timescales: rapid inference (perception), rapid policy selection (action selection), and slow structural adaptation (learning).

1.3 The Free Energy Principle

As discussed in the previous sections, perception can be understood as the inversion of a generative model. Formally, this is achieved by computing the posterior distribution $p(s|o,\pi)$ using Bayes' theorem. However, in most realistic settings, especially when dealing with high-dimensional or dynamic models, the posterior cannot be computed analytically. This is because the model evidence $p(o|\pi)$, which serves as a normalizing constant, involves an intractable sum over all possible hidden causes:

$$p(o|\pi) = \sum_s p(o|s)p(s|\pi) \tag{1.7}$$

To solve this computational problem, we can approximate the true posterior with a more tractable distribution $q(s|\pi)$, known as the *proposal*, and the surprise with an approximation called *variational free energy* F_π defined as follows:

$$F_\pi \coloneqq \mathbb{E}_{q(s|\pi)}\left[\log \frac{q(s|\pi)}{p(s,o|\pi)}\right] \tag{1.8}$$

The variational free energy F_π is computed separately for each possible policy. This is necessary because different policies influence the hidden states of the generative process, which in turn affect the likelihood of observing particular outcomes. For example, if I believe that there is a chair on my left and a table on my right, then, conditional on choosing the policy of looking left, observing a chair becomes more probable than observing a table[2] [3]. This is also the reason why both the approximate posterior $q(s|\pi)$ and the generative model $p(s,o|\pi)$ are conditioned on policies.

Now, we can prove that F_π is an upper bound for the surprise, and also that their difference is the KL divergence between the proposal and the true posterior (always non-negative, as KL divergences are). Indeed:

$$\begin{aligned}
F_\pi &= \mathbb{E}_{q(s|\pi)}\left[\log \frac{q(s|\pi)}{p(s,o|\pi)}\right] = \\
&= \mathbb{E}_{q(s|\pi)}\left[\log \frac{q(s|\pi)}{p(s|o,\pi)p(o|\pi)}\right] = \\
&= \mathbb{E}_{q(s|\pi)}\left[\log \frac{q(s|\pi)}{p(s|o,\pi)} - \log p(o|\pi)\right] = \\
&= \mathbb{E}_{q(s|\pi)}\left[\log \frac{q(s|\pi)}{p(s|o,\pi)}\right] - \mathbb{E}_{q(s|\pi)}[\log p(o|\pi)] = \\
&= D_{KL}[q(s|\pi)||p(s|o,\pi)] - \log p(o|\pi)
\end{aligned} \tag{1.9}$$

Since the surprise $-\log p(o|\pi)$ is independent of the choice of proposal, the value of F_π will become smaller as the value of $q(s|\pi)$ approaches the value of the true posterior $p(s|o,\pi)$. In other words, minimizing the KL divergence is equivalent to minimizing the variational free energy F_π. This leads to the formulation of the *Free Energy Principle* [10]: self-organizing systems, such as living organisms, can be described as minimizing the probability of finding themselves in surprising states; they do so by minimizing a tractable upper bound to surprise called variational free energy.

[2] At the same time, observing the chair also provides evidence for having chosen the policy to look left.

1.3.1 The variational free energy as an accuracy–complexity trade-off

As discussed earlier, the goal of both perception and learning is to minimize variational free energy to obtain an approximation of optimal posterior beliefs after each new observation. However, it is important to emphasize that minimizing F_π does not simply mean finding the best-fitting approximation. Sensory input is usually noisy, and if the agent always tried to fit each single trial perfectly, it would capture the noise instead of the real causes. This would lead to unnecessary and metabolically expensive updates. In mathematics, this phenomenon is known as overfitting. Fortunately, the way F_π minimization works naturally prevents this problem. Indeed, we could decompose variational free energy according to the following passages:

$$\begin{aligned} F_\pi &= \mathbb{E}_{q(s|\pi)}\left[\log \frac{q(s|\pi)}{p(s,o|\pi)}\right] = \\ &= \mathbb{E}_{q(s|\pi)}\left[\log \frac{q(s|\pi)}{p(s|\pi)p(o|s)}\right] = \\ &= \mathbb{E}_{q(s|\pi)}\left[\log \frac{q(s|\pi)}{p(s|\pi)}\right] - \mathbb{E}_{q(s|\pi)}[\log p(o|s)] = \\ &= D_{KL}[q(s|\pi)||p(s|\pi)] - \mathbb{E}_{q(s|\pi)}[\log p(o|s)] \end{aligned} \tag{1.10}$$

The first term represents the *complexity* of the model (i.e, how much the beliefs need to change to keep a high level of accuracy when new observations are received). The second term (without the negative sign) quantifies the *accuracy* of the model (i.e., how well it explains the data). Therefore, perception seeks to find the most parsimonious changes in beliefs about the causes of sensory input that are still sufficient to explain that input correctly (as in Occam's razor). Minimizing the variational free energy implies attempting to minimize complexity while trying, *at the same time*, to maximize accuracy, and can thus be interpreted as an accuracy–complexity trade-off.

Equation 1.10 can also be interpreted in the language of inverse problems. These problems are often *ill-posed*, and focusing solely on data fit may lead to overfitting random fluctuations rather than capturing the true underlying causes. The free-energy decomposition makes explicit how active inference avoids this issue: the KL term acts as a penalty on unwarranted deviations from the prior structure, thereby stabilizing the inversion of the generative model. From this perspective, priors are *regularisers*: their form and precision determine the geometry of the penalty that keeps inference well-posed. In particular, Gaussian priors implement quadratic (Tikhonov/$L2$) penalization, whereas Laplace priors correspond to $L1$ penalties that promote sparsity. The choice of a prior family thus specifies the regularization strategy [11].

1.3.2 The Free Energy Principle for active inference

As we mentioned earlier, active inference is not solely concerned with minimizing *current* prediction error in perception: action selection aims at choosing policies expected to minimize *future* variational free energy. Indeed, in addition to variational free energy F_π (which focuses on minimizing prediction error and explaining current observations), we need to define *expected variational free energy* G_π, which evaluates future outcomes that

have not yet been observed. The task of action selection is to choose policies π that minimize expected free energy [3].

Formally, the expected free energy for a given policy is defined as:

$$\begin{aligned} G_\pi &:= \mathbb{E}_{q(s,o|\pi)}\left[\log \frac{q(s|\pi)}{p(s,o|\pi)}\right] \\ &= \mathbb{E}_{q(s,o|\pi)}\left[\log \frac{q(s|\pi)}{p(s|o,\pi)p(o|\pi)}\right] \\ &= \mathbb{E}_{q(s,o|\pi)}\left[\log \frac{q(s|\pi)}{p(s|o,\pi)}\right] - \mathbb{E}_{q(o|\pi)}\left[\log p(o|\pi)\right] \\ &\approx \mathbb{E}_{q(s,o|\pi)}\left[\log \frac{q(s|\pi)}{q(s|o,\pi)}\right] - \mathbb{E}_{q(o|\pi)}\left[\log p(o|C)\right] \\ &= -\mathbb{E}_{q(s,o|\pi)}\left[\log \frac{q(s|o,\pi)}{q(s|\pi)}\right] - \mathbb{E}_{q(o|\pi)}\left[\log p(o|C)\right] \end{aligned} \tag{1.11}$$

The approximation in the penultimate line involves two main steps: (1) replacing the true posterior $p(s|o,\pi)$ with the approximate posterior $q(s|o,\pi)$, and (2) introducing *prior preferences* $p(o|C)$, which encode the observations that the agent expects or desires.

This decomposition reveals two fundamental components of G_π. Referring to the last line of the Equation 1.11, the first term represents the negative *expected information gain*; this term captures information-seeking behavior, such as exploring uncertain environments before acting to obtain rewards. The second term represents the *expected cost*, which evaluates how well the expected observations under a policy match the agent's preferred outcomes; this term drives reward-seeking behavior, pushing the agent toward policies that maximize the probability of achieving its goals [12]. Thus, the expected free energy associated with a policy is a value that represents a trade-off between the expected cost of that policy and its expected information gain [3]. When the agent's beliefs about (the hidden states of) its world are imprecise, actions are typically more of the information-seeking kind, aiming to reduce uncertainty. Conversely, when the agent is confident in its beliefs, actions become more reward-seeking, focusing on achieving preferred outcomes.

From a numerical point of view, the expected cost in Equation 1.11 penalizes policies expected to generate non-preferred outcomes and therefore acts as the regularizer in G_π. Therefore, the precision of the preferences distribution plays an important role in decision-making. When preferences are encoded with very high precision, G_π minimization can lead to risky behavior, where the agent strongly prioritizes obtaining the preferred sensations even when little information about how to achieve them is available. On the other hand, when the agent does not have strongly preferred observations (the distribution of the preferences is imprecise), it tends to act more to gather new information – out of *curiosity*, we could say. Thus, differences in the precision of prior preferences can help explain individual variability in these behavioral strategies.

Finally, it is important to remark that the posterior over policies depends on both F_π and G_π. While F_π evaluates how well a policy has worked so far, providing a retrospective assessment, G_π estimates how well the same policy is expected to perform in the future, giving a prospective evaluation. In intuitive terms, F_π answers the question "how good has this action plan turned out so far?", while G_π addresses "how good do I expect things to go if I continue to follow this action plan?" [3].

1.4 Partially observable Markov decision processes (POMDPs)

The expression POMDP (Partially Observable Markov Decision Process) refers to two main ideas. The first is *partial observability*, which means that observations give only probabilistic information about hidden states. The second is the *Markov property*, which says that all relevant information about the distant past is already summarized in the present belief about the state. This assumption is not always valid, but makes modeling more feasible and in many cases is sufficient [13, 14]. When the Markov assumption is not valid – for example when memory has to be modeled – it is necessary to include several interconnected Markovian processes that evolve over different timescales. However, in this Section, we will not treat these more complex models.

A POMDP is organized into trials and into time steps (or *epochs*) within each trial. These internal steps are usually denoted by τ. It is important to note that τ refers to the time points for which the agent holds beliefs. This must be distinguished from the variable t – although they are defined on the same temporal grid – which indicates the times when new observations are received. This distinction is necessary because an observation in the present may alter the belief about what has happened in the past. Imagine that you begin in one of two rooms, one green and one blue, without knowing the color of the walls. Later, if you open your eyes and perceive that the room is blue, you could revise your belief about the earlier situation: you infer that you were already in the blue room from the beginning. Formally, this means that the belief about the state at $\tau = 1$ is modified after the observation at $t = 2$. The inclusion of both t and τ in active inference means that the agent continuously updates beliefs about states at all time points whenever a new observation is acquired. In this way the model allows both retrospective inference, as in the room example, and prospective inference, in which expectations about future states are updated [3].

To make the concept easier to understand, it helps to look at some classical examples. A very well-known one is the 'Tiger Problem' [14]. In this situation, an agent stands in front of two doors: behind one there is a tiger, behind the other there is a treasure. The agent may open a door (risking to meet the tiger), or decide to listen. Listening has a cost – the value of the treasure decreases with decision time – but it provides some (uncertain, that is, probabilistic) information about where the tiger is located. This example shows clearly the tension in POMDPs between actions that collect information (*exploration*) and actions that try to maximize immediate reward (*exploitation*). Also in the previous room scenario the agent can choose an exploratory action, like turning on the lights or looking around, which costs something (e.g., in terms of the energy required) but reduces uncertainty about the color of the room. Both examples show the same trade-off between exploration and exploitation [15] and illustrate how POMDPs apply to many real-world problems, where information is incomplete and the agent must balance between exploring and exploiting.

Finally, it is necessary to underline that in POMDPs the agent never acts based on the direct knowledge of the hidden states, since these cannot be accessed by definition. Instead, the agent acts based on its *beliefs* about such states, which mathematically are just probability distribution over possible hidden states. For example, in the Tiger Problem, after listening the agent may believe that the tiger is behind the left door with 70% probability, and that it is behind the right door with 30% probability. Beliefs are updated after each observation according to Bayes' rule. This idea of acting based on

beliefs instead of acting based on the real but hidden states will be important later for understanding the relation between POMDPs and active inference [16].

1.4.1 Formal structure of a POMDP

Having introduced the basic intuition of POMDPs, we can now look more closely at their matrix structure, which is represented in Figure 1.2. In Table 1.1 the reader can find a summary of the symbols used in this Section.

Symbol	Definition	Dimension
s_τ	Hidden state at time τ.	one-hot vector in $\mathbb{R}^{\lvert S\rvert}$
o_τ	Observation at time τ.	one-hot vector in $\mathbb{R}^{\lvert O\rvert}$
π	Vector encoding the distribution over policies (i.e., sequence of actions), reflecting the predicted value of each policy.	$\mathbb{R}^{\lvert\Pi\rvert}$
$\mathbf{A}$	Matrix representing the probability that specific outcomes will be observed given specific hidden states.	$\mathbb{R}^{\lvert O\rvert\times\lvert S\rvert}$
$\mathbf{B}$	Tensor encoding beliefs about how hidden states evolve over time. At each time τ, the policy π prescribes a specific action, which selects the corresponding transition matrix $\mathbf{B}_{\pi,\tau}$	$\mathbb{R}^{\lvert S\rvert\times\lvert S\rvert\times\lvert\Pi\rvert\times T}$
$\mathbf{C}$	Matrix representing the degree to which some observed outcomes are preferred over others, technically modeled as prior expectations over outcomes. $\mathbf{C}$ has a column for each time point.	$\mathbb{R}^{\lvert O\rvert\times T}$
D	Vector encoding prior beliefs about initial hidden states.	$\mathbb{R}^{\lvert S\rvert}$
E	Vector representing the habits, technically modeled as a prior probability distribution over policies.	$\mathbb{R}^{\lvert\Pi\rvert}$
F	Variational free energy under the policies in π.	$\mathbb{R}^{\lvert\Pi\rvert}$
G	Expected free energy under the policies in π.	$\mathbb{R}^{\lvert\Pi\rvert}$
λ	Inverse temperature parameter reflecting the policy precision.	$\mathbb{R}^+$ (positive scalar)

Table 1.1. Summary of symbols used in the discrete POMDP formulation. Dimensions refer to the length of the trial (T), and the cardinalities of the outcome ($|O|$), state ($|S|$), and policy ($|\Pi|$) spaces.

At the first time point in a trial ($\tau = 1$), the model needs a prior distribution over the possible hidden states. This is encoded in a vector, usually called D, which contains one value for each possible state. If there are several factors of states, there will be one D vector for each factor. This allows to represent, for example, independent beliefs about the location of an object and about its sound.

When the process continues, the model needs to know how states evolve from one time to the next, depending on which action is selected under a policy. This is given by transition matrices $\mathbf{B}_{\pi,\tau}$, which specify the probability $p(s_{\tau+1}|s_\tau, \pi)$ of reaching a state at $\tau + 1$ given the state at τ. Each column of $\mathbf{B}_{\pi,\tau}$ corresponds to a state at τ and each row to a state at $\tau + 1$. In simple cases, there is a single $\mathbf{B}_\tau$ matrix for a given factor,

but when transitions are policy-dependent, there will be a $\mathbf{B}_{\pi,\tau}$ matrix for each possible action. Thus, we can define the four-dimensional tensor $\mathbf{B}$ as a concatenation of all $\mathbf{B}_{\pi,\tau}$ for each time τ and each possible action.

The link between hidden states and observations is given by the likelihood tensor $\mathbf{A}$. These arrays specify the probability $p(o_\tau|s_\tau)$ of observing a certain outcome under a given combination of hidden states. The dimension of $\mathbf{A}$ thus exceeds by one the number of hidden factors. When there are multiple sensory modalities (for example, visual and auditory), there is one $\mathbf{A}$ array for each modality.

The preferences about the outcomes, $\log p(o_\tau|C)$, are encoded in the matrix $\mathbf{C}$. This matrix contains one value for each possible observation, usually expressed in logarithmic preference terms. When there are multiple modalities, there is one $\mathbf{C}$ matrix for each of them. The agent tries to select the policies which are more likely to secure the most preferred observations, according to these $\mathbf{C}$ values.

The actions u can therefore be seen as selecting one of the possible $\mathbf{B}_{\pi,\tau}$ matrices for each state factor. The evaluation of policies uses two quantities, the vectors F and G. The vector F encodes the *variational free energy* of a policy given past observations, while the vector G encodes the *expected free energy* of the policy (according to its foreseen future observations). These are combined with a scalar parameter $\lambda \geq 0$, which controls the precision with which the policies are selected. This parameter can be thought of as the agent's prior belief about the confidence with which policies can be chosen on the basis of G. It can also be seen as an inverse temperature: when λ is high, the policy with the lowest G is chosen with high confidence, while when λ is low, the selection is more random.

The vector E, also shown in Figure 1.2, encodes prior *habits* about which policies are usually preferred, independently of their expected free energy. The presence of E is optional, but when included it allows to model habitual influences. In this way, both E and G contribute to the posterior distribution over policies π. Here, it is important to highlight the difference between E and $\mathbf{C}$. The first encodes prior beliefs over policies, with each entry specifying the prior probability assigned to a particular policy. In contrast, the latter encodes prior preferences over outcomes at each time step, specifying which observations are considered more desirable. In short, E represents priors over policies, while $\mathbf{C}$ represents priors over outcomes.

The distribution over policies π depends both on prior beliefs about habits, encoded in the E vector, and on the expected free energy G associated with each policy [2]. The influence of G is scaled by the precision parameter λ. The expected free energy G itself is defined in terms of prior preferences $\mathbf{C}$. The $\mathbf{C}$ matrices encode which observations are more desirable than others, and therefore the policies expected to generate these preferred outcomes will have lower G values. Formally, the prior over policies before any observation is received is written as:

$$\pi_0 = \sigma(\log E - \lambda G) \tag{1.12}$$

where σ denotes the softmax operator[3]. After an observation, the posterior over

[3]For a vector $x \in \mathbb{R}^K$, the softmax function is defined component-wise as:

$$\sigma(x)_i = \frac{\exp(x_i)}{\sum_{j=1}^{K} \exp(x_j)}, \qquad i = 1, \ldots, K.$$

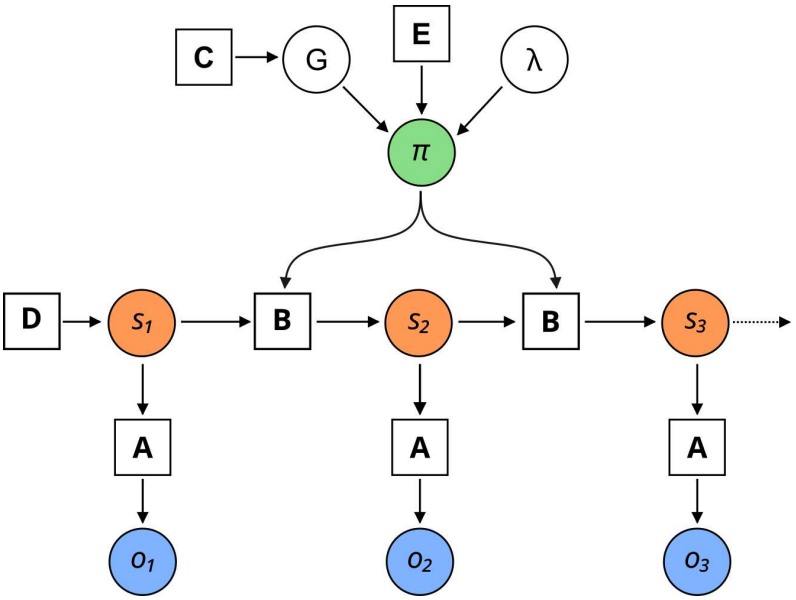

Figure 1.2. Bayesian network representations of perception and policy selection in a POMDP. Hidden states (orange nodes) evolve according to policy-dependent transition matrices **B**, while observations (blue nodes) are generated through the likelihood matrices **A**. Adapted from Smith, Friston, and Whyte [3, Figure 5].

policies incorporates also the variational free energy F, leading to:

$$\pi = \sigma(\log E - F - \lambda G) \tag{1.13}$$

In this way, both habits (E), past evidence (F), and expected future outcomes (G) contribute to policy selection. From the perspective of the generative model, once a policy is selected, it prescribes a sequence of actions; these actions determine which $\mathbf{B}_{\pi,\tau}$ matrices are applied, thereby generating hidden states over time, which in turn generate observations via likelihood mapping **A**. This is the sense in which the diagram in Figure 1.2 represents a generative process: it shows how sequences of states and observations can be generated given policies and preferences.

1.4.2 Message passing in POMDPs

Figure 1.2 shows how the previously introduced matrices and vectors are linked in the generative model using the notation of Bayesian networks. In these diagrams, the circles (or *nodes*) correspond to observations, hidden states, and policies, while the arrows (*edges*) show conditional dependencies between them [11]. Observations at each time step (blue nodes) are generated by hidden states (orange nodes) through a mapping encoded by the likelihood matrices **A**. The $\mathbf{B}_{\pi,\tau}$ matrices specify the dependencies between states in successive time steps, and the initial distribution over states is determined by the D vector.

Let $o_{1:T}$ and $u_{1:T-1}$ denote, respectively, the observed outcome and action sequences in a trial of length T. Generative models are formally defined as the joint probability distribution over observations, states, and policies of the POMDP across time:

$$p(o_{1:T}, s_{1:T}, \pi) \tag{1.14}$$

Using Markov property, this joint distribution can be factorized into a prior distribution over the initial state and the policy, together with likelihood terms and state transition terms:

$$p(o_{1:T}, s_{1:T}, \pi) = p(s_1)p(\pi) \prod_{\tau=1}^{T} p(o_\tau|s_\tau) \prod_{\tau=2}^{T} p(s_\tau|s_{\tau-1}, \pi) \tag{1.15}$$

Conditioning on the policy π gives:

$$p(o_{1:T}, s_{1:T}|\pi) = p(s_1) \prod_{\tau=1}^{T} p(o_\tau|s_\tau) \prod_{\tau=2}^{T} p(s_\tau|s_{\tau-1}, \pi) \tag{1.16}$$

Each distribution above can be expressed in matrix form using one-hot vectors (i.e., vectors of zeros with a one placed in the element corresponding to the state/observation of interest):

- The prior over the first state is $p(s_1) = s_1^\top D$, where s_1 is a one-hot vector selecting the relevant entry of vector D.
- The likelihood of an observation given a state is $p(o_\tau|s_\tau) = o_\tau^\top \mathbf{A}\, s_\tau$, where A is the likelihood matrix, and the bilinear form selects the entry corresponding to the realized state–outcome pair.
- The transition probability is $p(s_\tau|s_{\tau-1}, \pi) = s_\tau^\top \mathbf{B}_{\pi,\tau}\, s_{\tau-1}$, where $\mathbf{B}_{\pi,\tau}$ is the transition matrix determined by the policy π.

Using these identities, the joint distribution can be written as:

$$p(o_{1:T}, s_{1:T}|\pi) = s_1^\top D \prod_{\tau=1}^{T} o_\tau^\top \mathbf{A}\, s_\tau \prod_{\tau=2}^{T} s_\tau^\top \mathbf{B}_{\pi,\tau}\, s_{\tau-1} \tag{1.17}$$

where the one-hot vectors s_τ and o_τ act as selectors. Their role is simply to select the elements of $\mathbf{A}$, $\mathbf{B}_{\pi,\tau}$, and D corresponding to the relevant state/output pair or current/previous state pair.

Variational message passing is based on the *mean-field approximation*, which assumes that the approximate posterior (or proposal) factorizes into the product of independent distributions [11, 17]. In the POMDPs under discussion here, the mean-field approximation assumes that the proposal $q(s_{1:T}, \pi)$ factorizes as follows:

$$p(s_{1:T}|o_{1:T}, \pi) \approx q(s_{1:T}, \pi) = q(\pi) \prod_{\tau=1}^{T} q(s_\tau|\pi) \tag{1.18}$$

In variational message passing, the approximate posterior over each hidden state is updated by combining local contributions – called *messages* – coming from the factors of the generative model (i.e., square nodes in Figure 1.2) to the connected states. Specifically, each message $\mu_{f\to s_\tau}$ denotes the log-contribution sent from factor f to the state variable s_τ. To derive the update of beliefs about each hidden state from the mean-field factorization, we apply the standard variational, or *coordinate-ascent*, rule [17, 18]:

$$\log q(s_\tau|\pi) \;\propto\; \mathbb{E}_{q(s_{\zeta\neq\tau}|\pi)}[\log p(o_{1:T}, s_{1:T}|\pi)] \tag{1.19}$$

Now we can consider the logarithmic form of Equation 1.16:

$$\log p(o_{1:T}, s_{1:T}|\pi) = \log p(s_1) + \sum_{\tau=1}^{T} \log p(o_\tau|s_\tau) + \sum_{\tau=2}^{T} \log p(s_\tau|s_{\tau-1}, \pi) \tag{1.20}$$

Thus, inserting this term into the coordinate-ascent rule, we have:

$$\begin{aligned} \log q(s_\tau|\pi) \;\propto\; & \mathbb{E}_{q(s_{\zeta\neq\tau}|\pi)}[\log p(s_1)] + \mathbb{E}_{q(s_{\zeta\neq\tau}|\pi)}\left[\sum_{\zeta=1}^{T} \log p(o_\zeta|s_\zeta)\right] + \\ & + \mathbb{E}_{q(s_{\zeta\neq\tau}|\pi)}\left[\sum_{\zeta=2}^{T} \log p(s_\zeta|s_{\zeta-1}, \pi)\right] \end{aligned} \tag{1.21}$$

Therefore, we only need to retain terms that explicitly depend on the hidden variable s_τ. All other terms, which involve states at time steps $\zeta \neq \tau$, do not depend on s_τ and thus behave as constants that are absorbed by normalization. Consequently, three contributions remain:

$$\begin{aligned} \log q(s_\tau|\pi) \;\propto\; & \log p(o_\tau|s_\tau) + \mathbb{E}_{q(s_{\zeta\neq\tau}|\pi)}[\log p(s_\tau|s_{\tau-1}, \pi)] + \\ & + \mathbb{E}_{q(s_{\zeta\neq\tau}|\pi)}[\log p(s_{\tau+1}|s_\tau, \pi)] \end{aligned} \tag{1.22}$$

In other words, three messages can be identified: a *forward message* from the previous state, a *backward message* from the next state, and a *likelihood message* from the observation at the current time point. We can express these messages with the matrix notation used in Equation 1.17:

$$\text{forward:}\;\; \log \mu_{\mathbf{B}_{\pi,\,\tau-1}\to s_{\pi,\tau}} = \mathbb{E}_{q(s_{\tau-1}|\pi)}\left[\log p(s_\tau|s_{\tau-1}, \pi)\right] = \log(\mathbf{B}_{\pi,\tau-1}\, s_{\tau-1}) \tag{1.23}$$

$$\text{backward:}\;\; \log \mu_{\mathbf{B}_{\pi,\,\tau}\to s_{\pi,\tau}} = \mathbb{E}_{q(s_{\tau+1}|\pi)}\left[\log p(s_\tau|s_{\tau+1}, \pi)\right] = \log(\mathbf{B}_{\pi,\tau}^{\top}\, s_{\tau+1}) \tag{1.24}$$

$$\text{likelihood:}\;\; \log \mu_{\mathbf{A}\to s_{\pi,\tau}} = \log p(o_\tau|s_\tau) = \log(\mathbf{A}^{\top} o_\tau) \tag{1.25}$$

Adopting the matrix notation in Equation 1.22, belief updating can be written as[4]:

$$q(s_\tau|\pi) = \sigma\Big(\log \mathbf{B}_{\pi,\,\tau-1} s_{\tau-1,\pi} \;+\; \log \mathbf{B}_{\pi,\,\tau}^{\top} s_{\tau+1,\pi} \;+\; \log \mathbf{A}^{\top} o_\tau\Big) \tag{1.26}$$

where we used a softmax function to obtain a valid probability vector. Note that for $\tau = 1$, the element $\mathbf{B}_{\pi,\,\tau-1} s_{\tau-1,\pi}$ is replaced by the prior over initial states D. These updates could also be reformulated in terms of prediction errors that illustrate the biological plausibility of this message passing scheme [20].

