

Probing Mental Effort With Active Inference

Giuseppe Pagnoni

University of Modena and Reggio Emilia

24/06/2026, Napoli, Italy

What do we do when we “try harder”?

- Mental effort is a notion of frequent usage in daily life

What do we do when we “try harder”?

- Mental effort is a notion of frequent usage in daily life
- It is intrinsically related to several important areas of cognitive neuroscience
 - motivation and reward
 - attentional processes
 - willpower and self-control

What do we do when we “try harder”?

- Mental effort is a notion of frequent usage in daily life
- It is intrinsically related to several important areas of cognitive neuroscience
 - motivation and reward
 - attentional processes
 - willpower and self-control
- It has multiple meanings
 - a feeling of exertion, often of an aversive kind
 - the amount of mental resources *automatically* recruited by a difficult task
 - the capacity to *voluntarily* engage mental resources in a task

What do we do when we “try harder”?

- Mental effort is a notion of frequent usage in daily life
- It is intrinsically related to several important areas of cognitive neuroscience
 - motivation and reward
 - attentional processes
 - willpower and self-control
- It has multiple meanings
 - a feeling of exertion, often of an aversive kind
 - the amount of mental resources *automatically* recruited by a difficult task
 - the capacity to *voluntarily* engage mental resources in a task
- We still lack a clear mechanistic understanding of mental effort

A historical note

- In 1973, Daniel Kahneman published the seminal book *Attention and Effort*
- Mental effort was largely equated with a task's attentional demands: tasks requiring more attention were assumed to require greater mental effort
- No mention of the possibility of intentionally investing *more or less* effort in a task of a given difficulty

The active inference perspective

Whenever we engage in a demanding task, we face a tension between what we would tend to do naturally and what is required for a correct performance

In active inference, this tension is formulated as the *divergence* between two probability distributions over the possible courses of action (*policies*):

- the distribution that reflects our habitual tendencies
- the distribution that reflects the current, task-determined context

The larger this divergence, the greater the mental effort associated with task performance

Volitional effort

Researchers typically manipulate the degree of effort invested in a task by varying:

- the difficulty of the task
- the reward associated to a correct performance

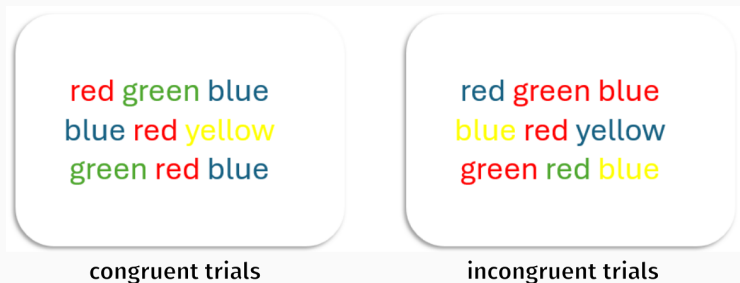
We were interested, however, in the capacity to **voluntarily** invest more or less effort, independently of external rewards or task demands

More precisely, we wanted to study the **effect of instructions** about effort: what do we do when we are *asked* to “try harder”?

This is a question of practical relevance and currently a much understudied topic

Experimental protocol

The **Stroop task** (color-word interference task) is designed to trigger an automatic response that participants may need to override. There are two types of trial:



where participants are asked to report either the word text or its font color

Since we tend to read words automatically, reporting the font color in incongruent trials **requires effort**

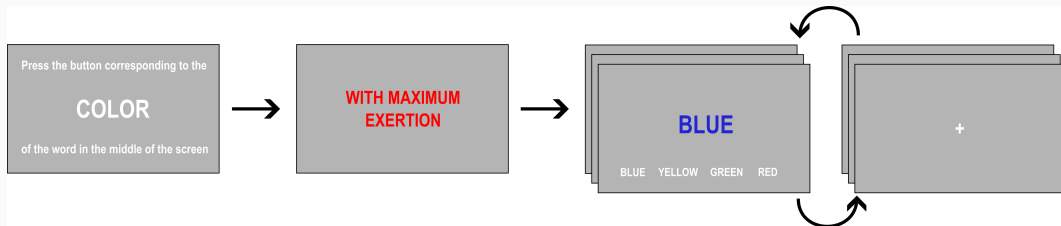
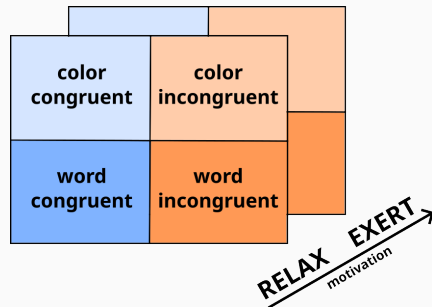
Study design