[4]Readers with specific interest in this topic may note that recent implementations of active inference have used a refined algorithm – called *marginal message passing* – that is more robust to problems of overconfidence that arise with variational message passing [19].

1.4.3 Parameter estimation in a POMDP

Having introduced the formal structure of discrete POMDPs, it is now necessary to describe how these models can be inverted on empirical data. In this context, one can distinguish between two types of generative models. The first is the *subjective model*, namely the POMDP we assume the participant is using in order to plan their responses. The second is the *objective model*, which is specified by the experimenter in terms of priors on the parameters of the subjective model. Model inversion consists in explaining the subjective model through a numerically tractable objective model.

Here, it is important to distinguish between hidden states s, which represent hidden variables evolving over time within the POMDP, and parameters θ, which encode the probabilistic mappings of the generative model (e.g., the likelihood matrix $\mathbf{A}$, the transition matrix $\mathbf{B}$, or the initial state prior $\mathbf{D}$). The aim of model inversion is to estimate the parameter values θ that maximize the posterior probability of the model, given observed behavior. Under the assumption of a flat prior distribution over models, this posterior is proportional to the likelihood, so that parameter estimation often reduces to maximizing the likelihood term. Thus, a classical approach is maximum likelihood estimation (MLE), in which one can perform a grid search over a predefined range of values, selecting the combination of parameters that best reproduces behavior. However, a grid search is feasible only for simple models with few parameters and is prone to overfitting.

In cognitive neuroscience a common approach is variational Bayes, in particular the Variational Laplace method [21, 22]. This method assumes that the proposal distribution $q(\theta)$ is Gaussian, and applies a second-order Taylor expansion around the mode of the variational free energy. In this way, closed-form update rules for the mean and covariance of the Laplace approximation F_{LAP} are obtained. The minimum of the variational free energy can then be found by solving:

$$\frac{\partial F_{\mathrm{LAP}}}{\partial \theta} = 0 \tag{1.27}$$

Through gradient descent on F_{LAP}, the approximate posterior is iteratively adjusted until convergence to a stable minimum, providing estimates of the parameters. This method is particularly suited for nonlinear but smooth generative models, which are typical in neurobiological systems. It has been implemented in toolboxes such as SPM [23], and it has been widely applied in Dynamic Causal Models (DCM) [24, 25]. Moreover, the Laplace approximation is considered neurobiologically plausible, since it can be related to predictive coding schemes where the brain updates beliefs about hidden states by local gradient descent on variational energy.

An important step in variational Bayes is the choice of estimation priors. These are not the priors of the participant, but initial values used during model fitting. The algorithm starts from the prior means and iteratively adjusts them to reproduce the participant's behavior, while penalizing excessive deviation from the priors. This reflects the trade-off between accuracy (fitting the data) and complexity (minimizing divergence from prior values). The prior variances play a crucial role in this balance: small variances give more weight to complexity minimization, while large variances privilege accuracy but increase the risk of overfitting. Consequently, the setting of estimation priors, both in terms of means and variances, is a sensitive step in variational Bayes, and it can strongly influence the resulting posterior estimates.

1.4.4 Hierarchical POMDPs

We can extend the POMDP framework to hierarchical, or so-called 'deep temporal', models. In the case of a two-level POMDP, the construction of such models is conceptually similar to the single-level case, since it involves specifying two generative models and arranging them one below the other. The additional step consists in defining how the two models exchange information. As shown in Figure 1.3, message passing between levels is bidirectional: hidden states at the second level provide prior beliefs about the initial states at the first level (top–down solid lines), while posterior beliefs over these initial states, obtained at the end of a first-level trial, are treated as observations by the second level (bottom–up dotted lines). In this sense, the second-level likelihood mapping $\mathbf{A}^2$ defines how posterior beliefs from the first level become observations at the higher level. This hierarchical structure implies that the second-level model evolves at a slower timescale: each time step in a second-level trial corresponds to the posterior beliefs after a complete trial in the first-level model. Thus, the number of first-level trials matches the number of time points in a second-level trial.

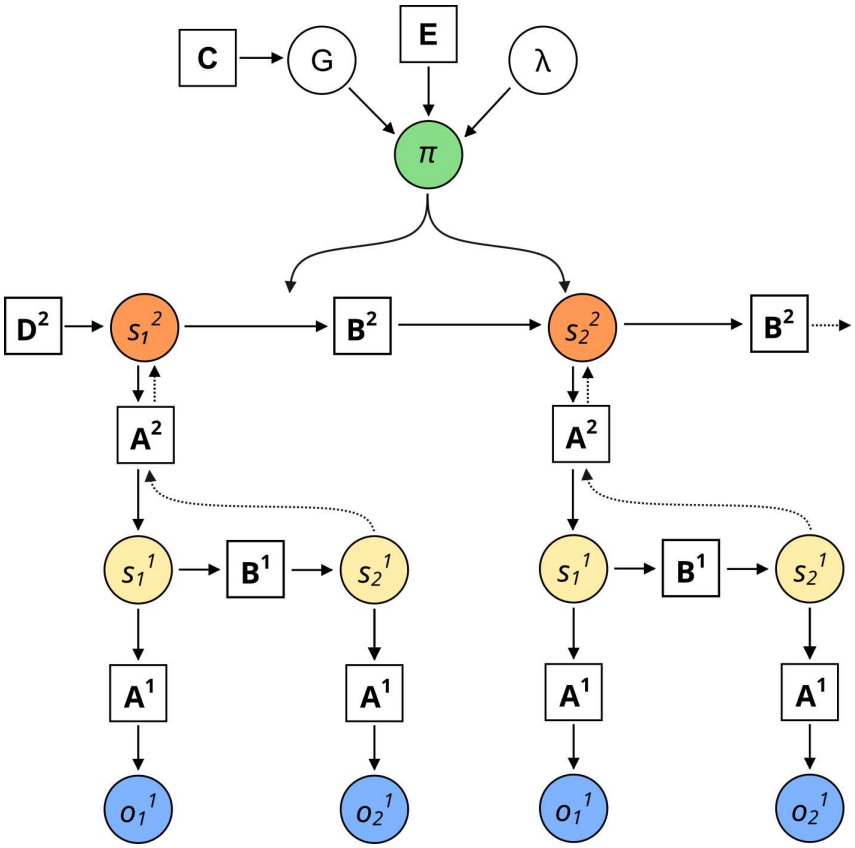

Figure 1.3. Bayesian network representation of a two-level POMDP. Hidden states at the second level provide prior beliefs about the initial states at the first level, while posterior beliefs over first-level states are treated as observations by the second level. In the example shown here, the first level includes two state transitions within each trial, so that second-level beliefs evolve on a slower timescale.

These hierarchical architectures are crucial to model perceptual and cognitive phenomena with nested temporal dynamics, or in contexts where local inferences must be combined to capture higher-order regularities. An intuitive example is reading: the first level infers the identity of individual words, while the second level integrates sequences of

words to infer the meaning of a sentence or a narrative [1, 26].

1.4.5 Software environments and libraries for active inference

From a practical perspective, active inference models can be implemented in both MATLAB and Python. On the MATLAB side, the SPM `DEM/MDP` toolbox provides mature routines for discrete-time, discrete-state POMDP formulations, with unified functions for simulation and inversion via variational message passing (e.g., `spm_MDP_VB_X.m`) [23, 3]. In the open-source ecosystem, `pymdp` offers a Python library targeted at discrete POMDPs, with a modular API for simulating agents, computing expected free energy, and performing policy selection and state inference [27]. More recently, the Julia package `ActiveInference.jl` has also been introduced [28].

At the algorithmic level, besides classical variational message passing, tutorials and research papers have promoted message-passing formulations designed to improve numerical robustness [29, 19]. In parallel, the use of modern automatic-differentiation (autodiff) frameworks (e.g., PyTorch, TensorFlow) is growing for optimization in deep and continuous active inference models; these approaches are promising but not yet standard in cognitive and computational neuroscience.

In terms of computational burden, forward simulations of discrete POMDP agents are typically lightweight – runtime tends to scale approximately linearly with the number of epochs, the size of the latent state space, and the number of evaluated policies. By contrast, model inversion is more demanding. Indeed, simple models with few parameters are often fitted in seconds to minutes on a modern laptop (using MATLAB/SPM or Python/`pymdp`), while inversions for hierarchical or long-horizon tasks can require hours [27, 29].

Chapter 2

Understanding mechanisms of voluntary engagement of mental effort using active inference

The theoretical framework outlined in Chapter 1 provides a formal basis for understanding perception and action as processes of inference under uncertainty. This chapter represents a transition from theory to data: it applies the active inference framework to model voluntary engagement of cognitive effort in a classical behavioral task. The results of this work have been submitted for publication and are currently under review.

2.1 Introduction

Mental effort is a widely used notion in daily life, often associated with the concepts of willpower and self-control. Students are told to put more effort into doing their homework. Adults may have to learn the appropriate balance in their job between insufficient engagement and excessive exertion that may result in burnout and be ultimately inefficient. Despite its common use, mental effort remains a somewhat elusive concept in cognitive neuroscience [30]. Part of the problem is that mental effort has at least three distinct connotations [31, 32, 33]. First, it may refer to how many cognitive resources are automatically recruited by a task, which varies according to task difficulty. This aspect of effort – which we will call *exogenous* effort and which is not necessarily conscious – was the subject of Daniel Kahneman's seminal work [34], where it was essentially equated to attentional allocation (see also [35]). The second aspect of effort, which we will call *endogenous*, is an executive one, related to self-control and the ability to voluntarily modulate the degree of engagement in a demanding task [30, 36]. The third dimension of effort is an *affective*, consciously experienced one [37], usually related to the aversive feeling (but not always, see [38, 39, 40]) associated with the performance of a laborious undertaking [41, 42]. These three aspects may not be easily dissociable and are often closely linked, as exemplified by the finding that emotional arousal not only facilitates physical effort but also decreases the perception of effort [43].

Two other factors are woven deeply into the fabric of mental effort, namely individual motivation and the foreseen consequences of our actions. Several researchers have argued for an intrinsic relationship between the amount of cognitive effort invested in a task and its expected reward, within a cost-benefit computational framework [44, 45, 46, 47], often from a neuroeconomic perspective [48]. A key role appears to be played by dopaminergic

transmission and mesolimbic circuits [49, 50, 51, 52, 53, 54], although cholinergic, noradrenergic, and serotonergic processes are also likely involved [55]. Other scholars have proposed that the subjective feeling of effort results from a prediction of the degree to which the ongoing task will disrupt homeostasis [56, 57].

The active inference approach licenses a fresh outlook on the topic of cognitive effort, by defining it as a divergence between the probability distribution over the courses of action (policies), conditioned on the current context, and the probability distribution over the same policies that reflect our habitual behavior [58]. Put more simply, whenever we perform a demanding task, we face a tension between the behavioral strategy we would follow 'automatically', and the one which is required for the correct performance. This tension, or divergence, is taken as a direct measure of the amount of effort associated with the performance. This approach is in line with descriptions of decision making as a process that pits effortful policies against habitual ones [59, 60], which have been more recently cast in terms of model-based vs. model-free strategy selection (see [61]).

Among the many experimental psychology paradigms designed to contrast habitual responses, one of the most widely used is the Stroop color-word interference task. In this task, participants are shown words that represent color names printed in different colors. In the 'word' condition, they are asked to read the words, that is, to report the displayed text. In the 'color' condition, participants are asked to report the font color of the presented word. Crucially, in different trials, the text and font color may be *congruent* (i.e., the word 'RED' in red fonts), or *incongruent* (i.e., the word 'RED' in green fonts). Given our habitual tendency to read words, responding to stimuli by stating their font color is a policy that requires more effort compared to the natural action of reading the word text.

An active inference model of the Stroop task was recently shown to reproduce various characteristics of actual behavioral and neurophysiological data [58]. The model relied on two key parameters, c and e, which were used to represent the participant's motivation to perform the task correctly and the strength of their habitual tendency to read words (rather than name their colors). The present study uses empirical behavioral data collected from a sample of volunteers performing the Stroop task to invert an adapted version of that model and recover individual estimates of c and e. Crucially, a modification of the experimental paradigm was introduced: participants were asked to perform the task on different blocks either "with maximum exertion" or "as relaxed as possible". This modification follows a previous neuroimaging study [31] that demonstrated measurable differences in both behavior and functional imaging findings with this intervention. Therefore, in addition to the *exogenous* effort demands of the classical Stroop task, an *endogenous* effort component was introduced as an explicit experimental manipulation by instructing participants to apply varying levels of voluntary effort. The aim of this study was to investigate the processes that take place when people respond to encouragement to exert more effort. More specifically, we want to assess the relative evidence for two hypotheses about the processes underlying the intentional engagement of cognitive effort: is voluntary effort mediated by (1) increasing the motivation for an optimal performance (e.g., "I'll do the task as if it was the most important thing today"), or (2) suppressing the automatic habitual response (i.e., reading the word)? A further two hypotheses that follow from this are that both (1) and (2) may be in play, or that neither (1) nor (2) provide adequate explanations for the deployment of voluntary effort. While this study is behavioral, with no direct neural measurements, and limited in terms of what we can say about the neural underpinnings of effort, there is a wide literature relating concepts

of effort to specific aspects of brain anatomy and physiology, to which we will relate in the Discussion (see Section 2.4).

2.2 Methods

2.2.1 Participants

Twenty volunteers (12 females; mean age: 27.9 ± 5.7 years; range: 18–43 years) took part in the study. A history of psychiatric or neurological disorders and use of psychoactive medications were considered exclusion criteria. The study was carried out according to the 2013 version of the Declaration of Helsinki, after approval by the local Ethics Committee (protocol number: CEAR 2024/0144289). Written informed consent to participate in the study was obtained from all volunteers.

2.2.2 Experimental design

A version of the color-word Stroop task with a finger-press response modality via a button box was implemented. Participants were instructed to focus on visual stimuli presented on a laptop screen using the PsychoPy software [62]. The stimuli consisted of 4 colored words ('RED', 'GREEN', 'YELLOW', and 'BLUE'), which were displayed either in a semantically matching font color for *congruent* trials (e.g., the word 'RED' in red fonts), or in a non-matching font color for *incongruent* trials (e.g., the word 'RED' in green fonts). The *Stroop interference effect*, which is deemed to reflect effortful cognitive control, refers to participants exhibiting longer response times and a higher number of errors during incongruent trials, compared to congruent ones.

Each participant completed four runs of the experimental task. Each run consisted of 96 trials, yielding a total of 384 trials. The intertrial interval (between a response and the onset of the following stimulus) was set at 1 s, and the participants were not constrained by a time limit for responding. Each run was divided into four blocks of 24 stimuli. In two of these blocks, participants were asked to respond to font color, while in the other two they were instructed to respond to the written text (Figure 2.1).

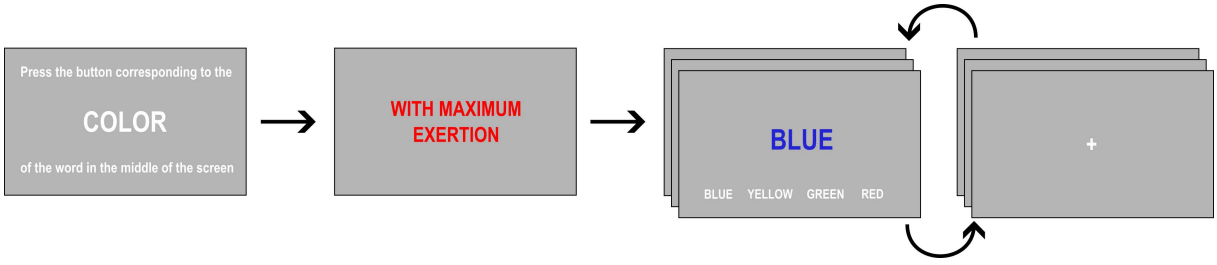

Figure 2.1. Structure of an experimental block. Prior to each block, participants were informed by textual cues about the target ("respond to the text" or "respond to the font color") and the degree of effort they were expected to invest in it. The word stimuli were presented centrally, with a bottom row of color names labels (in white ink) reminding the participant of the position of the corresponding buttons on the response box.

Crucially, as in Khachouf et al. [31], participants were instructed to perform alternating runs with two distinct levels of effort: (a) "with maximum exertion" (EXERT condition) or (b) "as relaxed as possible" (RELAX condition). As a consequence, our experimental

design had 3 factors: *effort* (EXERT or RELAX), *target* (word or color) and *congruency*. Note that the instructions to the participants focused on differentiating their mental attitude adopted in performing the task, rather than on achieving a better (more accurate and fast) performance in the EXERT vs. the RELAX condition. Thus, the ensuing changes in behavioral responses can be attributed to the participants' attempt to execute an intentional modulation of effort. The instructions were displayed in the center of the screen for a period of 2 seconds at the beginning of each block, reminding the participants to put in high or low effort right before performing the task (Figure 2.1). To avoid order effects, runs and blocks were counterbalanced across participants. In order to ensure full comprehension of the task, all participants completed a practice run prior to the start of the actual data collection.

2.2.3 Subjective task-load ratings

At the end of each run, participants were asked to rate their subjective workload using the NASA-TLX rating instrument [63]. This consists of six questions, presented in random order, addressing the following phenomenological dimensions:

1. *Mental demand*: how much thinking, deciding, or calculating was required
2. *Physical demand*: the amount and intensity of physical activity required to complete the task
3. *Temporal demand*: the amount of time pressure involved in completing the task
4. *Effort*: the degree of exertion required to maintain the participant's performance level
5. *Performance*: the perceived level of success in completing the task
6. *Frustration level*: how insecure, discouraged, or content the participant felt during the task

Each question was displayed in the center of the screen, followed by a horizontal line labeled 'very low' on the left and 'very high' on the right. Participants responded by positioning a cursor on the line and their selection was subsequently converted to a score of 0 to 10 (inverted scores were used for the *Performance* reports).

2.2.4 Descriptive and basic statistics of collected data

Accuracy and response time data were examined via standard descriptive statistics, stratified by all experimental conditions. To confirm the presence of the Stroop interference effect (*congruent* vs. *incongruent*), we performed a three-way repeated measures ANOVA on median response times. Then, to verify the effectiveness of the experimental manipulation, we performed Wilcoxon signed-rank tests comparing the average NASA-TLX ratings of the EXERT and RELAX blocks. This was done separately for each of the six NASA-TLX dimensions, and the results were corrected for multiple comparisons using the Bonferroni method, with a significance threshold set at $p_{corr} < 0.05$.

2.2.5 Generative model of the experimental task

We adopted a modeling approach based on the theoretical framework of active inference (see Chapter 1 for a full description). The model implemented in this study builds on a recently proposed active inference model for the Stroop task [58], with important adaptations tailored to our experimental design. In this section, the core structure of the model is discussed, emphasizing the modifications introduced with respect to [58]. For readers seeking a more comprehensive overview of the original modeling approach, I recommend consulting the cited work; however, Figures 2.2 and 2.3 are designed to provide some mechanistic intuition which should be sufficient to understand the core principles of the modeling used here. The model is a hierarchical POMDP, with a higher (*slow*) level that encodes task instructions and sequences, and a lower (*fast*) level that controls trial-specific actions.

In our POMDP model, the potential actions of participants in response to task instructions are represented by alternative *policies* π. These policies can be viewed as probabilistic beliefs about the type of response to issue (responding to the text or to the font color of the stimulus), which subsequently determine the button pressed on the button box. In our case, there are just two available policies, i.e., "report the word text" or "report the font color". Habitual actions – word reading, in this case – are represented by assigning a higher prior probability, which translates to being 'easier' to perform. In contrast, non-habitual actions – like font-color naming, here – are encoded with a lower prior probability, which corresponds to the requirement of a greater cognitive effort.

The prior over policies before any observation is received has been described in Equation 1.12:

$$\pi_0 = \sigma(\log E - \lambda G)$$

More specifically, the vector G represents the expected free energy and each one of its elements can be expressed as $G_\pi = o_\pi \cdot (\log o_\pi + C)$, where the probability distribution o_π describes the expected observations under the policy π, and C is a log-probability vector encoding prior preferences over observations[1]. G scores the implausibility (i.e., improbability given prior beliefs) of each policy, penalizing *risky* courses of action (i.e., whose anticipated observations differ from prior preferences). On the other hand, E quantifies the strength of habitual bias or mental habit (larger values of E indicate less habitual policies). Note that, while the elements of C score observations, the elements of E score policies. Cognitive effort can be defined as the divergence between *context-insensitive* prior beliefs (E) and *context-sensitive* beliefs about how to act in the present contingencies (G).

Figure 2.2 provides a graphical representation of the model's structure. The outcomes of the slow level function as priors for the fast level and, in turn, the outcomes of the fast level correspond to the observable data. At the slow level, the model features three hidden state factors: an *instruction* state, which determines whether the task involves reading words or identifying font color; a *narrative* state, which changes from an instructional to a response context; a *response modality*, which is the only policy-dependent term in the generative model and determines the strategy without influencing the observables directly

[1] In general, there is an additional term in the expected free energy, dealing with the expected entropy of the conditional distribution representing the beliefs about how hidden states generate observations, but this is suppressed here as it is a constant for our model.

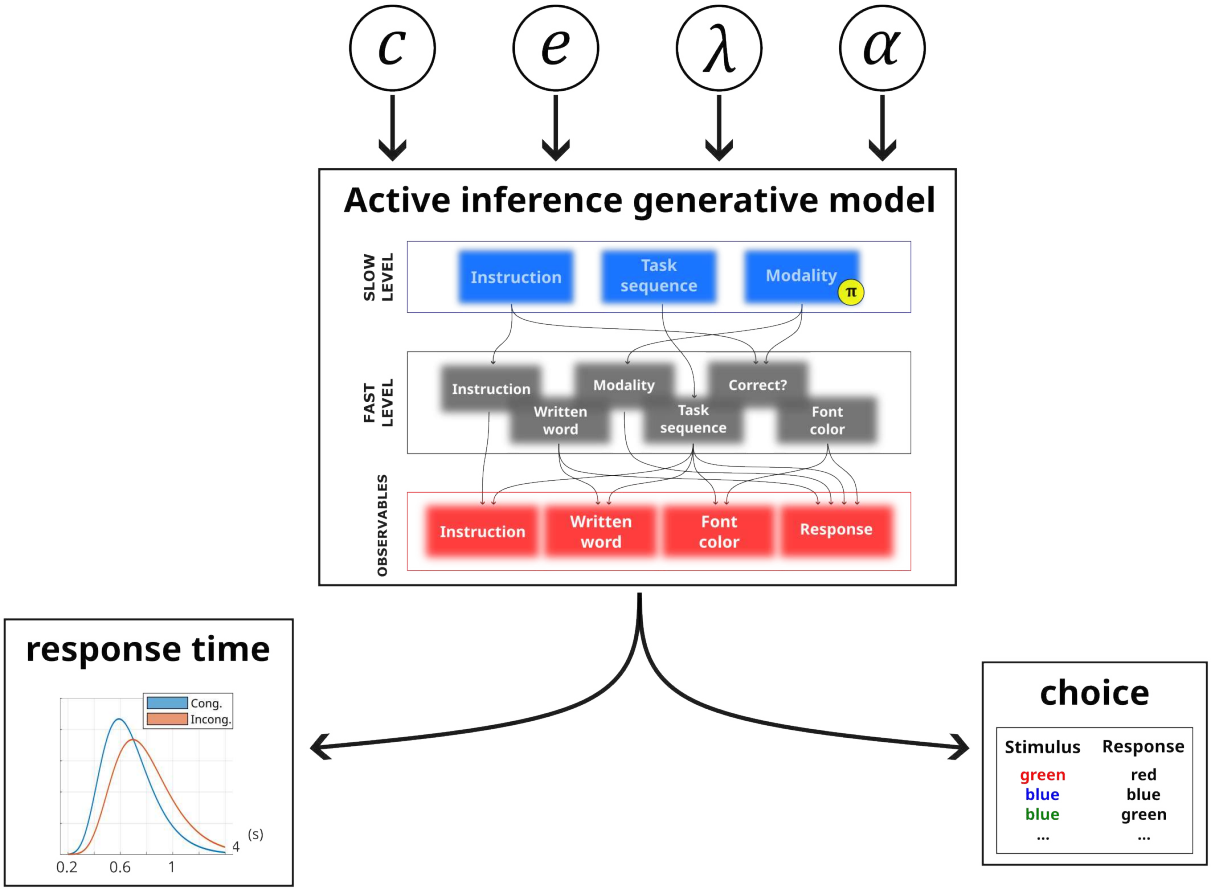

Figure 2.2. Hierarchical structure of the generative model, with a higher (slow) level encoding the task instructions and sequence, and a lower (fast) level encoding trial-specific features. Arrows indicate statistical dependencies: only the boxes related by a dependency are linked by an arrow. Given the values of the parameters c, e, λ and α (see main text), the generative model provides a probabilistic estimation of the action $\mathbf{u}(c, e, \lambda, \alpha)$ – i.e., a vector whose elements are the probabilities for each possible action conditioned upon the parameters – enabling the simulation of both choice and response time data. Indeed, the graph in the bottom-left represents the probability distributions of response times according to the computed confidence in the next response choice, while the inset in the bottom-right shows a simulated response choice for the 'report the font color' condition.

(actions that affect directly the observables are specified at the fast level). The outcomes of the factors of the slow level correspond to the three states of the fast level with the same labels. There are three other state factors in the fast level: the *color* of the font; the *written* word; and a state that reports the *correctness* of the chosen mental action, comparing the instruction and response modality at the slow level. Crucially, this state has no influence over the generated observations.

Figure 2.3 illustrates the dynamics of information processing in the model. In the figure, each level is annotated with a brief description. From the perspective of someone performing the task, we assume she has prior weightings as to which modality she expects to respond with, which are used to contextualize (1) transitions in intended response modalities from stimulus to stimulus as represented in slow level states. These states provide prior weighting (2) for faster states that deal with the transition from a new stimulus to a behavioral response (e.g., from observing the blue stimulus to declaring it

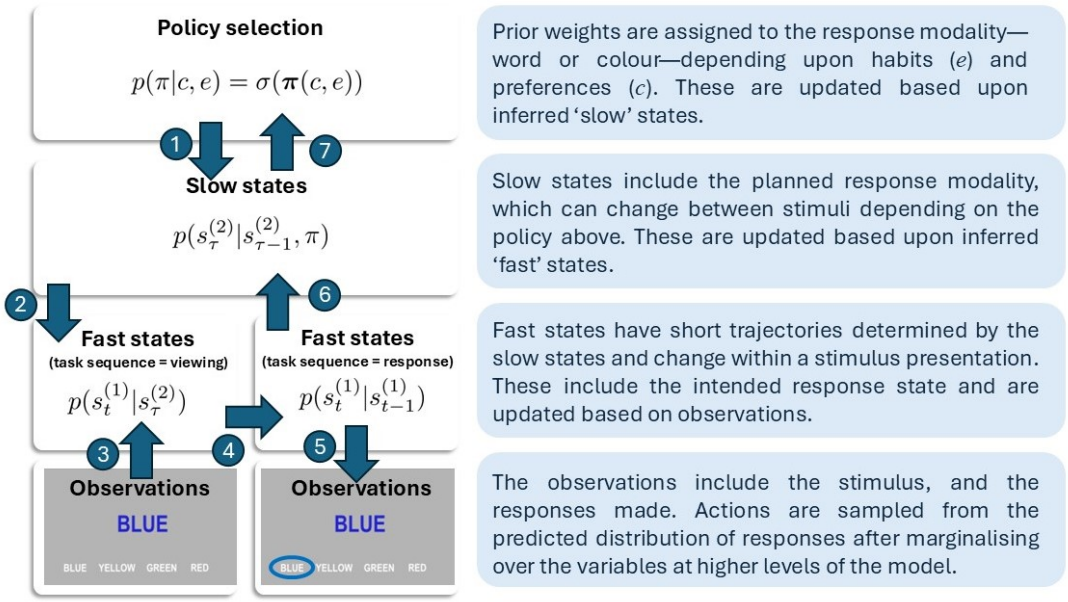

Figure 2.3. This figure is designed to provide some mechanistic detail on how the model assumes participants solve the task of deciding how to respond to a given stimulus presentation for the Stroop task. This is designed to be an accompaniment to Figure 2.2, which details more specifics of how the model was set up to estimate behavior in our analysis.

blue). The prior (2) and observational evidence (3) are combined through variational inference to update these faster states. Still in the fast level, after transitioning (4) from a *viewing* to a *response* state, the response is sampled (5) from a predictive distribution (which depends upon 1-4) over possible responses. The fast states inferred at this stage can be used to update the slower states (6) and the policy (7) through variational inference such that revised priors can be brought to bear on the next stimulus presentation.

The formulation of the generative model here has a few subtleties worth unpacking. The policy variable at the highest level of the model in Figure 2.3 represents decisions about the modality to which we should attend, and not the words and colors themselves. This influences the trajectories of the slow states to form weights for the two modalities. These weights influence the fast states such that the distribution of predicted response (from which the measured response is assumed to be sampled) is a weighted combination of the color and word displayed. As this is a probabilistic weighting, the weights assigned to each modality sum to one, meaning facilitation of one modality necessarily means the suppression of the other. In effect, the (potentially opposing) influence of prior habits and preferences jointly determines the relative weights.

Our analysis focused primarily on the probability vectors C and E which, as described above, reflect the motivation to perform the task well and the habitual bias toward reading the word (*vs.* stating the font color), respectively. In the computational model, these vectors were expressed via their log-scaled forms, c and e, so that:

$$\begin{aligned} C &= \exp(c) \cdot [1 \quad -1]^T \\ E &= \exp(e) \cdot [0.3 \quad -0.3]^T \end{aligned} \tag{2.1}$$

Higher values of c indicate a stronger preference for accurate performance, while higher

values of e indicate a greater strength of the habit to automatically read the word and thus the need for increased cognitive effort to suppress this habitual response. The interaction between these parameters reflects various individual scenarios, such as cases where a strong motivation for accuracy (c) can mitigate the impact of a strong habitual tendency (e) toward word reading over font color naming. Note that in [58], the E vector was defined as $E = \exp(e) \cdot [0.8; -0.8]$. For $e = 0$, it corresponded to a probability of responding to the word text (instead of to the font color) of 85%. In our empirical behavioral data, however, the difference in reaction times between incongruent and congruent trials was smaller than that in the simulations of [58]. Accordingly, we adopted a smaller multiplication factor, reflecting a reduced prior bias toward responding to the word text.

Response choice

The generative model enables simulation of response choices as *actions*. Instead of simply selecting the most probable action, actions are generated by sampling from a probability distribution $\mathbf{u}$, given by the expected observation at the next time step (itself determined by averaging observations conditioned upon policies under a distribution $\boldsymbol{\pi}$ that scores alternative policies based upon their expected free energy for the next time step). A softmax function σ is applied to the log outcome distribution weighted by a parameter λ to account for uncertainty in action not captured purely by this observation distribution:

$$\mathbf{u}_{t+1} \coloneqq \mathbf{u}(c, e, \lambda) = \sigma\left(\lambda \log \sum_{\pi} \boldsymbol{\pi}_{\pi} \cdot \boldsymbol{o}_{\pi,t+1}\right) \qquad \lambda \in \mathbb{R}^{+} \tag{2.2}$$

Effectively, this means that the next controllable observation o_{t+1} – i.e., the button press – is sampled from $\boldsymbol{u}_{t+1}$. The parameter λ, typically referred to as 'inverse temperature', regulates the level of stochasticity in action selection. Higher values of λ result in more deterministic actions, whereas lower values increase variability in decision-making.

Response times

Response times can also be simulated by the model, according to the level of confidence of the agent about her response choice, following a common approach in drift-diffusion models of decision-making [64]. More specifically, the response time is modeled here as a function of the entropy of the predicted response choice distribution at the next time step, $H_{t+1} = -\mathbf{u}_{t+1} \cdot \log(\mathbf{u}_{t+1})$. Higher entropy values correspond to longer response times:

$$r_t \coloneqq r(c, e, \lambda, \alpha) = \exp(\alpha) \exp(n + H_{t+1}) \qquad n \sim \mathcal{N}(0, \frac{1}{16}) \qquad \alpha \in \mathbb{R} \tag{2.3}$$

In this equation, the entropy term can be interpreted as the logarithmic drift rate that governs decision time, with the Gaussian random variable n accounting for the stochastic component of diffusion [64]. The constant $\exp(\alpha)$ represents the baseline response time under conditions of maximum confidence, where entropy approaches zero. In essence, α represents the minimum expected response time, i.e., when the subject is fully certain about her response choice; at each trial, within the active inference model, this minimum reaction time is adjusted by an amount represented by the entropy H.

Simulated behavior

A characteristic feature of many behavioral tasks, including the Stroop, is a *speed-accuracy trade-off*, whereby attempting to respond more quickly often causes a decrease in accuracy.

Since the question of whether investing more effort in a task affects speed, accuracy, or both is a relevant one, it is important that the simulated behavioral data from our generative model exhibit a realistic relationship between speed and accuracy.

Figure 2.4 shows the accuracy and response times of the data generated by the model using various values of the parameters c and e. In panels A and B, when c is slightly below 0 (leftmost columns), an increase in e results in decreased accuracy and faster response times. This means that when the motivation for performing the task well is low, the presence of strong habitual behaviors that are discordant with the task requirements will prevail, producing quick but inaccurate responses. In contrast, in the rightmost columns of the grid plots, where c is greater than 0, increasing e values lead to a (small) decline in both accuracy and response speed. In other words, when motivation is strong, we will generally observe fast and accurate responses, with only slight decreases in performance as the cognitive demands of the task (in terms of its deviation from habitual behavior) increase. Panel C of Figure 2.4 illustrates more explicitly the speed-accuracy trade-off for various values of the c and e parameters.

2.2.6 Model fitting and parameter estimation

We now turn to the central aim of the present approach: to infer the values of hidden causal factors of behavior from measures of task performance and use these to test hypotheses. In our case, this involves estimating the values of the parameters c, e, λ, and α for each participant, by fitting the generative model to the observed data (response choices and response times) via a Variational Laplace procedure (see Section 1.4.3). Our analysis was implemented in MATLAB 2024b, using the functions of SPM toolbox (see Section 1.4.5).

Single-subject level

Our modeling approach uses as data not only the overall accuracy of the responses, but also the sequence of choices (i.e., button presses), allowing for sequential effects to inform model fitting. Therefore, the log-likelihood $\mathcal{L}$ we used for model inversion depends on both response choices and response times, according to the following formula:

$$\mathcal{L}(r_t, o_t, c, e, \lambda, \alpha) = \underbrace{\sum_t \log(o_t \cdot \mathbf{u}_{t-1})}_{\text{choices}} \underbrace{- \sum_t \frac{1}{16}(\log r_t + \mathbf{u}_t \cdot \log \mathbf{u}_t - \alpha)^2}_{\text{response times}} \tag{2.4}$$

where o_t denotes the observed response choice (as a one-hot vector[2]) and r_t denotes the response time at time step t.

Bayesian inference requires the definition of prior distributions for the model parameters. In our case, all prior distributions were chosen as Gaussian, resulting in normal posteriors. The prior means of c and e were set at 0. In this way, the prior preference

[2]A one-hot vector is a vector that is zero everywhere except for a single position where it values one. Our response vector has five levels, corresponding to the four different button presses plus a null level, representing the absence of a response. As instance, a response issued by pressing the second button in the button box will be coded as $o_t = [0, 1, 0, 0, 0]$. As a consequence, the dot product $o_t \cdot \mathbf{u}_{t-1}$ in Equation 2.4 extracts the probability assigned to the chosen action.

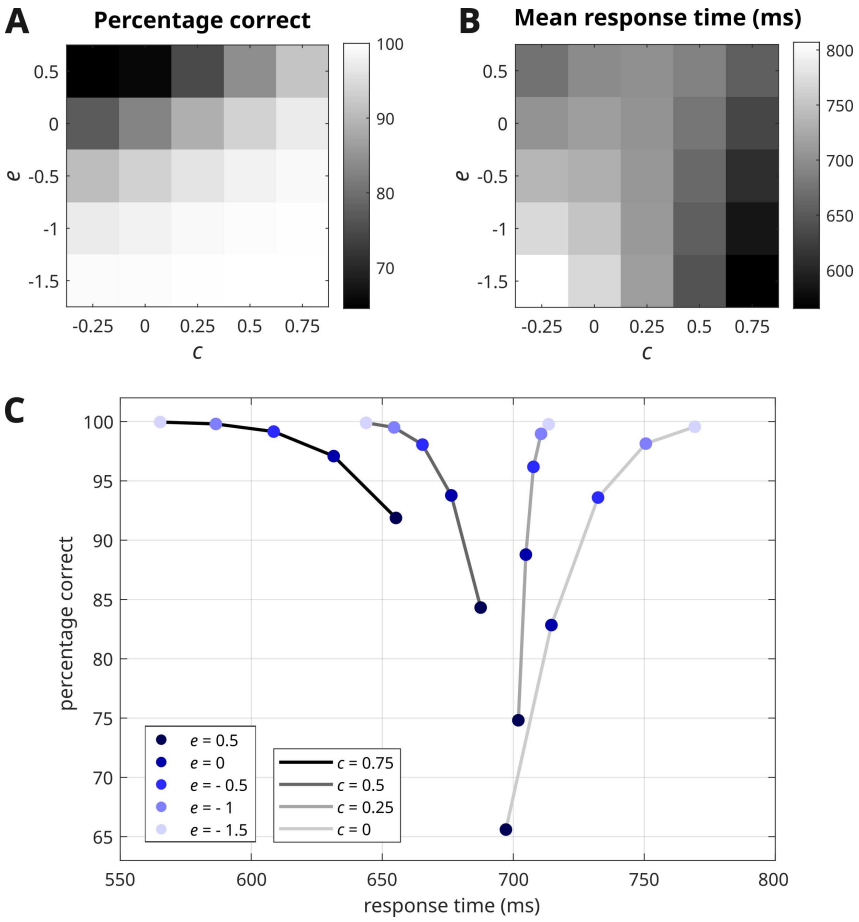

Figure 2.4. Panels A and B illustrate the average accuracy and response times, respectively, for simulated data under different prior beliefs. Panel C integrates this information into a two-dimensional plot, illustrating the relationship between speed and accuracy across different fixed levels of motivation (c) and habit (e). The gradient of the curve may be either positive or negative, indicating that an increase in speed can be associated with either a decrease or an increase in accuracy. In all these simulations, the other parameters were set as follows: $\lambda = \frac{1}{4}$, $\alpha = \ln(\frac{1}{2})$, which correspond to the values used in the simulations from Parr et al. [58]. Note that the graph includes some zones of very poor accuracy, which correspond to high values of e and low values of c. As our actual behavioral data did not exhibit such low levels of performance, we did not expect to recover the corresponding combination of parameter values, and indeed, these were not obtained.

(encoded as a probability) for being correct was set at $\exp(2) \approx 7.4$ times the prior preference for being incorrect[3]; as for the prior bias for habitual policies, our specification of the E vector implied that 65% of the time the participant expected to read the word, while 35% of the time she expected to name the color[4]. In order to ensure the positivity

[3]The prior preference for choosing the 'correct' response modality (the one prescribed by the instructions) at the slow level was defined as $C = \exp(c) \cdot [1; -1]^T$. As a consequence, a prior value of $c = 0$ corresponds to a log-probability vector $C = [1; -1]$, where the first event is $\exp(1)/\exp(-1) = \exp(2)$ more probable than the second.

[4]Cognitive demand is represented as a log-probability vector $E = \exp(e) \cdot [0.3; -0.3]^T$. A prior value of $e = 0$ corresponds to a log-probability vector $E = [0.3; -0.3]$. When applying the softmax operator, this results in a probability of $\exp(0.3)/(\exp(0.3) + \exp(-0.3)) \approx 0.65$ for the first event

of the parameter representing the stochasticity (or inverse temperature) in the model, we defined $\lambda = \exp(\zeta)$ and modeled ζ using a Gaussian prior with a mean of $\ln(\frac{1}{4})$. Finally, the parameter α (associated with the lower bound of response times) was modeled using a Gaussian prior of mean $\ln(\frac{1}{2})$.

As for the variance of the Gaussian priors, we used $\frac{1}{4}$ for all parameters. This is a rather large value when dealing with log parameters, meaning that we are not using strong priors. In other words, we are allowing these parameters significant flexibility to deviate from their initial values, ensuring that our model fitting is predominantly guided by the data.

The data from the EXERT and RELAX runs were used to fit the model separately for each of the 20 participants, yielding a total of 40 sets of model parameters. To evaluate the model's predictive reliability, we also conducted posterior predictive checks. Indeed, for each participant and each of the eight experimental conditions (*effort* × *target* × *congruence*), we generated simulated behavioral data by sampling from the generative model using parameter estimates obtained from the original model inversion. Simulated data were processed identically to empirical data to compute average response times (RTs) per participant–condition combination. To quantify the agreement between simulated and observed data, we regressed the mean observed RTs on the mean simulated RTs using a Bayesian linear mixed-effects model with random slopes for both participants and experimental conditions. Finally, a parameter recovery analysis was performed. For each participant, we simulated behavioral data using the fitted parameters (c and e) and re-inverted the model to obtain new parameter estimates.

Group-level

The group-level analysis aimed to determine whether the motivation parameter (c) and the bias parameter (e) differed between the two effort conditions. To address this, we applied Parametric Empirical Bayes (PEB; [65]), which updates individual estimates and predicts group-level c and e parameters, representing the difference between the EXERT and RELAX conditions. This approach mitigates overfitting and is well-suited for small sample sizes.

PEB requires the specification of a design matrix to account for sources of variability between subjects. While the first column contains all ones and represents the intercept, the second column typically refers to the effect of interest; in our case, it contained 1 for EXERT models and 0 for RELAX models. The remaining columns captured the subject-specific variability and were mean-centered[5] (Figure 2.5).

Bayesian Model Reduction (BMR; [65]) tests reduced versions of the full model obtained with PEB, selecting the most plausible one and highlighting key group-level effects. We used BMR to compare four models: the *full model*, which includes the effect of voluntary effort on both c and e; a *null model* with no effect of voluntary effort on either parameter; and two models with the effect of voluntary effort on only one parameter. In the BMR scheme, if a model has a posterior probability greater than 90% its parameter estimates are taken as the final values. Otherwise, the final estimates are computed as a weighted average of the estimates from all the models, where the weights are the posterior probabilities of the models (i.e., Bayesian model averaging). Credible intervals are also

[5]For instance, the second column assigns 1 to both cells corresponding to the first participant and $-2/(40-2) \approx -0.0526$ to all other cells.

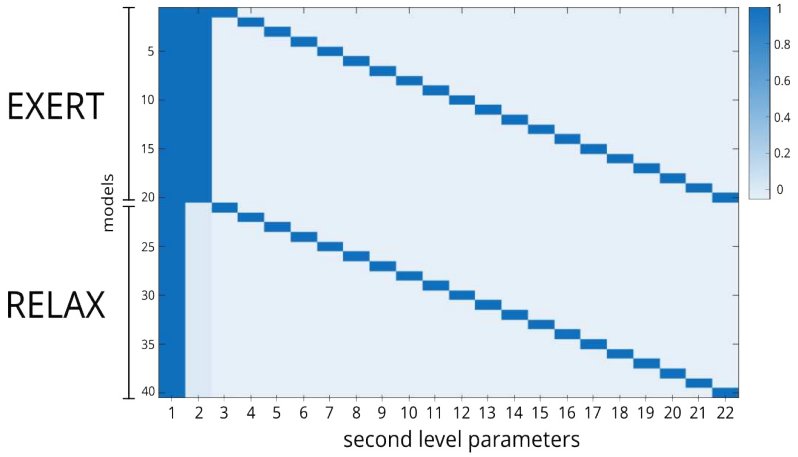

Figure 2.5. Design Matrix of the PEB analysis. The first column represents the intercept, while the second refers to the effect of interest, which is the effect of intentional effort on motivation c and bias e. The first twenty rows represent the parameters relative to the EXERT data, and the last twenty rows the parameters relative to the RELAX data.

computed using this procedure, thus we considered a posterior probability of being non-zero higher than 90% as a criterion for reasonable evidence of the effect of voluntary effort on the specified parameter. This allowed us to test if voluntary effort is mediated by (1) increasing motivation for being correct, (2) suppressing the automatic habitual response, by both (1) and (2), or by neither (1) nor (2).

2.3 Results

2.3.1 Descriptive and basic statistics of collected data

Accuracy and response times

The observed accuracy and response times are presented in Table 2.1 and Figure 2.6. In both the EXERT and RELAX blocks, the *color incongruent* condition – where participants had to respond to the font color and the stimulus was incongruent – yielded the lowest accuracy and the longest response times. Performance in the *word incongruent* condition – where participants had to respond to the stimulus text and the stimulus was incongruent – was comparatively better, even if it remained slower and less accurate than in all *congruent* conditions. In the RELAX condition, response times were generally slower compared to the EXERT condition, although a relevant reduction in accuracy was only observed in the color-incongruent condition (94.0% vs. 95.7%).

Within-subject statistical analysis on the Stroop effect

A three-way repeated-measures ANOVA was performed to investigate the effect of *effort* (EXERT vs. RELAX), *target* (color-naming vs. word-naming), and *congruency* (congruent vs. incongruent), as well as their interactions, on median response times. The results are shown in Table 2.2.

The analysis revealed a significant main effect of *effort*, indicating that participants responded faster during EXERT compared to RELAX trials ($p < 0.001$). A robust main

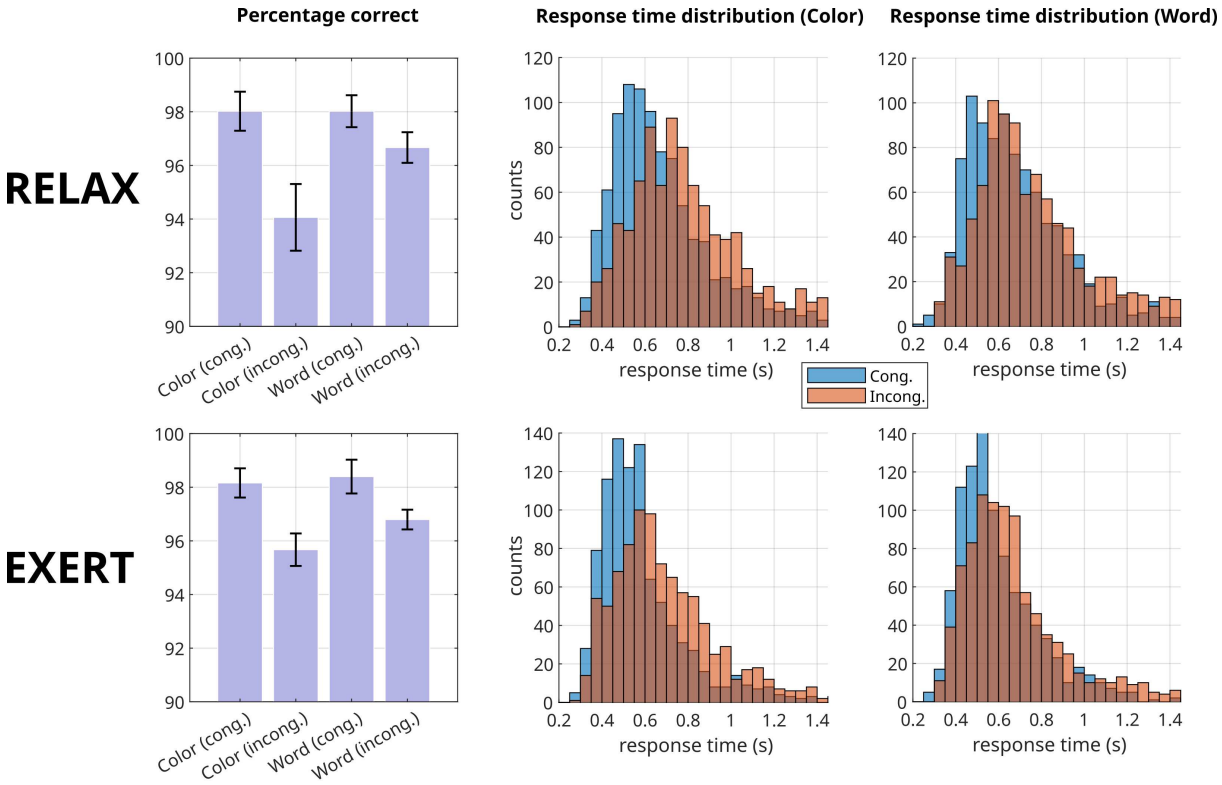

Figure 2.6. Graphical representation of the observed behavioral data. The left column illustrates accuracy across the various conditions (error bars represent within-subject standard errors), while the remaining two columns display the distributions of the observed response times for the two tasks ('report the font color' and 'report the word text').

Table 2.1. Average values for the observed accuracy and response times (RTs) across all eight conditions. For response times, within-subjects 95% confidence intervals (CI) were computed using the standard Student's t-based interval on the distribution of participant-level mean response times.

effort	target	congruency	average accuracy	average RT	RT 95% CI
EXERT	color	congruent	98.2%	594 ms	531 - 657 ms
EXERT	color	incongruent	95.7%	719 ms	647 -791 ms
EXERT	word	congruent	98.4%	607 ms	556 - 659 ms
EXERT	word	incongruent	96.8%	704 ms	638 - 770 ms
RELAX	color	congruent	98.0%	694 ms	628 - 761 ms
RELAX	color	incongruent	94.0%	859 ms	774 -945 ms
RELAX	word	congruent	98.0%	700 ms	638 - 762 ms
RELAX	word	incongruent	96.7%	806 ms	727 - 885 ms

effect of *congruency* was also observed, with slower responses for incongruent relative to congruent stimuli ($p < 0.001$). No significant main effect of *target* was found ($p = 0.559$).

Importantly, a significant interaction between *effort* and *congruency* ($p = 0.017$) suggests that the Stroop effect was modulated by effort, with a reduced interference effect in the EXERT condition. The interaction between *target* and *congruency* did not reach significance, although the result approached the threshold ($p = 0.063$), indicating a trend

Table 2.2. Within-subject effects obtained from a three-way repeated measures ANOVA on median response times.

Within-subject effect	Estimate	df	F	p
Effort	0.360	19	21.880	< 0.001
Target	0.005	19	0.353	0.559
Congruency	0.406	19	89.065	< 0.001
Effort * Target	0.001	19	0.310	0.584
Effort * Congruency	0.008	19	6.870	0.017
Target * Congruency	0.022	19	3.896	0.063
Effort * Target * Congruency	0.002	19	1.200	0.287

toward differential interference between color-naming and word-naming tasks.

To further investigate these effects, contrast analyses were performed comparing the Stroop effect (*incongruent* vs. *congruent*) across the four combinations of *target* and *effort*. Also, the effect of effort on the Stroop effect was tested. Results are reported in Table 2.3. The Stroop effect was found to be present for both levels of *target* and both levels of *effort*. However, the effect of effort on the Stroop effect was significant only in the color-naming task. Thus, only in the color-naming task the intentional increase in applied cognitive effort reduced the interference typically observed in the Stroop effect.

Table 2.3. Contrast analysis examining the Stroop effect (*incongruent* vs. *congruent* response times) across *effort* and *target* conditions.

Contrast	Estimate	df	t	p
color: EXERT (incong. vs. cong.)	0.103	19	8.625	< 0.001
color: RELAX (incong. vs. cong.)	0.146	19	6.269	< 0.001
word: EXERT (incong. vs. cong.)	0.071	19	6.713	< 0.001
word: RELAX (incong. vs. cong.)	0.083	19	3.549	0.002
color: RELAX (incong. vs. cong.) - EXERT (incong. vs. cong.)	0.044	19	2.306	0.033
word: RELAX (incong. vs. cong.) - EXERT (incong. vs. cong.)	0.012	19	0.732	0.473

Subjective task load measurements

Three of the six NASA-TLX dimensions had statistically significant differences between the EXERT and RELAX blocks. Specifically, in the Wilcoxon signed-rank tests, higher values for the EXERT, compared to the RELAX, runs were reported for *mental demand* (mean 6.57 vs. 5.15, effect size 0.87, 95% CI $[0.72, 0.96]$, $p < 0.001$), *temporal demand* (mean 5.24 vs. 3.87, effect size 0.96, 95% CI $[0.90, 0.99]$, $p < 0.001$), and *effort* (mean 6.74 vs. 4.83, effect size 0.99, 95% CI $[0.97, 0.99]$, $p < 0.001$).

2.3.2 Model fitting and parameter estimation

Single-subject level estimation

Figure 2.7 illustrates the model's parameter estimates across subjects. While the motivation-related parameter c had tight credible intervals and is consistently larger in the EXERT condition, the habit-related parameter e showed greater variability among participants.

We did not observe any significant correlation between the NASA-TLX ratings and the estimated values for the c and e parameters – both separately for the EXERT and RELAX conditions, and for the EXERT-RELAX differences.

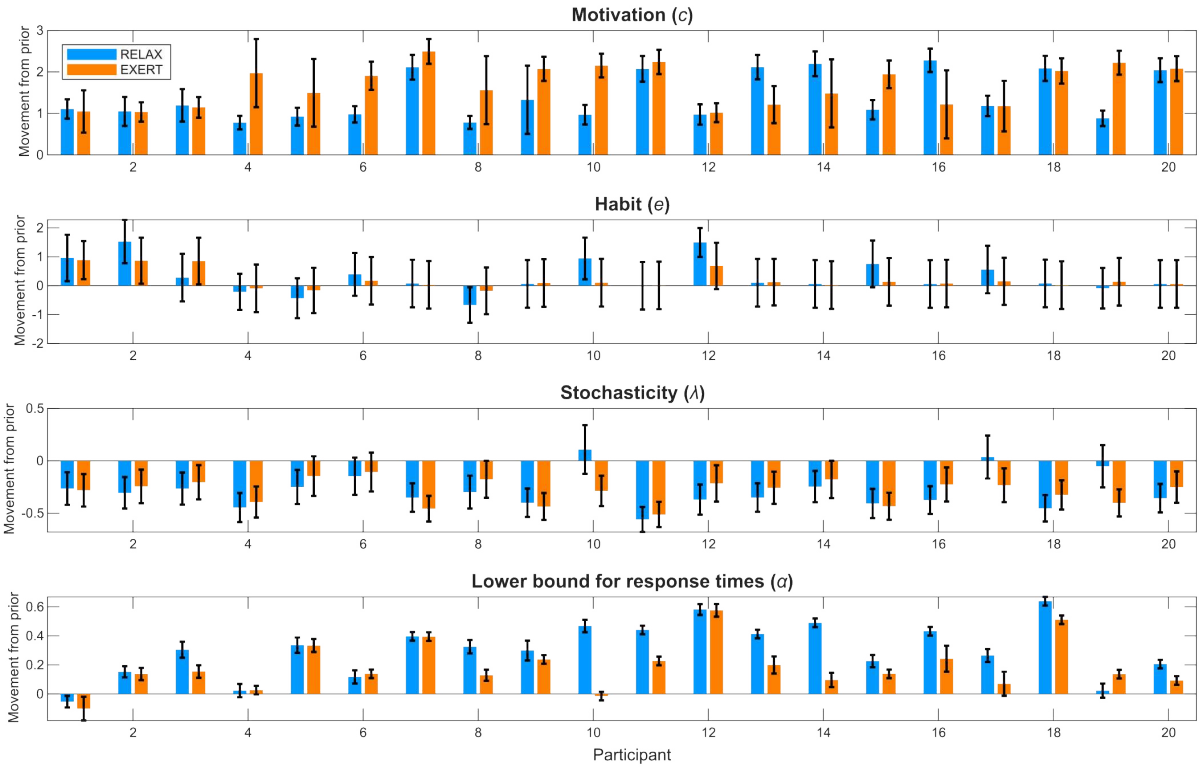

Figure 2.7. Posterior estimates and credible intervals for the parameters c, e, λ, and α in terms of their deviation from prior values.

Posterior predictive check and recovery analysis

The Bayesian linear mixed-effects model with random slopes for both participants and experimental conditions used for posterior predictive checks yielded a posterior median slope of 0.95 (95% CI [0.74, 1.17]) and an intercept of 0.11 (95% CI [–0.04, 0.26]), indicating a close correspondence without systematic bias. Figure 2.8 reports a scatterplot comparing observed and simulated response times.

Values obtained from the recovery analysis were strongly correlated with the original ones ($r = 0.90$, $p < 0.001$ for c; $r = 0.89$, $p < 0.001$ for e), indicating excellent preservation of the inter-individual ranking. A Bland–Altman analysis [66] further revealed negligible mean bias (–0.08, 95% limits of agreement [–0.81, 0.48] for c; –0.01, 95% limits of agreement [–0.70, 0.68] for e), suggesting no systematic over- or underestimation across the parameter space.

Note that also in [58] the authors simulated data based on various combinations of prior mean values for c and e. Then, they inverted the generative model and the estimates were compared to the hypothesized values of c and e. Their results showed that the recovered estimates for the c parameter were in good agreement with the original ones. For the e parameter, while their recovery was relatively poor, the rank order was preserved (i.e. higher hypothesized values of e were associated with higher posterior values of e). However, it is worthwhile to note that the values chosen in [58] for the prior mean of

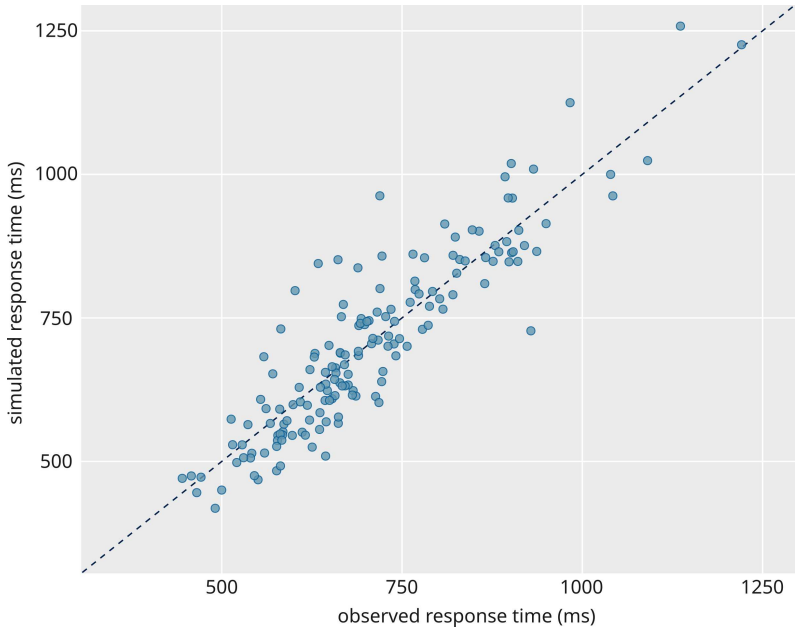

Figure 2.8. This scatterplot compares mean observed response times (RTs) and corresponding simulated RTs across all participants and experimental conditions (each point represents one participant–condition combination). The dashed diagonal line indicates perfect agreement between observed and simulated data.

the parameters were relatively small, ranging between −0.25 and 0.25. In contrast, the estimated values for the parameter in this study moved further away from zero. This might be due to a number of factors (including our specific manipulation of engaged effort), and it justifies the application of a specific recovery analysis on our real behavioral data.

Group level analysis

For a rigorous group analysis, we employed Parametric Empirical Bayes (PEB) followed by Bayesian Model Reduction (BMR), to investigate how the c and e parameters can explain the effect of intentional investment of effort in the task. This analysis identifies the model that best explains the observed differences between the EXERT and RELAX conditions.

The left panel of Figure 2.9 shows that, although the 'full model' that includes the effect of effort on both the c and e parameters exhibits the highest posterior probability (65.3%), the model with effort affecting only the c parameter also demonstrates a substantial posterior probability (34.7%), while the model considering only e has a negligible posterior probability. As all models have a posterior probability less than 90%, the final parameter estimates are computed as a weighted average of the estimates from all models, where the weights are their posterior probabilities.

Therefore, c is the only parameter whose variation with respect to endogenous effort has a probability of being non-zero higher than 90% (Figure 2.9, middle and right). This finding suggests that the intentional engagement of effort primarily affects motivation, as the parameter c, in contrast to the parameter e, is significantly higher in the EXERT condition compared to the RELAX one. Nevertheless, the findings of this group-level analysis do not preclude the possibility that, for certain individuals, endogenous effort may be mediated by changes in the e parameter.

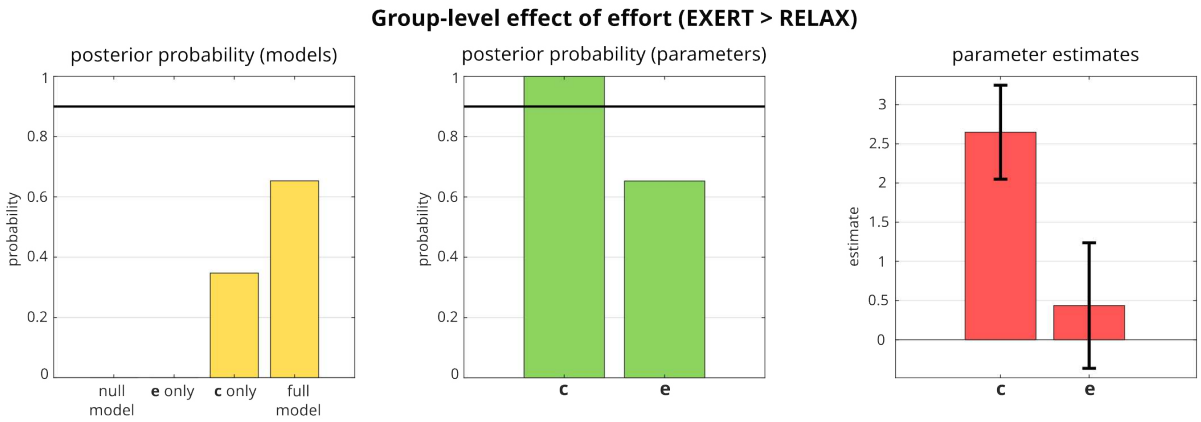

Figure 2.9. Results of group-level analysis. BMR was used to compute posterior probabilities of four models: a 'full model' that included both c and e, a 'null model' that excluded both, and two models that included only one of the parameters. The left panel shows the posterior probability for each model, with an horizontal line indicating the 90% probability level. The middle panel displays the posterior probability of being non-zero for each second-level parameter (representing the difference in c and e between EXERT and RELAX blocks). These values are computed as the sum of the posterior probabilities of the models where these parameters are present; for example the posterior probability of group-level c is the sum of the posterior probabilities of the full model and the 'only c' model. Note that the full model has a larger posterior probability than the one which includes only the c parameter, providing some evidence for the relevance of both parameters in explaining the data. However, this can be nuanced by computing final parameter estimates from a weighted average (i.e., Bayesian model averaging), where the weights are the posterior probabilities of the models. Indeed, the right panel of the Figure shows these weighted averages with 90% credible intervals. From this graph, we can see that the final posterior probability distribution for group-level e parameter has 90% credible intervals that include zero. As such, if we apply an arbitrary 90% thresholding, we would be unable to conclude that this parameter differ between EXERT and RELAX condition.

2.4 Discussion

We studied the effects of intentional investment of mental effort, using a Stroop task and asking participants to perform it with maximum exertion (EXERT) or as relaxed as possible (RELAX), in alternating runs. Incongruent stimuli in the classical Stroop task already require a degree of cognitive effort to suppress the automatic tendency to read the text in favor of identifying the font color. Thus, in addition to this *exogenous* effort, which reflects the cognitive load imposed by task demands, our experimental design introduced an explicit *endogenous* component of effort, namely a voluntary modulation of the participant's investment in performing the task. While exogenous effort is driven by task difficulty, endogenous effort involves self-regulation and intentional control. As the experimental conditions EXERT and RELAX differed only in the instruction about *how* to do the task (not about *what* to do), the observed behavioral differences between the two conditions can be interpreted as reflecting the cognitive processes associated with the attempt to intentionally vary the degree of invested mental effort.

Parr et al. [58] developed an active inference model of the word-color Stroop task simulating known features of task performance under a novel conceptualization of the construct of mental effort. We used a slightly modified version of this model and fitted it to actual performance data from the Stroop task. This approach allowed us to evaluate

the relative evidence of two hypothesized mechanisms of intentional effort: a weakening of the compulsive power of habitual policies, on the one hand, and an increase of intrinsic motivation, on the other. One could ask why, given that many Stroop paradigms make use of only the color-naming condition, we have elected to also include a word reading condition. The reason for this is that, in estimating the e parameter, dealing with the habitual effect of word reading, it is useful to be able to vary the demands placed on this parameter over and above the effects of some trials being incongruent.

The results showed that voluntary engagement of effort in the EXERT condition was associated with a significant increase in the motivation parameter (c) only, suggesting that what participants do when asked to engage maximum effort is to endogenously intensify their motivation, possibly by modulating precision weighting of connections in reward circuits. This seems plausible as the alternative – i.e., modifying directly the strength of our behavioral habits – may not be feasible in the sense that we may not have direct, operational access to the relevant mechanisms (or more simply that modifying habit strengths requires a longer time frame and cannot be performed in real time). It is also possible that this autonomous intensification of motivation is implemented by activating reward-related processes (not necessarily in an explicit, conscious manner), which would align with experimental data showing that increasing the magnitude of a reward leads to greater effort investments [67, 68] and improves executive function [69].

It is important to qualify some of the language used. Statements about motivation and demand here refer explicitly to the inferred parameters c and e, which may or may not reflect commonly held psychological definitions of these attributes – although we suggest that they do reflect a formalization of at least some definitions. In other words, stating that 'an instruction to voluntarily exert oneself led to an increase in their motivation' is really shorthand for saying that it led to an increase in the estimate of the c parameter that best explained their behavior. The meaning of the c and e parameters comes from their influence over decision making in the models in which they appear. The former determines the degree to which a decision is made to maximize the probability of a particular outcome, while the latter determines the degree to which a decision is biased in a context-independent manner, which may have been accumulated following repeated performance of those same decisions over time. Outside of the Stroop task, similar parameters have been estimated in the context of motivated decision making tasks, including studies of substance abuse disorders [70], pharmacological studies of serotonergic function [71], and even in saccadic exploration tasks [72].

Although the mechanisms corresponding to the hypotheses cited above can both be seen as instances of *mental action* – i.e., precision modulation from the point of view of active inference [73, 74] – they differ arguably in the neural locations where precision changes would be respectively implemented. In a previous fMRI study with a similar experimental design, Khachouf et al. [31] observed significant activity changes triggered by the instructional cue to apply intentional effort to the Stroop task in a wide mosaic of brain regions, including areas belonging to the salience network (anterior/middle cingulate and anterior insula cortex), to the fronto-parietal attentional network (superior parietal cortex, supplementary and pre-supplementary motor area, frontal eye fields and superior frontal gyrus, dorsolateral prefrontal cortex), to the corpus striatum of the basal ganglia, and to the midbrain arousal system. Research on the neural bases of motivational processes has consistently implicated the circuits supporting salience detection, attentional control and reward [75, 76], with a particular focus on dopaminergic transmission [77, 78]. In the clinic, the emergence of apathy in neurological conditions – especially in Parkinson's and

Alzheimer's diseases, but also in stroke – has been associated with functional impairment and anatomical atrophy in many of the same regions, in particular the medial frontal cortex and the striatum [79, 80]. Parr et al. [58] proposed a tentative mapping of the model's architecture onto a subset of the brain regions listed above – see Figure 4 in the cited reference. Also, a recent study using a transcranial stimulation protocol during an N-back task, demonstrated the causal role of dorsolateral prefrontal cortex in motivating the engagement of effortful cognitive control [81].

On this basis, it may be reasonable to hypothesize that the observed difference in the estimated values of the c parameter between the EXERT and RELAX conditions reflects a process of precision weighting of the connections among the midbrain, dorsomedial striatum, prefrontal cortex, anterior cingulate and insular cortex, primarily deployed through the neuromodulatory action of catecholamines. Habit-driven, context-independent behavior has been linked to the activity of the dorsolateral striatum – the posterior putamen in humans [82] – within a circuit including sensorimotor and premotor cortices. On the other hand, goal-directed, context-dependent behavior has been associated with the dorsomedial striatum – mainly the caudate nucleus – which is connected to lateral, medial, and orbital prefrontal regions, cingulate cortex, and other associative areas (see, e.g., [83, 84, 85]). In [31], the intentional engagement of effort in a Stroop task was associated with a markedly increased activation of the dorsomedial striatum along with other regions mentioned above but, notably, no significant change of activation was observed in the dorsolateral striatum. Although this aligns with the present findings of an increase of the motivation-related parameter c – rather than the habit-related parameter e – driven by intentional effort, the mapping of the observed effect onto specific neural circuits remains speculative at this stage and will have to be verified by future imaging studies with targeted functional connectivity analyses.

Several factors motivated the decision to model our data using an active inference approach. First, this framework is particularly well suited for studies with limited sample sizes, especially in clinical populations where recruitment may be challenging. This is because it explicitly captures uncertainty in parameter estimates, helping to quantify whether additional data are needed, but also because it uses the full sequence of behavioral measurements for each participant, rather than having to rely on summaries like overall accuracy and average response times. Furthermore, each individual can be well characterized in terms of precise individual estimates, improving the inferences at the group level if between-subject variability is not excessive. Second, active inference allows for the estimation of model parameters representing causal factors that are not directly observable but are meaningfully interpretable, thus making it possible to explicitly test specific hypotheses about behavior and decision-making that depend on such hidden factors. Third, the active inference framework naturally accounts for two sources of variability in behavior: the normal trial-to-trial variation (the choices are sampled from a probability distribution) and a more general random variability implemented via the inverse temperature parameter λ; this latter feature enables the model to capture phenomena such as distraction, which can lead participants to make errors that do not depend on motivation or task demands. Finally, our operational definition of effort in terms of the divergence of context-dependent actions from context-independent habitual policies – effectively a cost functional – is consistent with current characterizations of effort as a cost-benefit decision process within a neuroeconomic theoretical framework [45, 86, 87].

In summary, we have demonstrated the feasibility of a modeling approach based on

active inference to the estimation of parameters related to effort and motivation from behavioral data collected during a Stroop task. While this is one of the few studies to date in which active inference with model inversion has been applied to recover parameter estimates from actual observed data – and in particular, with the use of a deep temporal model – it is important to recognize some limitations. First, due to the use of a two-level generative model and the collection of hundreds of trials per participant, the computational time required for model inversion is substantial, even when utilizing a high-performance computing system. Second, the model was tested only on healthy participants, thus its applicability to patient populations – an extension we consider both promising and relevant – will need to be separately assessed and may require modifications to the model to accurately reproduce behavior. Third, in our study, the psychological meaning of the c parameter could not be directly confirmed by specific first-person ratings targeting motivation (which we did not collect), nor was motivation independently manipulated in the experimental paradigm (e.g., via different amounts of monetary rewards), thus our interpretation will need to be verified by future studies. Fourth, the employed experimental setup did not include a condition with a 'natural' (i.e., uninstructed) level of applied effort, thus both the EXERT and the RELAX conditions may represent a deviation from the natural level of effort investment. Fifth, since we did not include text stimuli without semantic content (e.g., the string 'XXXX' in colored fonts), we were not able to assess whether the manipulation of intentional effort influenced response facilitation (better performance on congruent trials), interference (worse performance on incongruent trials), or both.

It is also important to note that, while we found evidence for an association between the voluntary deployment of cognitive effort, our model was applied only to a specific cognitive task. Although the Stroop paradigm has often been used to study cognitive effort, the latter can involve various control mechanisms in different tasks or contexts – e.g., suppression of prepotent responses, increased attentional allocation to targets, enhanced suppression of distractors [88] – thus our findings may not be directly generalizable to all these different scenarios. In thinking about whether, and to what extent, the findings here generalize to other settings and tasks, consider if we had asked participants to specifically exert themselves to perform the task quickly. Here, we might expect to have to define preferences – and habits – over the timings of their responses. While such questions are vitally important in effort research, where speed-accuracy trade-offs are key measures of the deployment of effort, this appears to be a different sort of instruction. In principle, one could model this paradigm and might see a decrease in accuracy as the preference for faster responses increases (or as habitual biases favoring slower, more deliberative, responses are suppressed). Furthermore, the choice of the model and its parameters was informed by the general scheme of the active inference framework, leading to the working definition of mental effort as the divergence between habitual actions and context-dependent policies. This is a rather abstract definition and does not get into the details (and, consequently, nor does the model) of how the high-level constructs of motivation and habit exert a causal influence on specific control mechanisms. Indeed, there is evidence for a complex role of cognitive control in the Stroop task through a variety of potential mechanisms [89, 90, 91], but investigating this was beyond the scope of our study. To some extent, our model is also agnostic about the mechanisms of deployment of voluntary effort. One could, in principle, propose a further higher level for the model at which decisions about modifications of either automatic response suppression or preference enhancement are made. In other words, these results do not preclude the narrative

that "I have a preference for deploying more exertion because I have been asked to, and in doing so will enhance my preferences to suppress the automatic response". While this is an interesting direction for future research, it would likely be impractical to efficiently fit models of this size to data to answer these questions.

Despite the limitations, this work opens up significant opportunities for future research. As mentioned above, it can be potentially extended to clinical populations to investigate how conditions such as fatigue or psychological burnout are related to motivation and effortful engagement, which may have both important diagnostic and therapeutic implications. Furthermore, the computational framework developed in this study could be adapted to other neuropsychological tasks – e.g., the emotional Stroop [92, 93] – facilitating the development of normative models to support neurologists in the diagnostic pathway.

Chapter 3

Mathematical approaches for clinical research in neuroscience

This chapter presents some mathematical approaches for clinical problems. These works have been carried out within the European project *UnaWireD*, funded by the European Research Council, under the supervision of Prof. Giovanna Zamboni and in close interdisciplinary collaboration with neurologists and neuroscientists. Throughout the three years of my doctoral studies, my activities have covered the design, the collection, and the analysis of behavioral and neuroimaging data, as well as the development of computational models to explain the mechanisms underlying cognitive impairments.

3.1 Age-specific prevalence of the different clinical presentations of Alzheimer's disease and frontotemporal dementia in young-onset dementia

The first work presented refers to a project on the age-specific prevalence of Alzheimer's disease (AD) and frontotemporal dementia (FTD). My main contribution was the development and implementation of mathematical models used to estimate the disease prevalence across age groups. Specifically, I worked on the formulation and application of generalized logistic regression functions to characterize the trajectories of both typical and atypical clinical presentations of AD and FTD. This contribution was central to the results reported in [94].

3.1.1 Introduction

Of all the causes of late-onset dementia, AD is certainly the most common one, accounting for 70% of the cases [95]. The prevalence of AD increases with age across the entire age spectrum [96, 97] including ages before 65 [98, 99, 100]. However, it is not known whether the age-related increase in the prevalence of AD concerns only the most frequent amnestic presentation, or also affects atypical presentations such as the logopenic variant of primary progressive aphasia (lvPPA; [101]), posterior cortical atrophy (PCA; [102]), and the behavioral-dysexecutive variant (be-dAD; [103]).

In young-onset dementia (YOD; dementia occurring before the age of 65), some studies have reported similar prevalence for AD and FTD [104]. Epidemiological work focusing on YOD suggested that FTD prevalence peaks between 55 and 59 years [98, 99, 105,

106], whereas other studies indicated that prevalence may continue to rise at older ages [100, 107]. As with AD, it is unknown whether age-related trends are shared across FTD clinical presentations or primarily driven by its most common variant. Here, we focus on the three FTD syndromes: the behavioral variant (bvFTD) and the two language variants, nonfluent (nfPPA; [101]) and semantic (svPPA; [101]).

The purpose of the research presented here is to study the prevalence by age of onset of different presentations of young-onset AD and FTD in a population-based study. Focusing on YOD allowed us to have a clearer classification of different diagnosis, as co-pathology is less frequent than in older-onset dementia [108].

3.1.2 Methods

Data collection

Young-onset AD and FTD cases were identified in a previous study conducted in the province of Modena, Northern Italy, involving all the residents alive on census day (June 30th, 2019) who had received a diagnosis of YOD [109]. Exclusion criteria were: coexisting diagnoses of developmental disorders, longstanding history of schizophrenia and bipolar disorder, cognitive impairment in the context of a neurological disorder in which non-cognitive symptoms were the most disabling, age younger than 30, and residence outside the province of Modena on census day. Diagnoses were based on neurological examination, clinical history, neuropsychological assessment, brain magnetic resonance imaging, and also supported, when indicated, by measurement of cerebrospinal fluid and amyloid PET imaging biomarkers. The study was conducted in accordance with local clinical research regulations and conformed to the Declaration of Helsinki. All subjects provided informed written consent (Study Number 186/2016, approved by the ethical committee Area Vasta Emilia Nord, Italy).

Age-specific prevalence and cumulative occurrence

Age-specific prevalence (i.e., the prevalence stratified according to 1-year age groups) was calculated using as denominator the Modena province residing population on January 1st, 2019. The age-specific prevalence function

$$P(n) = \frac{\text{number of cases with symptoms onset } \mathbf{at} \text{ age } n}{\text{population of age } n}, \quad n \in \mathbb{N}$$

tended to increase with advancing age but showed fluctuations because of the small age groups. Therefore, to obtain more robustness, we defined the *cumulative occurrence* function:

$$F(n) = \text{number of cases with symptoms onset } \mathbf{by} \text{ age } n, \quad n \in \mathbb{N}$$

This function is non-decreasing and relates to age-specific prevalence as follows:

$$P(n) = \frac{F(n) - F(n-1)}{\text{population of age } n}. \tag{3.1}$$

It follows that age-specific prevalence can be estimated using a regression model on cumulative occurrence and then applying Equation 3.1.

Regression model

Fitting a function to a sample of points requires first a selection of the family of candidate functions. The ideal model should accommodate the variations of the true curve with the smallest number of parameters [110]. In our case, the *generalized logistic function* [111] provides a compromise between over-parameterized models that can fit data closely at the cost of a large variance in the predictions, and under-parameterized models that suffer from large lack-of-fit error. This function accommodates the sigmoidal shape expected for cumulative occurrence over age:

$$F(n) \approx D + \frac{A - D}{\left[1 + (n/C)^B\right]^E} \tag{3.2}$$

where the integer n represents the age in the interval 30–64 years. Figure 3.1 represents the trajectory of the generalized logistic function for different values of A, B, C, D, E. In particular, A controls the inferior asymptote. As by definition the cumulative occurrence must start at zero, and no cases were observed before age 42 in our sample, here A was constrained to zero.

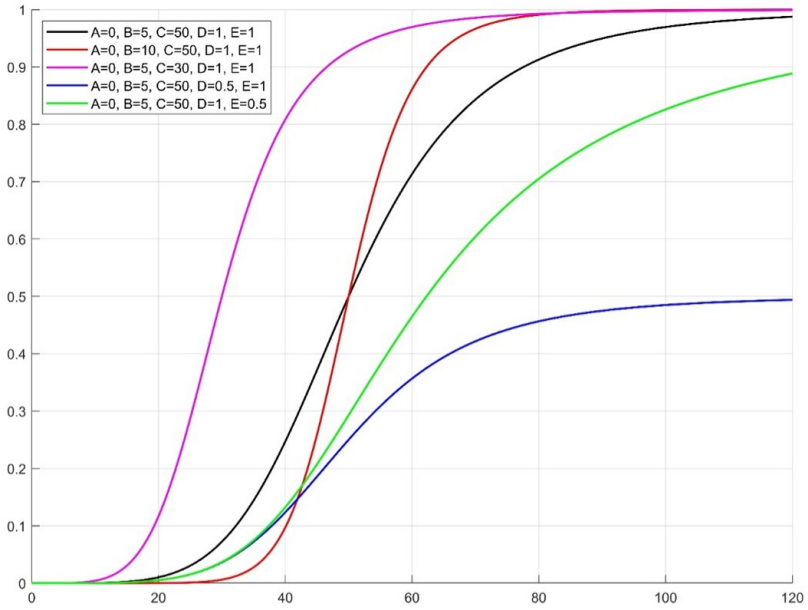

Figure 3.1. Effect of varying parameters on generalized logistic function. A controls the inferior asymptote. B refers to the steepness of the curve. C is related to the inflection point. D controls the superior asymptote. E is the asymmetry factor (when $E = 1$ the curve is symmetric around the inflection point).

The best fit was selected by minimizing the residual sum of squares. Having fixed $A = 0$, a four-parameter regression was obtained. The algorithm was implemented in MATLAB, using the function `fit` and adapting the `LP5` script of Cardillo [112]. The procedure was repeated for each of the following diagnostic groups: *amnestic AD*, *atypical AD* (lvPPA, PCA, and be-dAD), *bvFTD*, and *aphasic FTD* (svPPA and nfPPA). Age-specific prevalence of all-variants AD was calculated by summing up the estimates of age-specific prevalence of amnestic AD and atypical AD. Age-specific prevalence of FTD was calculated by summing up the estimates of age-specific prevalence of bvFTD and aphasic FTD.

3.1.3 Results

We identified 172 patients with a diagnosis of young-onset AD or FTD on census day. Table 3.1 reports the number and rate of cases for each clinical presentation, stratified by 5-year age group. The age-specific prevalence is plotted in Figure 3.2. Its oscillating behavior justified the use of cumulative occurrence for fitting purposes.

Table 3.1. Total cases and crude prevalence per 1,000,000 inhabitants (in brackets) of the different clinical presentations of young-onset AD and FTD.

Age range	**40–44**	**45–49**	**50–54**	**55–59**	**60–64**
			Alzheimer's disease (AD)		
All AD cases	1 (18)	2 (34)	14 (245)	31 (607)	58 (1322)
Amnestic AD	0 (0)	0 (0)	10 (175)	20 (392)	44 (1003)
lvPPA	0 (0)	2 (34)	2 (35)	5 (98)	10 (203)
PCA	1 (19)	0 (0)	2 (35)	1 (20)	4 (91)
be-dAD	0 (0)	0 (0)	0 (0)	5 (98)	0 (0)
			Frontotemporal dementia (FTD)		
All FTD cases	1 (18)	3 (51)	9 (158)	20 (392)	33 (752)
bvFTD	1 (18)	1 (17)	6 (105)	16 (313)	27 (616)
svPPA	0 (0)	2 (34)	1 (18)	2 (39)	6 (137)
nfPPA	0 (0)	0 (0)	2 (35)	2 (39)	0 (0)
Residents in Modena province	54,241	59,048	57,072	51,060	43,860

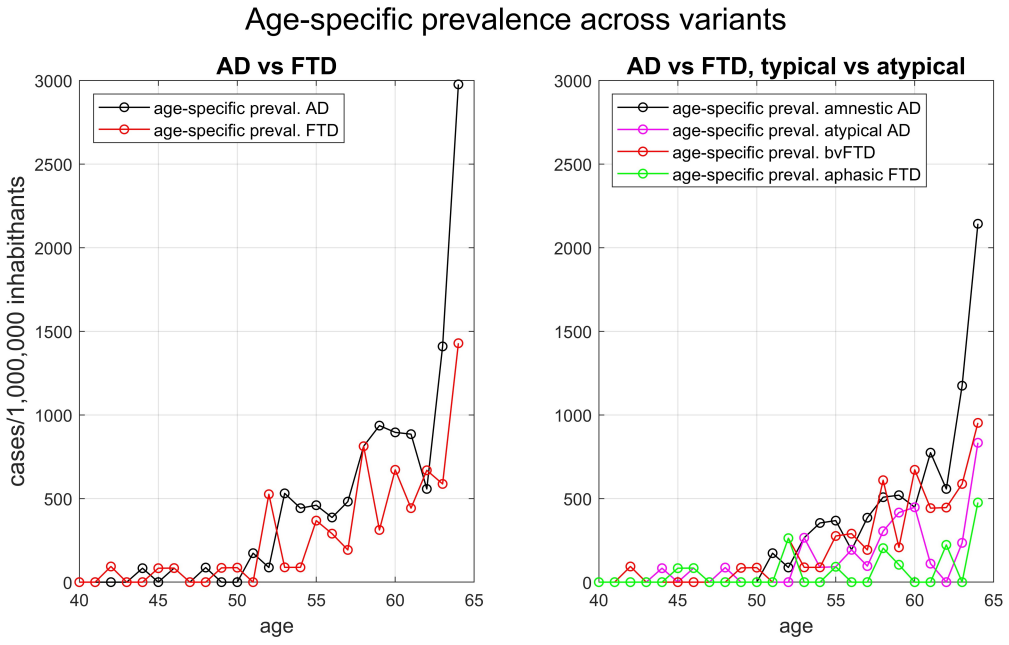

Figure 3.2. Age-specific prevalence across clinical presentations. The left panel shows all-variants AD and all-variants FTD. The right panel shows amnestic AD, atypical AD, behavioral FTD, and aphasic FTD.

In the regression models, all adjusted R^2_{adj} were above 0.96, indicating that all models

captured well the cumulative distributions. The estimated *normalized cumulative occurrence* functions[1] and their derivatives are represented in Figure 3.3.

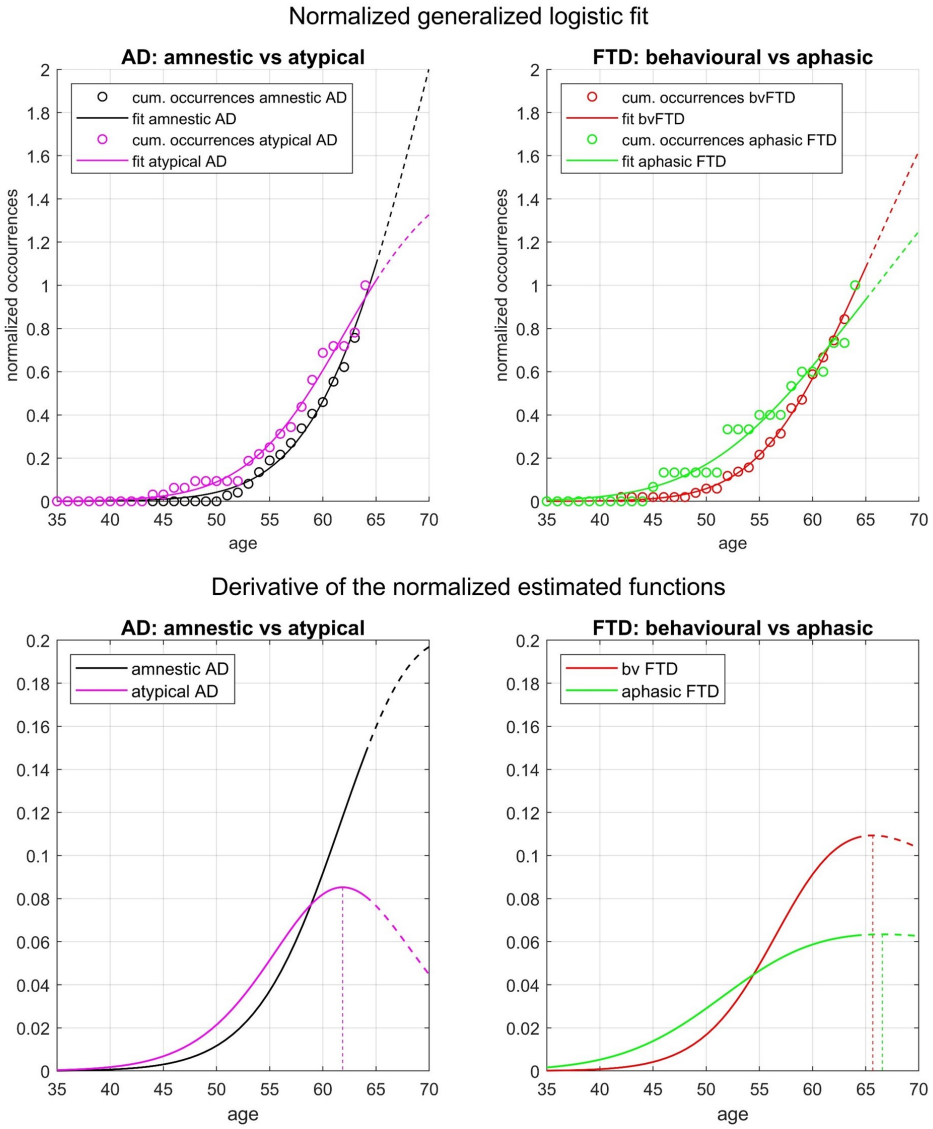

Figure 3.3. Top panels show the fit of regression models. The points indicate normalized cumulative occurrences, while the curves represent the estimates. These graphs are meant to compare the steepness of the curves, not the values on the y-axis, which were scaled to 1. Bottom panels indicate the first derivatives of the estimated functions. Note that the inflection point of a function corresponds to the maximum of its first derivative, thus the peaks, marked with vertical lines, can be interpreted as the age when the cumulative occurrence function slows its growth. For displaying purposes only, curves are extended beyond the age of 64 (dotted lines)

Figure 3.4 shows the estimated age-specific prevalence by age of onset obtained using Equation 3.1. The slope of this function increased with advancing age in both AD and FTD in early years, but the two curves started to separate after 50 years of age, with

[1]The normalized cumulative occurrence of x is the percentage of people with a specific clinical presentation who had developed symptoms by age x, among those with that clinical presentation of YOD.

AD increasing much more than FTD. This difference was mainly driven by amnestic AD. Indeed, the estimated age-specific prevalence functions of all the clinical presentations increased with advancing age in early ages, but the slope of such increase started to flatten after the age of 60 for all the clinical presentations except for amnestic AD. Atypical AD and aphasic FTD were the ones for which the estimated age-specific prevalence slowed the most after 60. More precisely, the delta for age-specific prevalence between age 60 and age 64 was +75.3% for amnestic AD (from 706 cases/1,000,000 to 1238/1,000,000), +30.2% for bvFTD (from 500/1,000,000 to 651/1,000,000), +17.3% for aphasic FTD (from 98/1,000,000 to 116/1,000,000), and +9.8% for atypical AD (from 287/1,000,000 to 315/1,000,000).

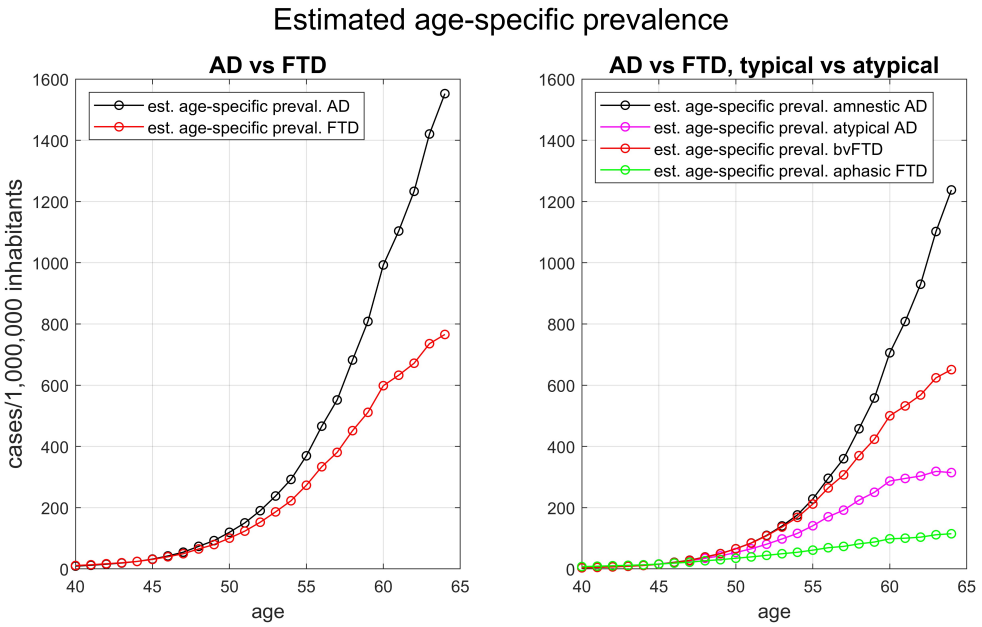

Figure 3.4. Estimated age-specific prevalence functions. Left panel compares all-variants AD and all-variants FTD, while the right panel compares amnestic AD, atypical AD, bvFTD, and aphasic FTD.

3.1.4 Discussion

The study presented in this section aimed to establish whether the prevalence of different clinical presentations of young-onset AD and FTD increases in the same way with age. This difference was clearly shown by their estimated age-specific prevalence, which were almost overlapping in the fifth decade, progressively separated in the sixth decade, and showed a sharp increase of AD in the first half of the seventh decade.

When looking at the different clinical presentations of AD and FTD, we found that the amnestic variant of AD increased disproportionally in age-specific prevalence compared to atypical AD (including PCA, lvPPA, and be-dAD) and all the presentations of FTD. Atypical AD and aphasic FTD were the clinical presentations for which the estimated age-specific prevalence grew the least and almost stabilized after the age of 60, suggesting that the most frequent clinical presentations are also those whose prevalence increases the most with advancing age. This pattern suggests that normal aging, with its impact

on hippocampal and frontal structures, amplifies the probability of memory-related presentations, whereas atypical variants are more likely to emerge earlier when additional neurodevelopmental vulnerabilities are present [108, 113, 114].

The study presented here has limitations. First, the number of prevalent cases was modest and may have biased the identification of associations between age and the less frequent presentations of AD or FTD. However, we deem this to be unlikely, as the atypical presentations of AD and aphasic presentations of FTD were pooled together to increase group sample sizes. Second, the analysis was performed on crude occurrences. The decision not to stratify by sex was based on the need to avoid reducing the model's statistical power. Furthermore, there is no consensus on whether YOD is differentially frequent between sexes [115].

In conclusion, with this research we showed that the trajectories of age-specific prevalence differ across clinical presentations of AD and FTD in YOD. With increasing age, atypical AD and aphasic FTD account for a progressively smaller proportion of cases, while bvFTD and, especially, amnestic AD become predominant. These observations support the view that age-related mechanisms may differentially affect the likelihood of specific clinical presentations, contributing to the phenotypical variability observed within neurodegenerative diseases.

3.2 Age-adjusted neurofilament light chain cutoffs for neurodegenerative dementia diagnosis: a machine learning approach

The second study presented in this chapter examines how serum and cerebrospinal fluid (CSF) neurofilament light chain (NfL) levels differ between individuals with and without neurodegenerative disorders, with the aim of developing an age-adjusted cutoff to help clinical diagnosis. My primary contributions included implementing a Bayesian regression to quantify the influence of diagnosis, age, and potential confounders on NfL concentrations, and applying a two-threshold Support Vector Machine (SVM) classifier to estimate the cutoff equations. Results of this work have been submitted for publication and are currently under review.

3.2.1 Introduction

Neurofilament light chain (NfL) proteins represent the most abundant intermediate filament protein expressed in mature neurons, providing structural stability and resistance to mechanical stress [116]. After neuronal and axonal damage, increased quantities of NfLs are released into the CSF and drained into the blood [117, 118]. Among other classes of neurofilaments (heavy and medium chain), NfLs are considered the most promising biomarkers, reflecting neuroaxonal damage in a wide variety of neurological diseases [119, 120]. For this reason, NfLs are considered as non-specific biomarkers of neurodegeneration and axonal damage. In particular, NfLs have considerable interest in the field of neurodegenerative dementias, including Alzheimer's disease (AD), frontotemporal dementia (FTD), and Lewy's body dementia (LBD). Its measurement in this context has proven valuable not only for diagnosis but also for monitoring disease progression, assessing treatment response, and predicting phenoconversion in genetic pathologies [121, 122]. Recent studies have also suggested that an altered permeability of the blood–brain barrier and

blood–CSF barrier can influence blood NfL concentrations, although the exact influence of such barrier permeability remains to be elucidated [123, 124].

Despite its promise, the use of NfL measurement has not been fully implemented into clinical practice yet, mainly because there are no reliable and universally applicable reference ranges to be used. Several confounding factors challenge the identification of cutoff values based solely on absolute NfL concentrations. An annual percentage increase in NfL concentrations of at least 2.2%, with faster rises in older populations is reported in [125]. This age-related increase has been attributed to neuronal loss and increased blood-barrier permeability that occur with aging [117]. Other physiological factors, including renal function, blood sugar levels, sex, and body mass index (BMI), have also been implicated in NfL levels variability, though findings are inconsistent, and their effect size appears smaller than that of age [126, 127, 128, 129]. With this study, we aimed to develop age-adjusted NfL reference values to enhance the clinical use of NfL in the diagnosis of neurodegenerative dementias.

3.2.2 Methods

Data collection

We retrospectively included all individuals who had presented to the Cognitive Neurology Clinic at Modena University Hospital, Northern Italy, due to cognitive and/or behavioral disturbances from 2009 to 2023. Inclusion criteria required participants to have completed a comprehensive baseline diagnostic workup, including neuropsychological assessment, blood test and a brain MRI scan. When clinically indicated, this also included lumbar puncture and amyloid-PET. Moreover, participants had to have subsequently undergone clinical follow-up for at least two years. Individuals younger than 30 years of age at the time of baseline evaluation were excluded. All participants had provided written consent for the storage of CSF and blood samples (Ethics Committee approval no 26/2017) and for their use for research purposes (Ethics Committee approval 372/2021).

Subjects who had eventually received a clinical and biomarker-based diagnosis of AD [130], FTD – including both behavioral and aphasic variants [101, 131], or LBD [132] were included in the Neurodegenerative group (NDG). Subjects whose diagnostic workup had not led to a diagnosis of a neurodegenerative condition and who had not clinically progressed or worsened for at least two years of clinical follow-up were stratified in the Other Etiology group (OE). They included subjects with a Primary Psychiatric Disorder (PPD), cognitive disturbances due to Obstructive Sleep Apnea Syndrome (OSAS) or vitamin deficiency, functional disorders, and iatrogenic disorders. Cognitively unimpaired individuals with available blood and CSF samples were also included as a control group (CTR) and aggregated with OE in the non-Neurodegenerative group (non-NDG) (Figure 3.5).

Demographic and clinical data were collected for all participants, as well as blood creatinine values as renal function indicator. NfL levels were measured in serum and CSF samples. Analyses were conducted at the Neuroimmunology Laboratory of Modena University Hospital using Ella™ (Ella Simple Plex assay technology, BioTechne, ProteinSimple; see also Section 3.3) [133, 134].

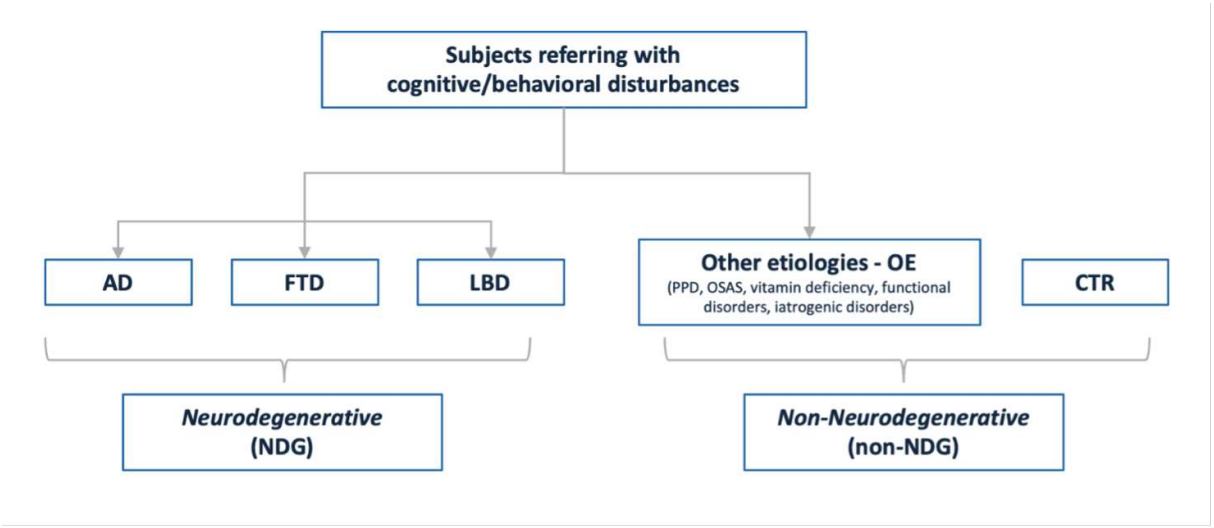

Figure 3.5. Sample characterization. Subjects referred for cognitive and/or behavioral disturbances were first classified according to clinical and biomarker-based diagnosis. Neurodegenerative dementias (i.e., AD, FTD, and LBD) were then classified in the NDG group. Subjects with other non-neurodegenerative etiologies and cognitively unimpaired controls were classified in the non-NDG group.

Comparison of NfL levels across groups

Median concentrations and interquartile ranges (IQR) of demographic and neurobiological variables were calculated for both the NDG and non-NDG group. Serum and CSF NfL levels were log-transformed to achieve normal distributions [135]. The Shapiro-Wilk test ($\alpha = 0.05$) was employed to assess the normality of these transformed distributions.

We examined the effects of age, disease duration, creatinine, gender, and diagnosis on log-transformed NfL levels using a Bayesian regression model. To investigate whether the relationship between diagnosis and NfL levels was moderated by other clinical variables, we included interaction terms between diagnosis and age, diagnosis and disease duration, and diagnosis and creatinine. Separate analyses were conducted for serum and CSF NfL levels. The model was implemented in R using the 'brm' function of the `brms` package [136], employing four Monte Carlo Markov Chains, each consisting of 2000 burn-in iterations followed by 4000 sampling iterations.

Identification of an age-dependent cutoff

A machine learning algorithm was employed to establish a threshold differentiating NDG and non-NDG groups based on age and log-transformed NfL levels. The exclusion of the other variables included in the regression presented in the previous paragraph was motivated by their lack of significant association with NfL levels (see results in Section 3.2.3). A Support Vector Machine (SVM) classification model with a linear kernel and soft-margin regularization was developed using the MATLAB function 'fitcsvm'. SVM classifiers trained on imbalanced datasets may yield suboptimal performance, typically favoring the majority class while underperforming on the minority class [137]. To address this issue, we implemented a weighted SVM approach. More precisely, we used inverse-frequency class weights, assigning each NDG case a weight equal to $\frac{1}{\text{number of NDG}}$ and each non-NDG case a weight equal to $\frac{1}{\text{number of non-NDG}}$, so that both classes contributed

approximately equally to the optimization problem.

To improve overall test accuracy and reduce the occurrence of false positive results, we adopted a two-threshold classification strategy [138]. This approach defines an upper threshold, above which individuals are classified as definitely NDG, and a lower threshold, below which individuals are classified as definitely non-NDG. Participants with assay values falling between these two thresholds were classified as intermediate risk and recommended for further clinical evaluations. Optimal thresholds were determined to maximize specificity while ensuring a sensitivity of at least 85% and limiting the proportion of intermediate results to less than 25% [139]. Model performance was evaluated using leave-one-out cross-validation (LOO-CV). For each participant, a weighted SVM model was trained on the remaining participants. Then, we applied the two-threshold rule defined by that model to classify the left-out patient. This yielded an out-of-sample predicted label for each individual. Sensitivity, specificity, accuracy, and the proportion of intermediate classifications were computed from these LOO-CV predictions.

Finally, to quantify the uncertainty of our results, we performed a non-parametric bootstrap with 1000 samples. For each bootstrap sample, we re-fitted the weighted SVM and re-calculated the lower and upper CSF NfL thresholds to derive the 90% bootstrap confidence intervals (5th–95th percentile).

3.2.3 Results

Data collection

A total of 217 individuals were recruited. Among them, 158 participants were classified as NDG, including 47 with AD, 106 with FTD, and 5 with LBD. The remaining 59 participants were categorized as non-NDG, comprising 29 OE and 30 CTR.

The median concentrations of CSF and serum NfL in the NDG group were 1432 pg/mL (IQR: 899–2502) and 34.6 pg/mL (23.5–53.5), respectively. In contrast, median concentrations of CSF and serum NfL in the non-NDG were 491 pg/mL (351–703) and 15.0 pg/mL (9.0–22.2). As expected, the median concentrations of both CSF and serum NfL did not differ significantly between CTR and OE groups (see Table 3.2). The Shapiro-Wilk test indicated that log-transformed NfL values were approximately normally distributed for both CSF ($p = 0.07$) and serum ($p = 0.19$).

Comparison of NfL levels across groups

Figure 3.6 presents the results of Bayesian regression models assessing the effects of age and diagnosis on NfL levels. In both biofluids, NfL levels were consistently higher in NDG individuals compared to non-NDG participants. Additionally, NfL levels exhibited a strong dependence by age, indicating that it is not possible to define an age-independent cutoff between the two groups. The impact of disease duration, creatinine, and gender was found to be not relevant in both models (see also Table 3.3 for more details).

As illustrated in Figure 3.6, the 90% credible interval for mean CSF NfL levels show no overlap across most of the age spectrum. In contrast, for serum NfL, credible intervals begin to overlap around age 70, indicating increased ambiguity at older ages. Consequently, within the current dataset, an age-dependent cutoff can be reliably established across the entire age range for CSF NfL but not for serum NfL. Therefore, the following analyses will focus only on CSF.

Table 3.2. Descriptive statistics of demographic and neurobiological variables. Data are reported as median score and interquartile range (IQR; in brackets).

	NDG	**AD**	**FTD**	**LBD**
num. of patients (males)	158 (80)	47 (18)	106 (59)	5 (3)
age (years)	65.1 (61.2 – 71.6)	63.5 (58.0 – 66.7)	66.0 (63.1 – 72.8)	64.4 (61.3 – 67.8)
disease duration (years)	2.8 (1.3 – 4.3)	3.0 (1.1 – 4.5)	2.8 (1.4 – 4.3)	1.2 (1.0 – 1.8)
creatinine (mg/dL)	0.80 (0.72 – 0.91)	0.80 (0.72 – 0.90)	0.80 (0.71 – 0.93)	0.83 (0.78 – 0.87)
CSF NfL (pg/mL)	1432 (899 – 2502)	1364 (888 – 1822)	1486 (926 – 3399)	1004 (833 – 1206)
serum NfL (pg/mL)	34.6 (23.5 – 53.5)	27.0 (22.8 – 41.2)	41.5 (24.0 – 58.9)	23.0 (22.0 – 25.0)

	non-NDG	**OE**	**CTR**
num. of patients (males)	59 (22)	29 (14)	30 (8)
age (years)	58.6 (46.7 – 65.3)	59.9 (53.6 – 65.1)	55.8 (44.3 – 66.7)
disease duration (years)	0.0 (0.0 – 1.8)	2.4 (1.0 – 3.9)	0.0 (0.0 – 0.0)
creatinine (mg/dL)	0.74 (0.68 – 0.84)	0.76 (0.68 – 0.86)	0.73 (0.68 – 0.81)
CSF NfL (pg/mL)	491 (351 – 703)	507 (400 – 717)	471 (337 – 682)
serum NfL (pg/mL)	15.0 (9.0 – 22.2)	16.0 (11.6 – 21.6)	14.3 (6.6 – 24.7)

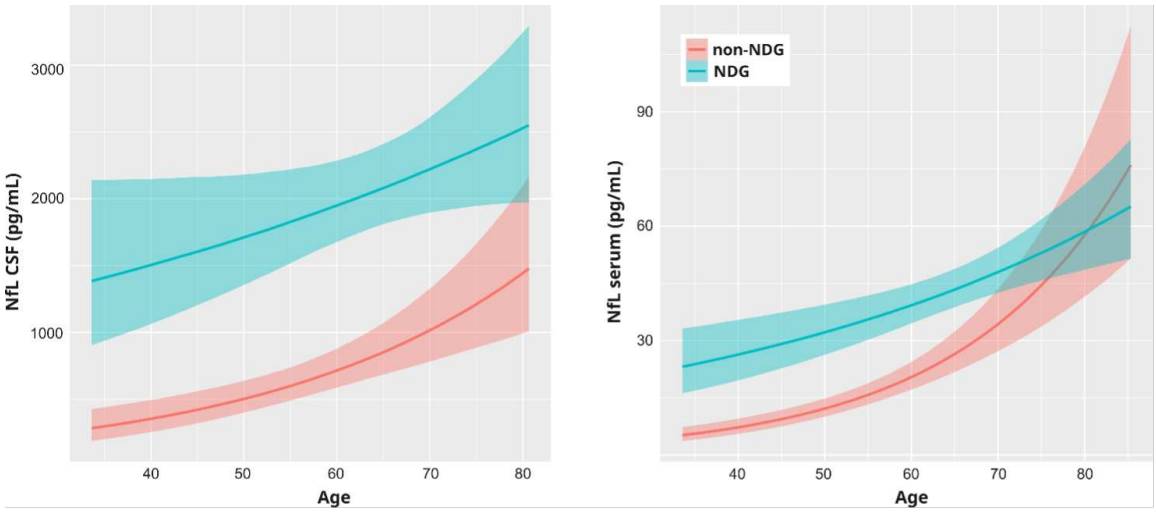

Figure 3.6. Estimated NfL concentrations in CSF (left) and serum (right) as a function of age and diagnosis. Solid lines indicate the expected mean levels, while the shaded regions represent the 90% credible intervals.

Identification of an age-dependent cutoff

To enhance overall test accuracy and reduce the incidence of false positive results compared to a single-threshold approach, we applied a two-threshold classification framework to NfL, implemented using a leave-one-out cross-validated SVM model. If y represents CSF NfL (pg/mL) and x the age in years, the lower cutpoint hyperplane was found to have the equation:

$$y = \exp(0.003584x + 6.3093) \tag{3.3}$$

while the upper one had the equation:

$$y = \exp(0.003557x + 6.7166) \tag{3.4}$$

Table 3.3. Estimates and 95% credible intervals for the parameters of the Bayesian regression model. The dependent variable was NfL (CSF or serum), while the predictors included age, diagnosis, disease duration, creatinine, gender, as well as interaction terms included to account for potential differential effects of diagnosis on age, disease duration, and creatinine. Variables whose credible intervals do not contain zero are denoted in bold.

	CSF (pg/mL)	**serum (pg/mL)**
Effect of diagnosis	**1.79 [0.02, 3.60]**	**1.80 [0.37, 3.21]**
Effect of age on non-NDG	**0.03 [0.02, 0.05]**	**0.04 [0.03, 0.06]**
Effect of age on NDG	**0.02 [0.01, 0.03]**	**0.02 [0.01, 0.03]**
Effect of age	**0.02 [0.01, 0.03]**	**0.03 [0.02, 0.04]**
Effect of disease duration on non-NDG	0.00 [-0.08, 0.08]	0.02 [-0.05, 0.09]
Effect of disease duration on NDG	0.02 [-0.03, 0.06]	0.02 [-0.02, 0.06]
Effect of disease duration	0.01 [-0.04, 0.05]	0.02 [-0.02, 0.06]
Effect of creatinine on non-NDG	-0.24 [-1.45, 0.98]	0.09 [-0.91, 1.11]
Effect of creatinine on NDG	-0.13 [1.05, 0.79]	0.20 [-0.38, 0.77]
Effect of creatinine	-0.18 [-0.99, 0.62]	0.14 [-0.47, 0.75]
Effect of female gender	-0.05 [-0.31, 0.23]	0.11 [-0.12, 0.33]

Participants with CSF NfL levels below the lower threshold can be classified as non-NDG, while those with levels exceeding the upper threshold are classified as NDG. The model achieved a specificity of 92.8% and a sensitivity of 85.6%, while restricting the proportion of participants classified in the intermediate zone to 22.5%. Table 3.4 and Figure 3.7 illustrate the variation of the two thresholds across the age spectrum.

Table 3.4. Age-dependent CSF NfL cutpoints for selected age categories. Values are expressed in pg/mL, with 90% bootstrap confidence intervals shown in parentheses. Participants with CSF NfL levels below the lower cutpoint are classified as non-NDG, whereas those with values above the upper cutpoint are classified as NDG; values falling between the two cutpoints define an intermediate zone of diagnostic uncertainty.

Age	Lower cutpoint (pg/mL)	Upper cutpoint (pg/mL)
40	635 (464 – 844)	952 (685 – 1319)
45	646 (510 – 799)	969 (753 – 1252)
50	657 (558 – 759)	987 (823 – 1200)
55	669 (598 – 738)	1005 (884 – 1163)
60	681 (599 – 761)	1023 (912 – 1168)
65	694 (574 – 800)	1041 (895 – 1229)
70	706 (545 – 858)	1060 (860 – 1320)
75	719 (515 – 933)	1079 (817 – 1421)
80	732 (487 – 1021)	1098 (771 – 1556)

3.2.4 Discussion

The study presented in this section is a comparison of NfL levels measured in CSF and serum between individuals with and without neurodegenerative conditions, with the aim of establishing an age-adjusted reference framework to support the identification of neurodegenerative pathologies in clinical settings. As expected, NfL levels were significantly higher in neurodegenerative dementias compared to both cognitively unimpaired controls

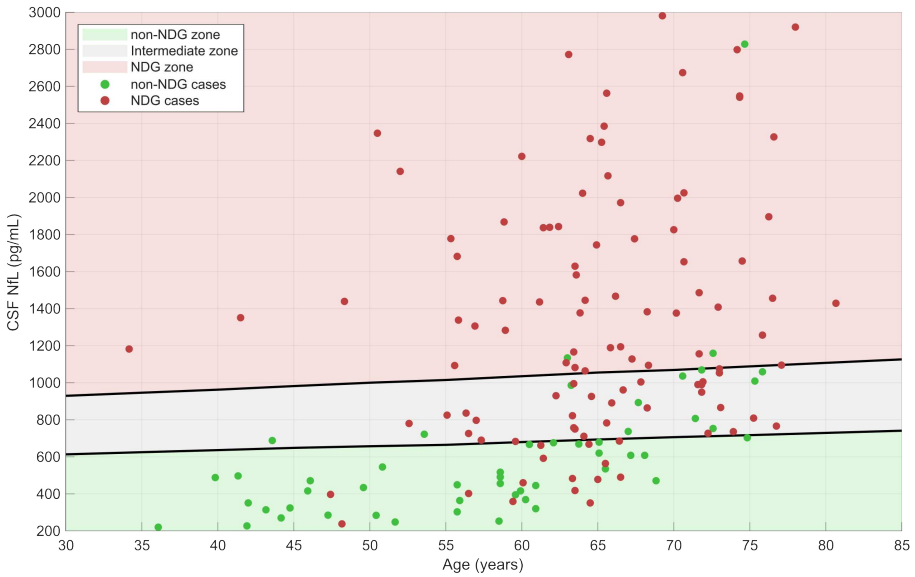

Figure 3.7. Two-threshold SVM classifier. Participants falling within the green region are classified as non-NDG, while those in the red region are classified as NDG. The gray intermediate region indicates an area of uncertainty. Colored dots represent participants from the study cohort: red dots correspond to NDG participants, and green ones represent non-NDG participants. For visual clarity, the y-axis scale is capped at 3000 pg/mL; however, several participants exhibited CSF NfL levels exceeding this value, all of whom belonged to the NDG group.

and individuals presenting with cognitive or behavioral disturbances but without evidence of an underlying neurodegenerative disease. This finding reinforces the role of NfL as a non-specific biomarker of neurodegeneration, supporting its potential use as a first-line diagnostic tool to identify individuals at risk. Moreover, we showed that both CSF and serum NfL levels did not differ significantly between CTR and OE groups, supporting the decision to aggregate them within the non-NDG group. Rather than being a limitation, this approach might more accurately reflect the clinical scenario, in which NfL measurement is typically applied to subjects with some neurological complaints, rather than to strictly healthy individuals.

When evaluating potential confounders, NfL concentrations increased significantly with advancing age in both groups and biofluids. This finding highlights the limitations of fixed, age-independent thresholds. Our model demonstrated that CSF NfL levels remained significantly different between individuals with and without a neurodegenerative condition across the entire age range. Conversely, serum NfL levels exhibited a substantial overlap between groups beyond the age of 70, mainly due to a steeper age-related increase among non-NDG individuals. This finding aligns with previous studies reporting a stronger effect of age on serum NfL and a greater inter-individual variability in its level among older individuals [118, 117]. These patterns have been attributed to both physiological and pathological factors, including a higher prevalence of co-pathologies in older adults [122, 125, 126]. For these reasons, we identified our age-dependent cutoff exclusively for CSF. Replicating our approach in a larger cohort will help to determine whether the absence of a significant difference in serum NfL levels among older individuals with and without a neurodegenerative condition is real or a mere consequence of insufficient amount of data. Should the overlap be confirmed, the use of serum NfL in older

age groups may be limited, suggesting that reliable application of NfL as a blood-based biomarker might be restricted to younger individuals.

From a clinical point of view, the two-threshold approach allows balancing diagnostic accuracy while ensuring that cases with borderline values are appropriately flagged for further clinical evaluation rather than being prematurely assigned to a diagnostic category. This approach aligns with real-world clinical decision-making where biomarker interpretation often follows a probabilistic rather than a strictly binary process. The use of intermediate ranges with two cutpoints has been successfully applied to other medical conditions such as prehypertension and prediabetes [140]. A two cutoff approach has also been applied to AD-specific biomarkers, including blood-based markers and amyloid-PET, demonstrating its effectiveness [138, 139].

The study has limitations. First, the non-NDG group is smaller and younger than the NDG group. However, to prevent the model from favoring the larger group in classification we implemented a weighted classification approach, which assigns greater importance to the minority class, improving classification balance. Second, although we included key confounding factors such as sex, renal function, and disease duration, we lacked data on other physiological (e.g., BMI) and pathological (e.g., diabetes) variables. These factors have been variably associated with NfL levels; however, it is plausible to hypothesize that their contribution is largely mediated by age, given their higher prevalence in older individuals.

In conclusion, this study proves the feasibility and advantages of an age-adaptive classification model for CSF NfL. Future research should focus on validating our approach in larger, independent cohorts and exploring its applicability to blood-based NfL measurements, with the goal of optimizing its utility across different clinical contexts and age groups.

3.3 Comparison of neurofilaments light chain concentrations measured by Ella^{TM} and Lumipulse^{TM} in patients with cognitive impairment

The third study presented in this chapter refers to a comparative analysis of neurofilament light chain (NfL) measurement in cerebrospinal fluid (CSF) and serum samples from patients with cognitive disorders. Two advanced technologies were compared, specifically the Ella^{TM} microfluidic platform and the Lumipulse^{TM} fully automated system. My main contribution includes the development and implementation of statistical models such as the Passing–Bablok regression. Results of this study are reported in [141].

3.3.1 Introduction

In the past three decades, ultrasensitive immunoassay technologies have been developed with the purpose of making biomarkers more clinically useful. With these improvements, NfLs have begun to be measured in blood, with studies reporting high correlations between serum and CSF levels [117, 142, 143]. The gold standard for NfL measurement in terms of sensitivity is the single-molecule array (SIMOA^{TM}) digital immunoassay technology from Quanterix. However, the assay cost often represents a limitation to its availability in clinical laboratories. A recently introduced and cheaper ultrasensitive assay based on

a microfluidic enzyme-linked immunosorbent assay (ELISA) system is Ella™ from Bio-Techne [133, 134]. Ella™ performs immunoassays in a microfluidic cartridge, measuring up to 72 samples in triplicate inside glass nanoreactors using a fluorescent substrate. Another recent platform is the assay based on the Lumipulse™ platform from Fujirebio. It is a fully automated chemiluminescent enzyme immunoassay (CLEIA), allowing for fully automated processing of samples.

3.3.2 Methods

Data collection

Thirty patients seen at the Cognitive Neurology Clinic of Modena University Hospital, Northern Italy, who had undergone blood testing and lumbar puncture as part of their diagnostic workup and had been clinically followed-up for two years, were recruited. They were included in the present study if they had one of the following clinical diagnoses: probable AD supported by abnormal CSF tau and amyloid biomarkers [130], FTD including its behavioral and aphasic variants [101, 131], and non-progressive mild cognitive impairment. The latter diagnostic group included subjects that had initially received a clinical diagnosis of Mild Cognitive Impairment (MCI) [144] or Mild Neurocognitive Disorder [145] but were found to have normal CSF biomarkers, normal structural magnetic resonance imaging of the brain and remained clinically stable for at least 2 years. All individuals had provided written informed consent for the use of CSF and serum samples and personal data for both diagnostic and research purposes before sample collection (Ethics Committee approvals no 26/2017, 372/2021).

CSF and serum NfL concentrations were assessed using two platforms: the Ella™ microfluidic platform (Bio-Techne, Minneapolis, USA) and the Lumipulse™ fully automated system for the CLEIA (Fujirebio Inc., Ghent, Belgium). Standard International Procedures for biobanking were followed [146].

Statistical analysis

Statistical analyses were performed using MATLAB. Median concentrations and interquartile ranges (IQR) of demographic and neurobiological variables were calculated. The Shapiro–Wilk test was used to check the normality of the distribution of NfL concentrations in CSF and serum for both assays. As none of the variables were found to be normally distributed, Spearman correlations were used to test the association between serum and CSF NfL concentrations in the same sample series.

Given the non-normal distribution of data, we used the non-parametric Passing–Bablok regression model to estimate a equation that explains the relationship between the two assays, allowing for the estimation of the values with one method when only the other is available and vice versa [147].

3.3.3 Results

The characteristics of the participants included in the study are reported in Table 3.5. Median concentrations of CSF and serum NfL were, respectively, 879 pg/mL (IQR 510–3641) and 21.0 pg/mL (13.3–58.4) with Ella™, and 735 pg/mL (365–2897) and 21.2 pg/mL (15.5–43.6) with Lumipulse™. As for the diagnostic group, NfL concentrations measured

through the two assays in both blood and CSF were higher in FTD compared to AD and non-progressive MCI, as well as in AD compared to non-progressive MCI (Table 3.5).

Table 3.5. Median and interquartile range (IQR; in brackets) of age, mini-mental state examination (MMSE) score, disease duration, and NfL concentrations according to platform used and divided by diagnostic subgroup.

	All	**MCI**	**AD**	**FTD**
Number of participants	30	10	10	10
Age (years)	61 (54–66)	51 (43–56)	60 (56–66)	67 (63–68)
MMSE score	27 (23–28)	29 (27–30)	25 (24–27)	23 (19–27)
Disease duration (months)	24 (13–45)	29 (12–57)	14 (11–38)	24 (18–43)
NfL CSF (pg/mL)				
Ella^{TM}	879 (510–3641)	411 (286–510)	879 (690–1166)	6235 (3641–7309)
Lumipulse^{TM}	735 (365–2897)	293 (209–365)	735 (648–859)	4785 (2897–6443)
NfL serum (pg/mL)				
Ella^{TM}	21.0 (13.3–58.4)	11.1 (7.6–17.4)	21.6 (15.8–27.7)	109.0 (62.6–137.0)
Lumipulse^{TM}	21.2 (15.5–43.6)	14.6 (13.7–16.5)	20.5 (16.7–25.5)	63.8 (43.6–91.6)

Table 3.6. Estimates and 95% confidence intervals (in brackets) of the Passing–Bablok coefficients.

	Intercept	**Slope**
NfL CSF model	−45.38 [−118.58, 29.42]	0.8261 [0.7038, 0.9289]
NfL serum model	7.543 [5.828, 8.935]	0.5775 [0.5372, 0.6254]

Spearman correlations showed very strong associations between the two assays ($r = 0.976$, 95% confidence interval [0.950, 0.989] for CSF; $r = 0.923$, 95% confidence interval [0.842-0.964] for serum). However, in the results of Passing–Bablok regression (Table 3.6 and Figure 3.8), the 95% confidence interval of the slope did not include 1, indicating a significant proportional difference between the two methods. Consequently, the application of a correction coefficient seems to be necessary. Conversion equations to estimate Ella^{TM} when only Lumipulse^{TM} is available are:

$$\begin{aligned} \text{Ella}^{TM} &= 0.8261 \times \text{Lumipulse}^{TM} - 45.38 \quad \text{for CSF} \\ \text{Ella}^{TM} &= 0.5775 \times \text{Lumipulse}^{TM} + 7.543 \quad \text{for serum} \end{aligned} \tag{3.5}$$

On the other hand, when only Ella^{TM} is available, the equations to obtain the corresponding Lumipulse^{TM} value are:

$$\begin{aligned} \text{Lumipulse}^{TM} &= 1.211 \times \text{Ella}^{TM} + 45.38 \quad \text{for CSF} \\ \text{Lumipulse}^{TM} &= 1.732 \times \text{Ella}^{TM} - 7.543 \quad \text{for serum} \end{aligned} \tag{3.6}$$

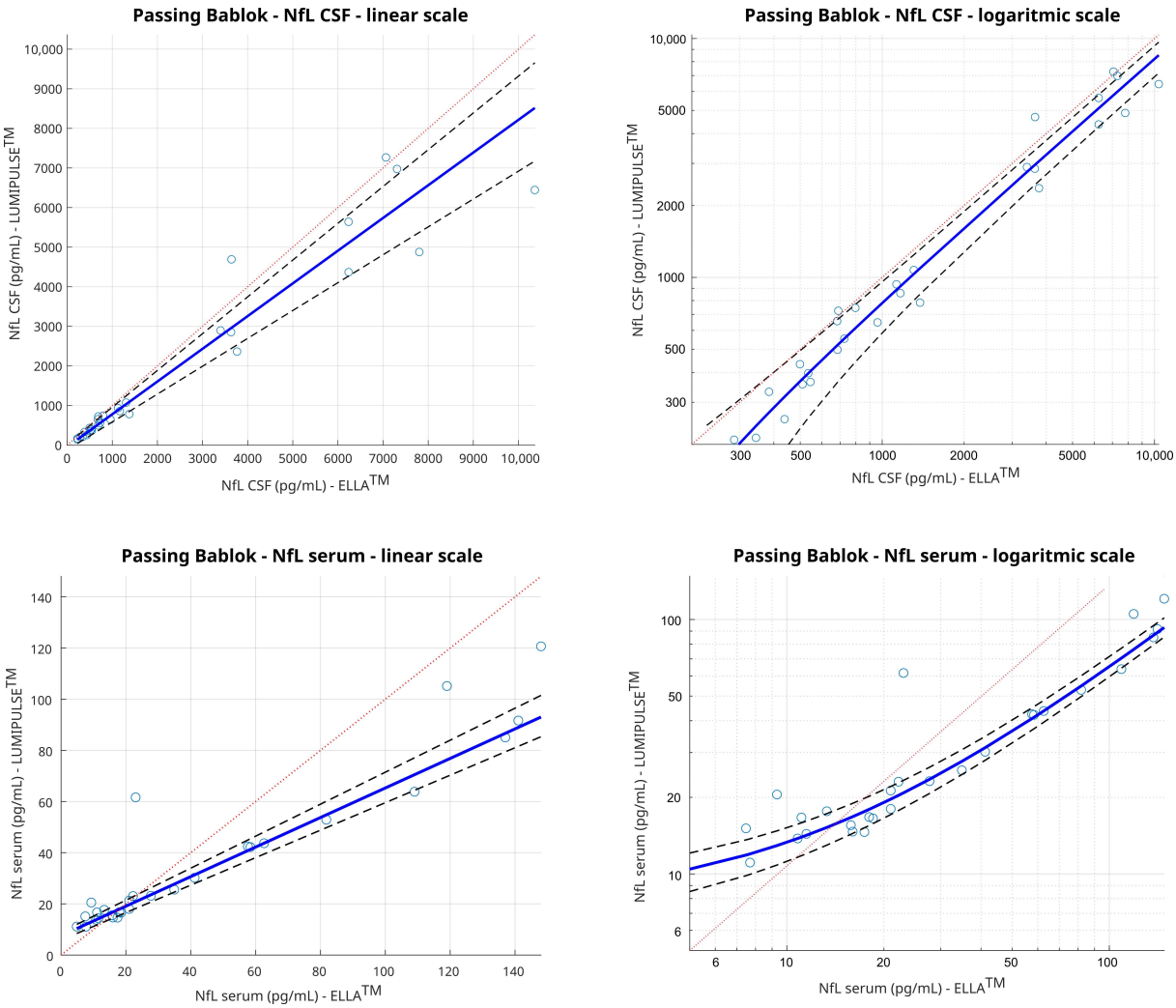

Figure 3.8. Passing–Bablok results, shown both on linear and logarithmic scales. The solid blue line represents the Passing–Bablok regression line, the dashed red line is the identity line, and the dashed grey lines represent the 95% confidence intervals. These graphics may be interpreted as follows: if the red identity line is outside of the confidence intervals of the Passing–Bablok regression line, this indicates a proportional difference between the two assays and thus a need for correction.

3.3.4 Discussion

In the study presented here, CSF and serum NfL levels in patients with cognitive disorders due to different diseases (AD, FTD, and non-progressive MCI) were measured with two different assays: Lumipulse™ and Ella™. Despite the high correlation, the two methods exhibited a significant proportional difference. We therefore estimated the equations allowing for the conversion from one to the other assay. These conversion equations enable cross-assay comparisons and dataset harmonization, facilitating meta-analyses on larger samples.

Before the study presented here, no previous similar comparisons had been made in patients with cognitive impairment or neurodegenerative diseases, as only one study to date has reported comparisons of NfL levels with Ella™ and Lumipulse™ in patients with multiple sclerosis [148]. Several studies, instead, have compared Ella™ with SIMOA™ technology. In these studies, the differences between the two assays were ascribed to

the use of different calibrators, namely naturally derived bovine NfLs for Ella™ and recombinant human NfLs for SIMOA™ [134, 149]. In our study, this should not be an issue as Lumipulse™, like Ella™, uses a naturally derived bovine NfL calibrator.

A theoretical limitation of our study was the size of our dataset, limited to 30 subjects. However, the univariate analysis previously described is also applicable in the context of a small sample size, and the narrow width of the confidence intervals suggests the stability of the regression models across subjects. Moreover, we purposely included a sample of patients with the two most common neurodegenerative dementias (AD and FTD) as well as a sample of subjects with cognitive concerns without any biomarker abnormality or clinical evidence of progression, allowing us to verify the correspondence between the two different methodologies for NfL measurement across a spectrum of patients representative of those usually presenting in cognitive neurology clinics.

In conclusion, as the quantification of NfLs is gaining an increasingly important role in neurodegenerative diseases associated with dementia (see Section 3.2), it is essential to know all the possible sources of variation in their measurement, including the type of assay. By establishing a reliable conversion factor, the results of different studies can be fully compared, even when values are obtained with different methods in different biological fluids. This could improve our understanding of the clinical importance of NfLs, which will, in turn, promote the wider adoption of this promising biomarker.

3.4 Abstracts of additional published studies

In the final section of this chapter, I present the abstracts of other research projects in which I was involved during my doctoral studies. These works are not discussed in detail because they are less related to the topic of the thesis and a full inclusion would have compromised the flow of the manuscript. In addition, most of these studies are primarily clinical rather than mathematical or computational. However, their abstracts are reported here to give a broader picture of my research activity and to show how my research contributions extended beyond active inference modeling and the computational analyses presented above. A complete list of published papers, along with oral and poster presentations, is available at the end of the thesis.

3.4.1 Cognitive reserve in young-onset cognitive impairment

Cognitive reserve (CR) reflects brain's resilience to pathology, enabling to maintain function despite structural damage. This study investigates its role in young-onset cognitive impairment (<65 years) beyond brain integrity and neurodegeneration.

Participants underwent neuropsychological assessment – including the Cognitive Reserve Index questionnaire (CRIq) –, magnetic resonance imaging (MRI), and blood neurofilaments light-chain (NfLs) measurement. Scores of global cognition and domain-specific cognition were derived from Principal Component Analyses of neuropsychological results. Linear regression models estimated CR's contribution to global and domain-specific cognition, alongside age, sex, MRI measures, and NfLs as predictors.

Among the 115 participants, global cognition was significantly explained by CR (effect size (ES) = 0.229), gray matter volume (ES = 0.348), and NfLs (ES = -0.302). The effect of CR was prominent on language and attentional-executive functions: while the CRIq subscore related to education predicted performance in both these domains, the subscore related to leisure activities was positively associated with the language domain

only. These findings highlight CR's protective role in young-onset cognitive impairment, particularly for non-amnestic cognitive domains. Since a high CR can mask or compensate for neurological cognitive disorders delaying its diagnosis, our results suggest that measures of CR, including time spent on leisure activities, should be considered when interpreting neuropsychological tests.

3.4.2 Resting-state networks and anosognosia in Alzheimer's disease

Recent evidence suggests that anosognosia or unawareness of cognitive impairment in Alzheimer's Disease (AD) may be explained by a disconnection between brain regions involved in accessing and monitoring information regarding self and others. It has been demonstrated that AD patients with anosognosia have reduced connectivity within the default mode network (DMN) and that anosognosia in people with prodromal AD is positively associated with bilateral anterior cingulate cortex (ACC), suggesting a possible role of this region in mechanisms of awareness in the early phase of disease. We hypothesized that anosognosia in AD is associated with an imbalance between the activity of large-scale resting-state functional magnetic resonance imaging (fMRI) networks, in particular the DMN, the salience network (SN), and the frontoparietal network (FPN).

Sixty patients with MCI and AD dementia underwent fMRI and neuropsychological assessment including the Anosognosia Questionnaire Dementia (AQ-D), a measure of anosognosia based on a discrepancy score between patient's and carer's judgments. After having applied Independent Component Analysis (ICA) to resting fMRI data we performed: (i) correlations between the AQ-D score and functional connectivity in the DMN, SN, and FPN, and (ii) comparisons between aware and unaware patients of the DMN, SN, and FPN functional connectivity.

We found that anosognosia was associated with (i) weak functional connectivity within the DMN, in posterior and middle cingulate cortex particularly, (ii) strong functional connectivity within the SN in ACC, and between the SN and basal ganglia, and (iii) a heterogenous effect concerning the functional connectivity of the FPN, with a weak connectivity between the FPN and PCC, and a strong connectivity between the FPN and ACC. The observed effects were controlled for differences in severity of cognitive impairment and age. Our results suggest that anosognosia in the AD continuum is associated with a dysregulation of the functional connectivity of three large-scale networks, namely the DMN, SN, and FPN.

3.4.3 Repeated brain MRI utility in identifying neurodegenerative disorders at the pre-dementia stage

Despite the increasing availability of biomarkers in clinical settings, diagnosing individuals with subtle cognitive/behavioral symptoms and normal structural brain imaging remains challenging. In real-world settings, it is not feasible to subject all such patients to costly and invasive second-level assessments, and there are no clear guidelines for identifying those who should undergo these procedures beyond clinical follow-up and structural imaging. The present study explores the potential of repeated magnetic resonance imaging (MRI) in this context. We investigated whether detecting pathological brain volume (BV) changes exceeding normal aging could aid in deciding whether advanced biomarker testing is reasonable.

Individuals with subtle cognitive/behavioral complaints underwent baseline and 18-month follow-up assessments (neuropsychological evaluation, MRI, and biomarker tests). Annualized percentage of BV change (PBVC/y), adjusted for age, sex, and scanner, was calculated with SIENA, and classified whether exceeding the 80th/95th percentiles of normative data. At follow-up, participants were classified (I) Converters/Non-Converters based on clinical information, (II) Non-converters+/Non-Converters- based on biomarkers status, and (III) Underlying disease based on both clinical follow-up and biomarkers. Logistic regressions assessed PBVC/y, baseline BV, and Mini-Mental State Examination change as predictors of participants' classification.

Among 110 participants, PBVC/y exceeding both the 80th/95th percentiles increased the probability of being classified Converters or Alzheimer's disease/Frontotemporal dementia. Exceeding the 80th percentile increased the odds of being classified Non-Converters+. Our results suggest that the application of SIENA normative data to repeated MRIs may help in deciding whether or not performing more invasive and high-cost second-level procedures in individuals with early-stage and subtle cognitive/behavioral changes.

3.4.4 Neural correlates of emotional valence for faces and words

Stimuli with negative emotional valence are especially apt to influence perception and action because of their crucial role in survival, a property that may not be precisely mirrored by positive emotional stimuli of equal intensity. The aim of this study was to identify the neural circuits differentially coding for positive and negative valence in the implicit processing of facial expressions and words, which are among the main ways human beings use to express emotions.

Thirty-six healthy subjects took part in an event-related fMRI experiment. We used an implicit emotional processing task with the visual presentation of negative, positive, and neutral faces and words, as primary stimuli. Dynamic Causal Modeling (DCM) of the fMRI data was used to test effective brain connectivity within two different anatomo-functional models, for the processing of words and faces, respectively.

In our models, the only areas showing a significant differential response to negative and positive valence across both face and word stimuli were early visual cortices, with faces eliciting stronger activations. For faces, DCM revealed that this effect was mediated by a facilitation of activity in the amygdala by positive faces and in the fusiform face area by negative faces; for words, the effect was mainly imputable to a facilitation of activity in the primary visual cortex by positive words. These findings support a role of early sensory cortices in discriminating the emotional valence of both faces and words, where the effect may be mediated chiefly by the subcortical/limbic visual route for faces, and rely more on the direct thalamic pathway to primary visual cortex for words.

3.4.5 Angry facial expressions elicit a late attentional withdrawal

Facial expressions of emotion are salient social cues that can trigger adaptive behavioral responses – such as facilitated approach or avoidance movements in response to positive or negative stimuli, respectively.

We conducted a behavioral experiment testing the hypothesis that while both happy and angry faces initially capture attention, angry expressions subsequently prompt a withdrawal of attention – a pattern consistent with aversive behavioral tendencies. Using a target-detection paradigm, we varied the spatial location and timing (50 ms vs. 350

ms) of targets relative to face stimuli. Results supported the hypothesis: angry faces led to faster responses to targets appearing nearer to the observer (i.e., farther from the face) at 350 ms, suggesting attentional retraction.

A follow-up fMRI experiment replicated this design with only the 350 ms delay. Behaviorally, responses were faster to targets near happy faces and away from angry ones. These effects correlated with subjective reports of being attracted or repelled by the expressions, as well as with individual differences in empathy and trait avoidance. Neuroimaging data revealed an interaction between emotion and target location in early visual cortices, modulated by subjective attraction-retraction ratings.

These findings support the idea of attention disengaging from angry faces, with both behavioral and neural correlates linked to individual affective dispositions.

Conclusions

This thesis aimed to show how mathematical models can provide insight into both cognitive mechanisms and clinical decisions. The first chapter introduced the framework of active inference, starting from the assumption that every agent is equipped with a generative model – a probabilistic internal model used to simulate how the world generates sensory inputs and how one's own actions can shape future observations. According to this view, perception is based on the minimization of variational free energy, which quantifies the discrepancy between sensory evidence and brain's predictions. However, active inference is not solely concerned with minimizing prediction error: agents also minimize expected free energy, a quantity that evaluates possible future outcomes that have not yet been observed.

Active inference in Partially Observable Markov Decision Processes (POMDPs) was then presented. In this framework, behavior is seen as a trade-off between actions that collect useful information (*exploration*) and actions that try to maximize immediate reward (*exploitation*). In POMDPs, concepts such as preferences (C) and policy priors (E) can be interpreted as latent parameters of cognition, potentially measurable from behavior. By viewing perception and action as two complementary parts of the same inferential process, active inference frames cognition as a continuous loop of prediction and correction. This idea provided the theoretical background for the empirical work subsequently presented in Chapter 2.

The study presented in Chapter 2 aimed at investigating how voluntary effort modulates cognitive control. Participants performed a Stroop color-word interference task under two conditions: EXERT, in which they were asked to engage maximal effort, and RELAX, in which they were encouraged to perform the task in a more relaxed manner. Within the active inference framework, effort can be formally defined in terms of the divergence between context-dependent goal-directed actions and context-independent habitual policies. To capture this computational distinction, participants' behavior was modeled using a two-level POMDP including two latent cognitive parameters, c and e – corresponding to the vectors C and E introduced in Chapter 1. These parameters represented, respectively, the individual's motivation to perform the task correctly and the strength of their habitual tendency to read words instead of naming their colors.

At the behavioral level, participants responded faster in the EXERT condition than in the RELAX condition. A group-level Bayesian model inversion revealed that this difference was best explained by an increase in the motivational parameter (c), rather than by a suppression of habit bias (e). This suggests that what participants do when asked to exert maximum effort is to intensify their motivation. In other words, when people 'try harder', they do not necessarily inhibit automatic responses more strongly; instead, they seem to enhance the subjective value of performing the task correctly. This interpretation reframes effort as a form of meta-control, consistent with other studies linking

effortful engagement to the activation of the dorsomedial striatum and other brain regions associated with motivational control. From a computational perspective, voluntary effort thus appears as an operation that increases the confidence in desired outcomes, thereby sharpening policy selection.

The third chapter extended the inferential framework from cognitive processes to clinical problems. Its general aim was to develop computational models to make diagnostic inference more explicit and quantitative. Three main problems were addressed: modeling the prevalence of different clinical presentations of early-onset dementia, defining an age-adjusted decision threshold for clinical biomarkers, and assessing the agreement between two measurement platforms for clinical biomarkers.

More specifically, in the first analysis, generalized logistic functions were fitted to population data to estimate age-specific prevalence curves for different clinical presentations of young-onset Alzheimer's disease (AD) and frontotemporal dementia (FTD). The results showed that the most common form of both diseases (amnesic AD and behavioral FTD) increased in prevalence across the entire age range, whereas atypical variants reached a plateau earlier. These findings indicate that the prevalence of different phenotypes is not uniform, suggesting that diagnostic priors should be conditioned on age.

The second study also analyzed the influence of age on a clinical variable. In this case, a Support Vector Machine model was applied to define a two-threshold, age-adjusted cutoff for a clinical biomarker, namely cerebrospinal fluid neurofilament light chain (CSF NfL). The model divided the diagnostic space into three zones: negative, positive and uncertain. This structure reflects the fact that borderline values are better flagged for further assessment rather than prematurely assigned to a diagnostic category. The resulting model achieved satisfactory specificity and sensitivity, while also limitating the width of the intermediate gray zone.

While the second study focused on the effect of age on biomarker interpretation, the third addressed the measurement of NfL itself, both in serum and CSF. A Passing-Bablok regression was applied to compare Ella^{TM} and Lumipulse^{TM} technologies, deriving a conversion equation between the two platforms. This formula allows measurements to be harmonized across different systems, enabling consistent inference in multi-site studies. Moreover, these results can be integrated with those of the second study, where NfL was measured using the Ella^{TM} platform: through the conversion equation, it becomes possible to recover the diagnostic thresholds also for Lumipulse^{TM} data.

All three clinical studies shared the fact that the guiding logic was inferential. Moreover, the explicit quantification of uncertainty – whether through credible intervals or intermediate diagnostic zones – aligns with ethical imperatives in clinical practice: an estimate only gains meaning when accompanied by a quantification of its uncertainty.

The active inference framework may offer powerful new tools for future clinical applications. For instance, in a task-based fMRI study, researchers may compare the latent variables estimated by an active inference model with measures of functional activation or effective connectivity (e.g., using Dynamic Causal Modeling). This approach would allow testing whether the computational quantities that explain behavior are also reflected in neural responses. Also, when using Dynamic Causal Modeling, latent cognitive variables may be incorporated as subject-specific regressors in the group-level Parametric Empirical Bayes analysis [150]. This would enable testing whether individual differences in policy evaluation also predict effective connectivity changes.

In a longer-term perspective, one could imagine defining a set of normative active inference models for widely used neuropsychological tests (such as Stroop, Go/No-Go, N-back, and Trail-Making), calibrated on healthy populations and stratified by age, education, and other demographical features. These models would provide clinicians with personalized diagnostic tools, allowing parameters derived from a patient's test to be compared against normative distributions. More broadly, a curated library of validated task models, shared code, and reference distributions would promote multi-site reproducibility and facilitate integration with clinical decision-support systems. In this way, active inference modeling could become a practical instrument in clinical workflows, offering not only explanations for why a patient performs abnormally, but also quantitative estimates of how much their control processes deviate from the normative population.

An example of a promising direction for future research concerns the application of active inference to the study of implicit awareness (or anosognosia) in AD. Our group is currently developing a model of the Emotional Stroop task – a behavioral experiment designed to investigate how emotional and self-relevant information affects decision-making. In this task, participants are asked to name the ink color of words that can be neutral, negative, or disease-related (e.g., *dementia*, *forgetful*), while ignoring their meaning. A typical finding is that, in the presence of explicit or implicit awareness, responses tend to be slower for emotionally relevant words, reflecting interference driven by implicit salience [92]. At the neural level, anosognosia has been linked to dysfunctions within large-scale networks such as the default mode and the salience networks [151]. In a POMDP model of this task, individual parameters could describe the subjective salience assigned to disease-related and negative words, allowing the model to reproduce the characteristic selective-slowing observed behaviorally. These model-derived "implicit salience" parameters could provide mechanistic markers of altered self-referential processing, and serve as quantitative indicators of awareness. In the long term, the active inference framework may contribute to transforming anosognosia from a descriptive clinical concept into a measurable computational phenotype. However, it is important to consider that, from a methodological point of view, clinical deployment also requires parameter identifiability and robustness to prior assumptions. Before a model can be used for inference, its parameters must demonstrate strong parameter recovery and posterior predictive validity under variations of prior means and variances. This check can be performed similarly to the analysis presented in Section 2.3.2 of this thesis.

More generally, it should be emphasized that, like any computational model, active inference is a formal tool for investigating empirical phenomena (in this case, the behavior of a biological system), rather than a direct representation of reality. Consequently, the results of an active inference model should not be read as "true" or "false" statements about the world: they are model-dependent inferences, whose meaning and reliability depend on the quality of the data and the assumptions encoded in the model. From this perspective, a key methodological risk is to over-trust quantitative parameters simply because they are formally well-defined. For this reason, any prospective use of active inference in clinical contexts requires systematic validation of the approach, including stability across sites and evidence that model-derived quantities provide reliable information beyond established descriptive measures. These considerations do not diminish the value of the framework. Active inference remains a particularly versatile modeling strategy: it can generate testable predictions and mechanistic hypotheses, while also providing a principled way to quantify uncertainty and to make explicit the assumptions that guide inference.

Finally, it is important to note that, from a mathematical point of view, active inference priors play a role similar to regularizers in numerical optimization. In active inference, the variational free energy comprises an accuracy term and a complexity term, which depend on prior beliefs over hidden states and policies (see Equation 1.10); this complexity term acts as a penalty, ensuring that problems remain well-posed and preventing the model from overfitting noisy observations. In this sense, choosing priors corresponds to selecting a regularization strategy: Gaussian priors act as $L2$ or Tikhonov penalties, while Laplace priors correspond to $L1$ penalties that promote sparsity in parameters or policies (see Section 1.3.1). Therefore, from a numerical standpoint, priors control ill-conditioning and stabilize inference, while from a neuroscientific point of view they encode assumptions about habits, policies, and precisions that guide behavior. This analogy also clarifies why active inference can be useful for medical studies: the free-energy criterion penalizes excessively flexible models in a way similar to classical regularizers, but also with an explicit interpretation that can be examined and, when necessary, adapted to the clinical question under investigation.

Taken together, the work presented in this thesis outlines a possible path toward a stronger integration between computational neuroscience and medicine. By modeling how individuals infer, decide, and adapt, the active inference framework offers a way to connect behavioral experiments, clinical diagnostics, and theoretical neuroscience. The next important step will be to embed these models into collaborative environments, where clinicians and data scientists can jointly develop tools that are both mathematically rigorous and clinically relevant. Such an approach could transform active inference from a purely computational framework into a clinical instrument, capable of guiding decisions in the same probabilistic terms that the brain itself uses to act upon the world.

List of publications

Published articles

1. **Angry facial expressions elicit a late attentional withdrawal**
   Ballotta D., ***Maramotti R.(first co-author)****, Borelli E., Lui F., Pagnoni G.*
   2025, Scientific Reports, 15, 41632. DOI: 10.1038/s41598-025-25609-w.

2. **Cognitive reserve in young-onset cognitive impairment**
   Carbone C., ***Maramotti R.****, Balboni E., Beltrami D., Ballotta D., Bedin R., Gallingani C., Tondelli M., Salemme S., Gasparini F., Vinceti G., Marti A., Chiari A., Nocetti L., Pagnoni G., Zamboni G.*
   2025, Brain and cognition, Jun:186:106297. DOI: 10.1016/j.bandc.2025.106297

3. **Repeated brain MRI utility in identifying neurodegenerative disorders at the pre-dementia stage**
   Carbone C., Gallingani G., Balboni E., Lucchetti L., Gentile G., ***Maramotti R.****, Ballotta D., Iacovino N., Cossutti S., Marti A., Tondelli M., Chiari A., Battaglini M., Zamboni G.*
   2025, Neurobiol Aging, Nov:155:35-43. DOI: 10.1016/j.neurobiolaging.2025.07.012

4. **Comparison of serum and cerebrospinal fluid neurofilament light chain concentrations measured by Ella™ and Lumipulse™ in patients with cognitive impairment**
   Urbano T., ***Maramotti R. (first co-author)****, Tondelli M., Gallingani C., Carbone C., Iacovino N., Vinceti G., Zamboni G., Chiari A., Bedin R.*
   2024, Diagnostics, 14(21), 2408. DOI: 10.3390/diagnostics14212408

5. **Age-specific prevalence of the different clinical presentations of AD and FTD in young onset dementia**
   Zamboni G., ***Maramotti R.****, Salemme S., Tondelli M., Adani G., Vinceti G., Carbone C., Filippini T., Vinceti M., Pagnoni G., Chiari A.*
   2024, J Neurol, Apr 21. DOI: 10.1007/s00415-024-12364-7

6. **Resting-state networks and anosognosia in Alzheimer's disease**
   Tondelli M., Ballotta D., ***Maramotti R.****, Carbone C., Gallingani C., MacKay C., Pagnoni G., Chiari A., Zamboni G.*
   2024, Front Aging Neurosci, 5:16:1415994. DOI: 10.3389/fnagi.2024.1415994

7. **Neural correlates of emotional valence for faces and words**
   Ballotta D., ***Maramotti R.****, Borelli E., Lui F., Pagnoni G.*
   2023, Front Psychol, 23:14:1055054. DOI: 10.3389/fpsyg.2023.1055054

Oral contributions

1. **Invited speaker in a symposium**
   Active inference and cognition: computational models of effort and awareness.
   Symposium "From Classical to Data-Driven: Modern Trends in Optimization and Regularization", 01-05 Sep 2025, SIMAI congress, Trieste.

2. **Invited speaker in a symposium**
   Active inference and psychological tests: modeling cognition and its impairments.
   Symposium "Active Inference in Psychiatry: Understanding Mechanisms of Psychopathology", 18-21 Jun 2025, 50th Annual Conference Psychology and the Brain (PuG), Würzburg.

3. **Oral presentation in a public engagement event**
   Does a "hidden" awareness of illness exist? How can we assess it?
   Event "La ricerca scientifica su disturbi cognitivi e demenze (IV edition)", 30 Nov 2024, Modena.

4. **Oral presentation in a PhD contest**
   Event "The transition of the near future", 28 Nov 2024, University of Ferrara, Ferrara.

5. **Oral presentation in a workshop**
   Age-specific prevalence of different clinical variants in young onset dementia
   Workshop "Optimization Algorithms and Software for Inverse Problems X-mas", 15 Dec 2023, Ferrara.

Posters

1. *Voluntary modulation of mental effort: an Active Inference model.*
   Maramotti R., Parr T., Tondelli M., Ballotta D., Zamboni G., Pagnoni G., 27-31 Jul 2025, Alzheimer's Association International Conference (AAIC), Toronto.

2. *Implicit awareness in Alzheimer's Disease: an Active Inference model.*
   Ballotta D., **Maramotti R.**, Parr T., Carbone C., Iacovino N., Tondelli M., Pagnoni G., Zamboni G., 27-31 Jul 2025, Alzheimer's Association International Conference (AAIC), Toronto.

3. *Resting-state functional connectivity is altered in hypernosognosia within the Alzheimer's Disease continuum.*
   Maramotti R., Ballotta D., Carbone C., Gallingani C., Pagnoni G., Zamboni G., Tondelli M., 24-28 Jun 2025, 31st Annual Meeting of the Organization for Human Brain Mapping (OHBM), Brisbane.

4. *An Active Inference model on implicit awareness in Alzheimer's disease.*
   Maramotti R., Parr T., Ballotta D., Carbone C., Iacovino N., Tondelli M., Pagnoni G., Zambini G., 24-28 Jun 2025, 31st Annual Meeting of the Organization for Human Brain Mapping (OHBM), Brisbane.

5. *Comparison of serum and cerebrospinal fluid neurofilament light chain concentrations measured by ELLA and LUMIPULSE in patients with cognitive impairment.*
   Maramotti R., Tondelli M., Urbano T., Gallingani C., Carbone C., Iacovino N., Vinceti G., Zamboni G., Chiari A., Bedin R., 23-25 Oct 2024, NeuroMI, Milano.

6. *The neural basis of self-awareness in the present and in the past.*
   Torrenti F., Ballotta D., **Maramotti R.**, Carbone C., Iacovino N., Tondelli M., Zamboni G., 23-25 Oct 2024, NeuroMI, Milano.

7. *Serum and cerebrospinal fluid NfL measurement in neurodegenerative dementias: identification of age-related cut-offs.*
   Bushaj R., Gallingani C., **Maramotti R.**, Bedin R., Carbone C., Chiari A., Tondelli M., Zamboni G., 23-25 Oct 2024, NeuroMI, Milano.

8. *The relationship between premorbid personality and anosognosia in Alzheimer's Disease.*
   Passov M., Carbone C., Iacovino N., Ballotta D., **Maramotti R.**, Tondelli M., Zamboni G., 23-25 Oct 2024, NeuroMI, Milano.

9. *Modeling motivation in Stroop Task using Active Inference.*
   Maramotti R., Parr T., Tondelli M., Ballotta D., Zamboni G., Manohar S., Pagnoni G., 4-6 Sep 2024, 5th Optimization Techniques for Inverse Problems workshop, Modena.

10. *Anosognosia in Alzheimer's disease: the interplay between resting state Default Mode Network and Salience Network.*
    Tondelli M., Ballotta D., Carbone C., **Maramotti R.**, Chiari A., Zamboni G., 28 Jul - 1 Aug 2024, Alzheimer's Association International Conference, Philadelphia.

11. *Markers of atrophy progression in patients with young-onset cognitive complaints not due to Alzheimer's Disease.*
    Carbone C., Balboni E., Luchetti L., Urbano T., Beltrami D., **Maramotti R.**, Tondelli M., Gallingani C., Gasparini F., Vinceti G., Marti A., Chiari A., Zamboni G., 28 Jul - 1 Aug 2024, Alzheimer's Association International Conference, Philadelphia.

12. *Functional internetworks connectivity in Alzheimer's disease and Mild Cognitive impairment patients with Anosognosia.*
    Ballotta D., Tondelli M., Carbone C., **Maramotti R.**, Chiari A., Pagnoni G., Zamboni G., 29 Jun - 2 Jul 2024, 10th Congress of the European Academy of Neurology, Helsinki.

13. *Comparison of serum and cerebrospinal fluid neurofilament light chain concentrations measured by ELLA and LUMIPULSE in patients with neurocognitive disorders.*
    Tondelli M., Urbano T., **Maramotti R.**, Gallingani C., Carbone C., Iacovino N., Vinceti G., Zamboni G., Chiari A., Bedin R., 27-28 June 2024, CSF Society meeting, Barcelona.

14. *Voluntary modulation of mental effort affects the attentional retraction induced by angry faces.*
    Ballotta D., **Maramotti R.**, Borelli E., Lui F., Pagnoni G., 23-27 June 2024, 30th Annual Meeting of the Organization for Human Brain Mapping (OHBM), Seul.

15. *The impact of Cognitive Reserve on cognitive performance in early onset Mild Cognitive Impairment.*
    Carbone, C., **Maramotti, R.**, Balboni, E., Ballotta, D., Bedin, R., Tondelli, M., Beltrami, D., Gasparini, F., Vinceti, G., Marti, A., Chiari, A., Nocetti, L., Pagnoni, G., Zamboni, G., 14-16 Feb 2024, SINdem4Juniors, Bressanone.

16. *The impact of Cognitive Reserve on neuropsychological performance in young onset Mild Cognitive Impairment.*
Maramotti R., Carbone C., Balboni E., Tondelli M., Beltrami D., Gasparini F., Vinceti G., Marti A., Chiari A., Nocetti L., Pagnoni G., Zamboni G., 5-9 Jun 2023, Bayesian Statistical Analysis for the Human, Social and Cognitive Science, Verona.

Bibliography

[1] Karl J. Friston et al. "Deep Temporal Models and Active Inference". In: *Neuroscience & Biobehavioral Reviews* 77 (2017), pp. 388–402. DOI: 10.1016/j.neubiorev.2017.04.009.

[2] Karl J. Friston et al. "Active inference: A process theory". In: *Neural Computation* 29.1 (2017), pp. 1–49. DOI: 10.1162/NECO_a_00912.

[3] Ryan Smith, Karl J. Friston, and Christopher J. Whyte. "A Step-by-Step Tutorial on Active Inference and Its Application to Empirical Data". In: *Journal of Mathematical Psychology* 107 (2022), p. 102632. DOI: 10.1016/j.jmp.2021.102632.

[4] David C. Knill and Alexandre Pouget. "The Bayesian brain: the role of uncertainty in neural coding and computation". In: *Trends in Neurosciences* 27.12 (2004), pp. 712–719. DOI: 10.1016/j.tins.2004.10.007.

[5] Kenji Doya et al., eds. *Bayesian Brain: Probabilistic Approaches to Neural Coding*. Cambridge, MA: MIT Press, 2007. ISBN: 9780262042383.

[6] Thomas Parr, Giovanni Pezzulo, and Karl J. Friston. *Active Inference: The Free Energy Principle in Mind, Brain, and Behavior*. The MIT Press, 2022. DOI: 10.7551/mitpress/12441.001.0001.

[7] Rajesh P. N. Rao and Dana H. Ballard. "Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects". In: *Nature Neuroscience* 2.1 (1999), pp. 79–87. DOI: 10.1038/4580.

[8] Karl J. Friston, Jean Daunizeau, and James Kilner. "Active inference and agency: optimal control without cost functions". In: *Biological Cybernetics* 106.8–9 (2012), pp. 523–541. DOI: 10.1007/s00422-012-0512-8.

[9] Giovanni Pezzulo, Francesco Rigoli, and Karl J. Friston. "Hierarchical active inference: a theory of motivated control". In: *Trends in Cognitive Sciences* 22.4 (2018), pp. 294–306. DOI: 10.1016/j.tics.2018.01.009.

[10] Karl J. Friston. "The free-energy principle: a rough guide to the brain?" In: *Trends in Cognitive Sciences* 13.7 (2009), pp. 293–301. DOI: 10.1016/j.tics.2009.04.005.

[11] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. New York: Springer, 2006. ISBN: 978-0387310732.

[12] Thomas Parr and Karl J. Friston. "Uncertainty, epistemics and active inference". In: *Journal of the Royal Society Interface* 14.136 (2017), p. 20170376. DOI: 10.1098/rsif.2017.0376.

[13] Karl Johan Åström. "Optimal control of Markov processes with incomplete state information". In: *Journal of Mathematical Analysis and Applications* 10.1 (1965), pp. 174–205. DOI: 10.1016/0022-247X(65)90154-X.

[14] Leslie Pack Kaelbling, Michael L. Littman, and Anthony R. Cassandra. "Planning and acting in partially observable stochastic domains". In: *Artificial Intelligence* 101.1–2 (1998), pp. 99–134. DOI: 10.1016/S0004-3702(98)00023-X.

[15] Iadine Chadès et al. "A primer on partially observable Markov decision processes". In: *Methods in Ecology and Evolution* 12.12 (2021), pp. 2519–2532. DOI: 10.1111/2041-210X.13692.

[16] Rajesh P. N. Rao. "Decision making under uncertainty: a neural model based on partially observable Markov decision processes". In: *Frontiers in Computational Neuroscience* 4 (2010), p. 146. DOI: 10.3389/fncom.2010.00146.

[17] John Winn and Christopher M. Bishop. "Variational Message Passing". In: *Journal of Machine Learning Research* 6 (2005), pp. 661–694. URL: https://jmlr.org/papers/v6/winn05a.html.

[18] David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. "Variational Inference: A Review for Statisticians". In: *Journal of the American Statistical Association* 112.518 (2017), pp. 859–877. DOI: 10.1080/01621459.2017.1285773.

[19] T. Parr et al. "Neuronal message passing using Mean-field, Bethe, and Marginal approximations". In: *Scientific Reports* 9.1 (2019), p. 1889. DOI: doi.org/10.1038/s41598-018-38246-3.

[20] Karl J. Friston, Thomas Parr, and Bert de Vries. "The graphical brain: Belief propagation and active inference". In: *Network Neuroscience* 1.4 (2017), pp. 381–414. DOI: 10.1162/NETN_a_00018.

[21] Karl J. Friston et al. "Variational free energy and the Laplace approximation". In: *NeuroImage* 34.1 (2007), pp. 220–234. DOI: 10.1016/j.neuroimage.2006.08.035.

[22] Karl J. Friston. "Variational filtering". In: *NeuroImage* 41.3 (2008), pp. 747–766. DOI: 10.1016/j.neuroimage.2008.03.017.

[23] William D. Penny et al., eds. *Statistical Parametric Mapping: The Analysis of Functional Brain Images*. London: Elsevier/Academic Press, 2007. ISBN: 9780123725608.

[24] Jean Daunizeau, Karl J. Friston, and Stefan J. Kiebel. "Variational Bayesian identification and prediction of stochastic nonlinear dynamic causal models". In: *Physica D: Nonlinear Phenomena* 238.21 (2009), pp. 2089–2118. DOI: 10.1016/j.physd.2009.08.002.

[25] Klaas E. Stephan et al. "Ten simple rules for dynamic causal modeling". In: *NeuroImage* 49.4 (2010), pp. 3099–3109. DOI: 10.1016/j.neuroimage.2009.11.015.

[26] Thomas Parr and Karl J. Friston. "Working memory, attention, and salience in active inference". In: *Scientific Reports* 7.1 (2017), p. 14678. DOI: 10.1038/s41598-017-15249-0.

[27] Conor Heins et al. "pymdp: A Python library for active inference in discrete state spaces". In: *Journal of Open Source Software* 7.73 (2022), p. 4098. DOI: 10.21105/joss.04098.

[28] Tim N. Nisslbeck et al. "Introducing ActiveInference.jl: A Julia Library for Simulation and Model Inversion in Active Inference". In: *Entropy* 27.1 (2025), p. 62. DOI: 10.3390/e27010062.

[29] Thijs van de Laar and Bert de Vries. "Simulating Active Inference Processes by Message Passing". In: *Frontiers in Robotics and AI* 6 (2019), p. 20. DOI: 10.3389/frobt.2019.00020.

[30] Joshua Shepherd. "Conscious cognitive effort in cognitive control". In: *WIREs Cognitive Science* 14.2 (2023), e1629. DOI: 10.1002/wcs.1629.

[31] Omar T. Khachouf et al. "Voluntary modulation of mental effort investment: an fMRI study". In: *Scientific Reports* 7.1 (2017), p. 17191. DOI: 10.1038/s41598-017-17519-3.

[32] Brian Bruya and Yi-Yuan Tang. "Is Attention Really Effort? Revisiting Daniel Kahneman's Influential 1973 Book". In: *Frontiers in psychology* 9 (2018), p. 1133. DOI: 10.3389/fpsyg.2018.01133.

[33] Noham Wolpe, Richard Holton, and Paul C. Fletcher. "What Is Mental Effort: A Clinical Perspective". In: *Biological Psychiatry* 95.11 (2024), pp. 1030–1037. DOI: 10.1016/j.biopsych.2024.01.022.

[34] Daniel Kahneman. *Attention and Effort*. Prentice-Hall, 1973. ISBN: 0130505188.

[35] Martin Sarter, William J. Gehring, and Rouba Kozak. "More attention must be paid: the neurobiology of attentional effort." In: *Brain Res Rev* 51.2 (2006), pp. 145–160. DOI: 10.1016/j.brainresrev.2005.11.002.

[36] M. Muraven and R. F. Baumeister. "Self-regulation and depletion of limited resources: does self-control resemble a muscle?" In: *Psychol Bull* 126.2 (2000), pp. 247–259. DOI: 10.1037/0033-2909.126.2.247.

[37] Maria M. Robinson and Ezequiel Morsella. "The subjective effort of everyday mental tasks: Attending, assessing, and choosing." In: *Motivation and Emotion* 38.6 (2014), pp. 832–843. DOI: 10.1007/s11031-014-9441-2.

[38] Michael Inzlicht, Amitai Shenhav, and Christopher Y. Olivola. "The Effort Paradox: Effort Is Both Costly and Valued." In: *Trends in Cognitive Sciences* 22 (4 2018), pp. 337–349. DOI: 10.1016/j.tics.2018.01.007.

[39] Marcell Székely and John Michael. "The Sense of Effort: a Cost-Benefit Theory of the Phenomenology of Mental Effort". In: *Review of Philosophy and Psychology* 12.4 (2021), pp. 889–904. DOI: 10.1007/s13164-020-00512-7.

[40] Peter Carruthers and David M. Williams. "Model-free metacognition." In: *Cognition* 225 (2022), p. 105117. DOI: 10.1016/j.cognition.2022.105117.

[41] W. P. Morgan. "Psychological components of effort sense." In: *Med Sci Sports Exerc* 26.9 (1994), pp. 1071–1077.

[42] Robert Kurzban. "The sense of effort". In: *Current Opinion in Psychology* 7 (2016), pp. 67–70. DOI: 10.1016/j.copsyc.2015.08.003.

[43] Liane Schmidt et al. "Get aroused and be stronger: emotional facilitation of physical effort in the human brain." In: *Journal of Neuroscience* 29.30 (2009), pp. 9450–9457. DOI: 10.1523/JNEUROSCI.1951-09.2009.

[44] Paula L. Croxson et al. "Effort-Based Cost-Benefit Valuation and the Human Brain". In: *J. Neurosci.* 29.14 (2009), pp. 4531–4541. DOI: 10.1523/jneurosci.4515-08.2009.

[45] Amitai Shenhav et al. "Toward a Rational and Mechanistic Account of Mental Effort." In: *Annual review of neuroscience* 40 (2017), pp. 99–124. DOI: 10.1146/annurev-neuro-072116-031526.

[46] Juliette Bénon et al. "The online metacognitive control of decisions". In: *Communications Psychology* 2.1 (2024), p. 23. DOI: 10.1038/s44271-024-00071-y.

[47] Sanjay G. Manohar et al. "Reward Pays the Cost of Noise Reduction in Motor and Cognitive Control". In: *Current Biology* 25.13 (2015), pp. 1707–1716. DOI: 10.1016/j.cub.2015.05.038.

[48] Wouter Kool and Matthew Botvinick. "Mental labour". In: *Nature Human Behaviour* 2 (12 2018), pp. 899–908. DOI: 10.1038/s41562-018-0401-9.

[49] John D. Salamone et al. "Activational and effort-related aspects of motivation: neural mechanisms and implications for psychopathology." In: *Brain* 139.5 (2016), pp. 1325–1347. DOI: 10.1093/brain/aww050.

[50] Mark E. Walton et al. "Functional specialization within medial frontal cortex of the anterior cingulate for evaluating effort-related decisions." In: *J Neurosci* 23.16 (2003), pp. 6475–6479. DOI: 10.1523/JNEUROSCI.23-16-06475.2003.

[51] Mark E. Walton et al. "Weighing up the benefits of work: behavioral and neural analyses of effort-related decision making." In: *Neural Netw* 19.8 (2006), pp. 1302–1314. DOI: 10.1016/j.neunet.2006.03.005.

[52] Mark E. Walton and Sebastien Bouret. "What Is the Relationship between Dopamine and Effort?" In: *Trends in Neurosciences* 42 (2 2019), pp. 79–91. DOI: 10.1016/j.tins.2018.10.001.

[53] A. Westbrook and T. S. Braver. "Dopamine Does Double Duty in Motivating Cognitive Effort." In: *Neuron* 89 (4 2016), pp. 695–710. DOI: 10.1016/j.neuron.2015.12.029.

[54] A. Westbrook et al. "Dopamine promotes cognitive effort by biasing the benefits versus costs of cognitive work." In: *Science* 367 (6484 2020), pp. 1362–1366. DOI: 10.1126/science.aaz5891.

[55] Jay G. Hosking, Stan B. Floresco, and Catharine A. Winstanley. "Dopamine antagonism decreases willingness to expend physical, but not cognitive, effort: a comparison of two rodent cost/benefit decision-making tasks." In: *Neuropsychopharmacology : official publication of the American College of Neuropsychopharmacology* 40 (4 2015), pp. 1005–1015. DOI: 10.1038/npp.2014.285.

[56] Timothy D. Noakes. "Fatigue is a Brain-Derived Emotion that Regulates the Exercise Behavior to Ensure the Protection of Whole Body Homeostasis." In: *Front Physiol* 3 (2012), p. 82. DOI: 10.3389/fphys.2012.00082.

[57] Michael Inzlicht and Samuele M. Marcora. "The Central Governor Model of Exercise Regulation Teaches Us Precious Little about the Nature of Mental Fatigue and Self-Control Failure." In: *Frontiers in psychology* 7 (2016), p. 656. DOI: 10.3389/fpsyg.2016.00656.

[58] Thomas Parr et al. "Cognitive effort and active inference". In: *Neuropsychologia* 184 (2023), p. 108562. DOI: 10.1016/j.neuropsychologia.2023.108562.

[59] A. Dickinson. "Actions and Habits: The Development of Behavioural Autonomy". In: *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 308.1135 (1985), pp. 67–78. DOI: 10.1098/rstb.1985.0010.

[60] Daniel Kahneman. "Maps of Bounded Rationality: Psychology for Behavioral Economics". In: *The American Economic Review* 93.5 (2003), pp. 1449–1475. DOI: 10.1257/000282803322655392.

[61] Wouter Kool, Samuel J. Gershman, and Fiery A. Cushman. "Cost-Benefit Arbitration Between Multiple Reinforcement-Learning Systems." In: *Psychological science* 28 (9 2017), pp. 1321–1333. DOI: 10.1177/0956797617708288.

[62] Jonathan Peirce et al. "PsychoPy2: Experiments in behavior made easy". In: *Behavior Research Methods* 51.1 (2019), pp. 195–203. DOI: 10.3758/s13428-018-01193-y.

[63] Sandra G. Hart and Lowell E. Staveland. "Development of NASA-TLX (task load index): Results of empirical and theoretical research". In: *Human Mental Workload*. Vol. 52. Advances in Psychology. Elsevier, 1988, pp. 139–183. DOI: 10.1016/S0166-4115(08)62386-9.

[64] Robert Ratcliff and Gail McKoon. "The diffusion decision model: theory and data for two-choice decision tasks". In: *Neural Comput.* 20.4 (2008), pp. 873–922. DOI: 10.1162/neco.2008.12-06-420.

[65] Karl J. Friston et al. "Bayesian model reduction and empirical Bayes for group (DCM) studies". In: *NeuroImage* 128 (2016), pp. 413–431. DOI: 10.1016/j.neuroimage.2015.11.015.

[66] J. Martin Bland and Douglas G. Altman. "Statistical methods for assessing agreement between two methods of clinical measurement". In: *The Lancet* 327.8476 (1986), pp. 307–310. DOI: 10.1016/S0140-6736(86)90837-8.

[67] Colin F. Camerer and Robin M. Hogarth. "The Effects of Financial Incentives in Experiments: A Review and Capital-Labor-Production Framework". In: *Journal of Risk and Uncertainty* 19.1 (1999), pp. 7–42. DOI: 10.1023/A:1007850605129.

[68] Koji Jimura, Hannah S. Locke, and Todd S. Braver. "Prefrontal cortex mediation of cognitive enhancement in rewarding motivational contexts." In: *Proceedings of the National Academy of Sciences of the United States of America* 107 (19 2010), pp. 8871–8876. DOI: 10.1073/pnas.1002007107.

[69] Ruth M. Krebs, Carsten N. Boehler, and Marty G. Woldorff. "The influence of reward associations on conflict processing in the Stroop task." In: *Cognition* 117 (3 2010), pp. 341–347. DOI: 10.1016/j.cognition.2010.08.018.

[70] Navid Hakimi et al. "Computational Mechanisms of Learning and Forgetting Differentiate Affective and Substance Use Disorders." In: *Research Square* (2024). DOI: 10.21203/rs.3.rs-4682224/v1.

[71] Elizabeth L. Fisher et al. "Psilocybin increases optimistic engagement over time: computational modelling of behaviour in rats". In: *Translational Psychiatry* 14.1 (2024), p. 394. DOI: 10.1038/s41398-024-03103-7.

[72] M. Berk Mirza et al. "Human visual exploration reduces uncertainty about the sensed world." In: *PloS one* 13 (1 2018), e0190429. DOI: 10.1371/journal.pone.0190429.

[73] Jakub Limanowski and Karl J. Friston. "'Seeing the Dark': Grounding Phenomenal Transparency and Opacity in Precision Estimation for Active Inference." In: *Frontiers in psychology* 9 (2018), p. 643. DOI: 10.3389/fpsyg.2018.00643.

[74] Lars Sandved-Smith et al. "Towards a computational phenomenology of mental action: modelling meta-awareness and attentional control with deep parametric active inference." In: *Neuroscience of consciousness* 2021 (2 2021), niab018. DOI: 10.1093/nc/niab018.

[75] Cameron Parro, Matthew L. Dixon, and Kalina Christoff. "The neural basis of motivational influences on cognitive control". In: *Human Brain Mapping* 39.12 (2018), pp. 5097–5111. DOI: 10.1002/hbm.24348.

[76] Stefano I. Di Domenico and Richard M. Ryan. "The Emerging Neuroscience of Intrinsic Motivation: A New Frontier in Self-Determination Research". In: *Frontiers in Human Neuroscience* 11 (2017). DOI: 10.3389/fnhum.2017.00145.

[77] Michael T. Treadway and John D. Salamone. "Vigor, Effort-Related Aspects of Motivation and Anhedonia." In: *Current topics in behavioral neurosciences* 58 (2022), pp. 325–353. DOI: 10.1007/7854_2022_355.

[78] John D. Salamone and Mercè Correa. "The Neurobiology of Activational Aspects of Motivation: Exertion of Effort, Effort-Based Decision Making, and the Role of Dopamine." In: *Annual review of psychology* 75 (2024), pp. 1–32. DOI: 10.1146/annurev-psych-020223-012208.

[79] Richard Levy and Bruno Dubois. "Apathy and the functional anatomy of the prefrontal cortex-basal ganglia circuits." In: *Cerebral cortex (New York, N.Y. : 1991)* 16 (7 2006), pp. 916–928. DOI: 10.1093/cercor/bhj043.

[80] C. Le Heron, M. A. J. Apps, and M. Husain. "The anatomy of apathy: A neurocognitive framework for amotivated behaviour." In: *Neuropsychologia* 118 (Pt B 2018), pp. 54–67. DOI: 10.1016/j.neuropsychologia.2017.07.003.

[81] Alexander Soutschek and Philippe N. Tobler. "Causal role of lateral prefrontal cortex in mental effort and fatigue." In: *Human brain mapping* 41 (16 2020), pp. 4630–4640. DOI: 10.1002/hbm.25146.

[82] Bernard W. Balleine and John P. O'Doherty. "Human and rodent homologies in action control: corticostriatal determinants of goal-directed and habitual action." In: *Neuropsychopharmacology : official publication of the American College of Neuropsychopharmacology* 35 (1 2010), pp. 48–69. DOI: 10.1038/npp.2009.131.

[83] Melissa Malvaez. "Neural substrates of habit." In: *Journal of neuroscience research* 98 (6 2020), pp. 986–997. DOI: 10.1002/jnr.24552.

[84] Elizabeth Tricomi, Bernard W. Balleine, and John P. O'Doherty. "A specific role for posterior dorsolateral striatum in human habit learning." In: *The European journal of neuroscience* 29 (11 2009), pp. 2225–2232. DOI: 10.1111/j.1460-9568.2009.06796.x.

[85] Eike K. Buabang et al. "Leveraging cognitive neuroscience for making and breaking real-world habits." In: *Trends in Cognitive Sciences* 29 (1 2025), pp. 41–59. DOI: 10.1016/j.tics.2024.10.006.

[86] Robert Kurzban et al. "An opportunity cost model of subjective effort and task performance." In: *The Behavioral and brain sciences* 36 (6 2013), pp. 661–679. DOI: 10.1017/S0140525X12003196.

[87] Wouter Kool and Matthew Botvinick. "A labor/leisure tradeoff in cognitive control." In: *Journal of experimental psychology. General* 143 (1 2014), pp. 131–141. DOI: 10.1037/a0031048.

[88] Harrison Ritz, Xiamin Leng, and Amitai Shenhav. "Cognitive Control as a Multivariate Optimization Problem". In: *J Cogn Neurosci* 34.4 (2022), pp. 569–591. DOI: 10.1162/jocn_a_01822.

[89] Julie M. Bugg, Larry L. Jacoby, and Jeffrey P. Toth. "Multiple levels of control in the Stroop task." In: *Memory & cognition* 36 (8 2008), pp. 1484–1494. DOI: 10.3758/MC.36.8.1484.

[90] Corentin Gonthier, Todd S. Braver, and Julie M. Bugg. "Dissociating proactive and reactive control in the Stroop task." In: *Memory & cognition* 44 (5 2016), pp. 778–788. DOI: 10.3758/s13421-016-0591-1.

[91] Avishai Henik, Julie M. Bugg, and Liat Goldfarb. "Inspired by the past and looking to the future of the Stroop effect". In: *Acta Psychologica* 189 (2018), pp. 1–3. DOI: 10.1016/j.actpsy.2018.06.007.

[92] A. Martyr et al. "Dissociation between implicit and explicit manifestations of awareness in early stage dementia: evidence from the emotional Stroop effect for dementia-related words." In: *Int J Geriatr Psychiatry* 26 (1 2011), pp. 92–99. DOI: 10.1002/gps.2495.

[93] Manuela Tondelli et al. "Eliciting Implicit Awareness in Alzheimer's Disease and Mild Cognitive Impairment: A Task-Based Functional MRI Study". In: *Frontiers in Aging Neuroscience* 14 (2022), p. 816684. DOI: 10.3389/fnagi.2022.816648.

[94] Giovanna Zamboni et al. "Age-specific prevalence of the different clinical presentations of Alzheimer's disease and frontotemporal dementia in young-onset dementia". In: *Journal of Neurology* 271.7 (2024), pp. 4326–4335. DOI: 10.1007/s00415-024-12364-7.

[95] R. Brookmeyer et al. "National estimates of the prevalence of Alzheimer's disease in the United States". In: *Alzheimer's & Dementia* 7.1 (2011), pp. 61–73. DOI: 10.1016/j.jalz.2010.11.007.

[96] K. B. Rajan et al. "Population estimate of people with clinical Alzheimer's disease and mild cognitive impairment in the United States (2020–2060)". In: *Alzheimer's & Dementia* 17 (2021), pp. 1966–1975. DOI: 10.1002/alz.12362.

[97] H. Niu et al. "Prevalence and incidence of Alzheimer's disease in Europe: a meta-analysis". In: *Neurologia* 32 (2017), pp. 523–532. DOI: 10.1016/j.nrl.2016.02.016.

[98] M. Kvello-Alme et al. "The prevalence and subtypes of young onset dementia in central Norway: a population-based study". In: *Journal of Alzheimer's Disease* 69 (2019), pp. 479–487. DOI: 10.3233/JAD-181223.

[99] A. Withall et al. "The prevalence and causes of younger onset dementia in eastern Sydney, Australia". In: *International Psychogeriatrics* 26 (2014), pp. 1955–1965. DOI: 10.1017/S1041610214001835.

[100] C. Ikejima et al. "Prevalence and causes of early-onset dementia in Japan: a population-based study". In: *Stroke* 40 (2009), pp. 2709–2714. DOI: 10.1161/strokeaha.108.542308.

[101] M. L. Gorno-Tempini et al. "Classification of primary progressive aphasia and its variants". In: *Neurology* 76.11 (2011), pp. 1006–1014. DOI: 10.1212/WNL.0b013e31821103e6.

[102] S. J. Crutch et al. "Consensus classification of posterior cortical atrophy". In: *Alzheimer's & Dementia* 13.8 (2017), pp. 870–884. DOI: 10.1016/j.jalz.2017.01.014.

[103] R. Ossenkoppele et al. "The behavioural/dysexecutive variant of Alzheimer's disease: clinical, neuroimaging and pathological features". In: *Brain* 138.9 (2015), pp. 2732–2749. DOI: 10.1093/brain/awv191.

[104] E. Ratnavalli et al. "The prevalence of frontotemporal dementia". In: *Neurology* 58 (2002), pp. 1615–1621. DOI: 10.1212/WNL.58.11.1615.

[105] S. Hendriks et al. "Global prevalence of young-onset dementia: a systematic review and meta-analysis". In: *JAMA Neurology* 78.9 (2021), pp. 1080–1090. DOI: 10.1001/jamaneurol.2021.2161.

[106] R. J. Harvey, M. Skelton-Robinson, and M. N. Rossor. "The prevalence and causes of dementia in people under the age of 65 years". In: *Journal of Neurology, Neurosurgery & Psychiatry* 74 (2003), pp. 1206–1209. DOI: 10.1136/jnnp.74.9.1206.

[107] K. Wada-Isoe et al. "Epidemiological survey of frontotemporal lobar degeneration in Tottori prefecture, Japan". In: *Dementia and Geriatric Cognitive Disorders Extra* 2 (2012), pp. 381–386. DOI: 10.1159/000342972.

[108] J. Rahimi and G. G. Kovacs. "Prevalence of mixed pathologies in the aging brain". In: *Alzheimer's Research & Therapy* 6 (2014), p. 82. DOI: 10.1186/s13195-014-0082-1.

[109] A. Chiari et al. "Epidemiology of early onset dementia and its clinical presentations in the province of Modena, Italy". In: *Alzheimer's & Dementia* 17 (2021), pp. 81–88. DOI: 10.1002/alz.12177.

[110] P. G. Gottschalk and J. R. Dunn. "The five-parameter logistic: a characterization and comparison with the four-parameter logistic". In: *Analytical Biochemistry* 343 (2005), pp. 54–65. DOI: 10.1016/j.ab.2005.04.035.

[111] F. J. Richards. "A flexible growth function for empirical use". In: *Journal of Experimental Botany* 10 (1959), pp. 290–300. DOI: 10.1093/jxb/10.2.290.

[112] G. Cardillo. *Five parameters logistic regression—there and back again*. 2012. URL: http://www.mathworks.com/matlabcentral/fileexchange/38043.

[113] E. J. Rogalski et al. "Increased frequency of learning disability in patients with primary progressive aphasia and their first-degree relatives". In: *JAMA Neurology* 71.12 (2014), pp. 1564–1567. DOI: 10.1001/jamaneurol.2014.1710.

[114] J. Persson et al. "Longitudinal structure-function correlates in elderly reveal MTL dysfunction with cognitive decline". In: *Cerebral Cortex* 22.10 (2012), pp. 2297–2304. DOI: 10.1093/cercor/bhr306.

[115] G. Peeters et al. "Sex differences in the incidence and prevalence of young-onset Alzheimer's disease: a meta-analysis". In: *International Journal of Geriatric Psychiatry* 37 (1 2022), pp. 1–9. DOI: 10.1002/gps.5612.

[116] P. Bomont. "The dazzling rise of neurofilaments: Physiological functions and roles as biomarkers". In: *Current Opinion in Cell Biology* 68 (2021), pp. 181–191. DOI: 10.1016/j.ceb.2020.10.009.

[117] C. Bridel et al. "Diagnostic Value of Cerebrospinal Fluid Neurofilament Light Protein in Neurology: A Systematic Review and Meta-analysis". In: *JAMA Neurology* 76 (2019), pp. 1035–1048. DOI: 10.1001/jamaneurol.2019.1534.

[118] C. Barro, T. Chitnis, and H.L. Weiner. "Blood neurofilament light: A critical review of its application to neurologic disease". In: *Annals of Clinical and Translational Neurology* 7 (2020), pp. 2508–2523. DOI: 10.1002/acn3.51234.

[119] M. Khalil et al. "Neurofilaments as biomarkers in neurological disorders". In: *Nature Reviews Neurology* 14 (2018), pp. 577–589. DOI: 10.1038/s41582-018-0058-z.

[120] A. Yuan and R.A. Nixon. "Neurofilament proteins as biomarkers to monitor neurological diseases and the efficacy of therapies". In: *Frontiers in Neuroscience* 15 (2021), p. 689938. DOI: 10.3389/fnins.2021.689938.

[121] L. Gaetani et al. "Neurofilament light chain as a biomarker in neurological disorders". In: *Journal of Neurology, Neurosurgery & Psychiatry* 90.8 (2019), pp. 870–881. DOI: 10.1136/jnnp-2018-320106.

[122] M. Khalil, C. E. Teunissen, S. Lehmann, et al. "Neurofilaments as biomarkers in neurological disorders — towards clinical application". In: *Nature Reviews Neurology* 20.5 (2024), pp. 269–287. DOI: 10.1038/s41582-024-00955-x.

[123] M. Kalm et al. "Serum concentrations of the axonal injury marker neurofilament light protein are not influenced by blood-brain barrier permeability". In: *Brain Research* 1668 (2017), pp. 12–19. DOI: 10.1016/j.brainres.2017.04.018.

[124] T. Uher et al. "Neurofilament levels are associated with blood-brain barrier integrity, lymphocyte extravasation, and risk factors following the first demyelinating event in multiple sclerosis". In: *Multiple Sclerosis* 27 (2021), pp. 220–231. DOI: 10.1177/1352458520912379.

[125] M. Khalil, L. Pirpamer, E. Hofer, et al. "Serum neurofilament light levels in normal aging and their association with morphologic brain changes". In: *Nature Communications* 11.1 (2020), p. 812. DOI: 10.1038/s41467-020-14612-6.

[126] S. Coppens et al. "Neurofilament-Light, a Promising Biomarker: Analytical, Metrological and Clinical Challenges". In: *International Journal of Molecular Sciences* 24.14 (2023), p. 11624. DOI: 10.3390/ijms241411624.

[127] J. Sarto, D. Esteller-Gauxax, A. Tort-Merino, et al. "Impact of demographics and comorbid conditions on plasma biomarkers concentrations and their diagnostic accuracy in a memory clinic cohort". In: *Journal of Neurology* 271.4 (2024), pp. 1973–1984. DOI: 10.1007/s00415-023-12153-8.

[128] A. Rodero-Romero, E. Monreal, R. Sainz-Amo, et al. "Establishing Normal Serum Values of Neurofilament Light Chains and Glial Fibrillary Acidic Protein Considering the Effects of Age and Other Demographic Factors in Healthy Adults". In: *International Journal of Molecular Sciences* 25.14 (2024), p. 7808. DOI: 10.3390/ijms25147808.

[129] K. C. Fitzgerald et al. "Contributors to Serum NfL Levels in People without Neurologic Disease". In: *Annals of Neurology* 92.4 (2022), pp. 688–698. DOI: 10.1002/ana.26446.

[130] Guy M. McKhann et al. "The diagnosis of dementia due to Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease". In: *Alzheimer's & Dementia* 7.3 (2011), pp. 263–269. DOI: 10.1016/j.jalz.2011.03.005.

[131] Katya Rascovsky et al. "Sensitivity of revised diagnostic criteria for the behavioural variant of frontotemporal dementia". In: *Brain* 134.9 (2011), pp. 2456–2477. DOI: 10.1093/brain/awr179.

[132] I. G. McKeith, B. F. Boeve, D. W. Dickson, et al. "Diagnosis and management of dementia with Lewy bodies: Fourth consensus report of the DLB Consortium". In: *Neurology* 89.1 (2017), pp. 88–100. DOI: 10.1212/WNL.0000000000004058.

[133] M. Notzel et al. "Ella versus Simoa Serum Neurofilament Assessment to Monitor Treatment Response in Highly Active Multiple Sclerosis Patients". In: *International Journal of Molecular Sciences* 23 (2022), p. 12361. DOI: 10.3390/ijms232012361.

[134] M. Truffi et al. "Neurofilament-light chain quantification by Simoa and Ella in plasma from patients with dementia: A comparative study". In: *Scientific Reports* 13 (2023), p. 4041. DOI: 10.1038/s41598-023-29704-8.

[135] J. Gaiottino, N. Norgren, R. Dobson, et al. "Increased neurofilament light chain blood levels in neurodegenerative neurological diseases". In: *PLoS ONE* 8.9 (2013), e75091. DOI: 10.1371/journal.pone.0075091.

[136] P. C. Bürkner. "brms: An R Package for Bayesian Multilevel Models Using Stan". In: *Journal of Statistical Software* 80.1 (2017), pp. 1–28. DOI: 10.18637/jss.v080.i01.

[137] H. He and E. A. Garcia. "Class Imbalance Learning Methods for Support Vector Machines". In: *Imbalanced Learning: Foundations, Algorithms, and Applications*. Hoboken, NJ: Wiley, 2013. Chap. 5, pp. 83–99. DOI: 10.1002/9781118646106.ch5.

[138] W. S. Brum, N. C. Cullen, S. Janelidze, et al. "A two-step workflow based on plasma p-tau217 to screen for amyloid β positivity with further confirmatory testing only in uncertain cases". In: *Nature Aging* 3.9 (2023), pp. 1079–1090. DOI: 10.1038/s43587-023-00471-5.

[139] D. J. Figdore, M. Griswold, J. A. Bornhorst, et al. "Optimizing cutpoints for clinical interpretation of brain amyloid status using plasma p-tau217 immunoassays". In: *Alzheimer's & Dementia* 20.9 (2024), pp. 6506–6516. DOI: 10.1002/alz.14140.

[140] C. R. Jack, D. A. Bennett, K. Blennow, et al. "NIA-AA Research Framework: Toward a biological definition of Alzheimer's disease". In: *Alzheimer's & Dementia* 14.4 (2018), pp. 535–562. DOI: 10.1016/j.jalz.2018.02.018.

[141] T. Urbano et al. "Comparison of Serum and Cerebrospinal Fluid Neurofilament Light Chain Concentrations Measured by Ella™ and Lumipulse™ in Patients with Cognitive Impairment". In: *Diagnostics (Basel)* 14.21 (2024), p. 2408. DOI: 10.3390/diagnostics14212408.

[142] S. Abu-Rumeileh et al. "The multifaceted role of neurofilament light chain protein in non-primary neurological diseases". In: *Brain* 146 (2023), pp. 421–437. DOI: 10.1093/brain/awac319.

[143] P. Lopes das Neves et al. "Serum Neurofilament Light Chain in the Diagnostic Evaluation of Patients with Cognitive Symptoms in the Neurological Consultation of a Tertiary Center". In: *Journal of Alzheimer's Disease* 95 (2023), pp. 391–397. DOI: 10.3233/JAD-220727.

[144] B. Winblad et al. "Mild cognitive impairment – beyond controversies, towards a consensus: Report of the International Working Group on Mild Cognitive Impairment". In: *Journal of Internal Medicine* 256 (2004), pp. 240–246. DOI: 10.1111/j.1365-2796.2004.01380.x.

[145] American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders*. 5th ed. Washington, DC, USA: American Psychiatric Association Publishing, 2013. ISBN: 9780890425558.

[146] E. A. Willemse and C. E. Teunissen. "Biobanking of cerebrospinal fluid for biomarker analysis in neurological diseases". In: *Biobanking in the 21st Century*. Ed. by J. Merkle and A. Hofman. Cham, Switzerland: Springer, 2015, pp. 79–93. DOI: 10.1007/978-3-319-20579-3_7.

[147] H. Passing and W. Bablok. "A new biometrical procedure for testing the equality of measurements from two different analytical methods. Application of linear regression procedures for method comparison studies in clinical chemistry, Part I". In: *Journal of Clinical Chemistry and Clinical Biochemistry* 21.11 (1983), pp. 709–720. DOI: 10.1515/cclm.1983.21.11.709.

[148] D. Vecchio et al. "Serum and cerebrospinal fluid neurofilament light chains measured by SIMOA, Ella, and Lumipulse in multiple sclerosis naive patients". In: *Multiple Sclerosis and Related Disorders* 82 (2024), p. 105412. DOI: 10.1016/j.msard.2023.105412.

[149] A. Gauthier et al. "Comparison of Simoa™ and Ella™ to assess serum neurofilament-light chain in multiple sclerosis". In: *Annals of Clinical and Translational Neurology* 8.6 (2021), pp. 1141–1150. DOI: 10.1002/acn3.51355.

[150] Jean Daunizeau, Olivier David, and Klaas E. Stephan. "Dynamic Causal Modelling: A critical review of the biophysical and statistical foundations". In: *NeuroImage* 58.2 (2011), pp. 312–322. DOI: 10.1016/j.neuroimage.2009.11.062.

[151] Manuela Tondelli et al. "Resting-state networks and anosognosia in Alzheimer's disease". In: *Frontiers in Aging Neuroscience* 16 (2024), p. 1415994. DOI: 10.3389/fnagi.2024.1415994.