20 healthy subjects (12 females; age 27.9 ± 5.7)
were asked to perform the task alternately:

- with maximum effort (**EXERT**)
- as relaxed as possible (**RELAX**)

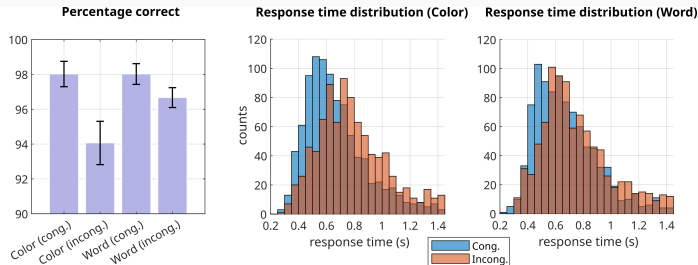
and to report either:

- the **word** text
- the font **color**

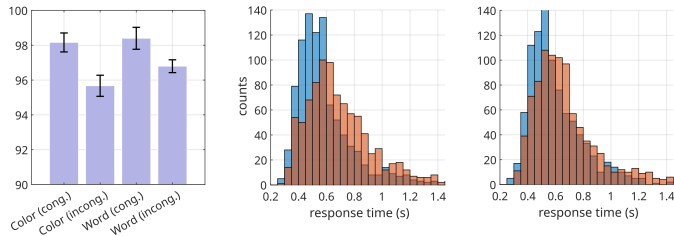


Behavioral data

RELAX



EXERT



The *Color incongruent* condition showed the worst performance (accuracy and RTs) in both effort conditions

Responses were generally faster under the EXERT condition

Motivation and habits

In the AIF perspective, **effort** is construed as the divergence between the **required** behavior (given the context) and the **habitual** behavior

For example, in the **Color incongruent** condition of the Stroop task, the required behavior is stating the font color and the habitual one is reading the word text

Motivation and habits

In the AIF perspective, **effort** is construed as the divergence between the **required** behavior (given the context) and the **habitual** behavior

For example, in the **Color incongruent** condition of the Stroop task, the required behavior is stating the font color and the habitual one is reading the word text

Two key model parameters underwrite this:

- **preferences c** $\rightarrow$ motivation to perform well
- **habits e** $\rightarrow$ tendency to respond to text rather than color

Motivation and habits

In the AIF perspective, **effort** is construed as the divergence between the **required** behavior (given the context) and the **habitual** behavior

For example, in the **Color incongruent** condition of the Stroop task, the required behavior is stating the font color and the habitual one is reading the word text

Two key model parameters underwrite this:

- **preferences c** $\rightarrow$ motivation to perform well
- **habits e** $\rightarrow$ tendency to respond to text rather than color

Does the volitional investment of effort modulate c , e , both, or neither one?

A 2-level AIF generative model

The upper level represents states that evolve over a **slower** time scale (trial-to-trial):

- **instruction**: respond to color or text
- **task sequence**: instructional or response context
- **modality**: response strategy selected

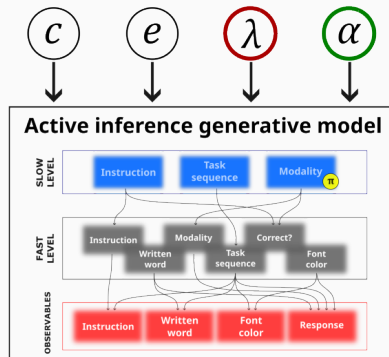
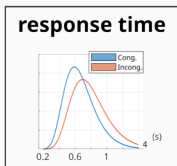
The lower level represents states that evolve over a **faster** time scale (single trial); in addition to receiving the outcomes of the upper level states as its own state factors, it includes:

- **color** of the font
- **text** of the word
- **correctness** of the chosen response strategy

The model layout: causal dependencies

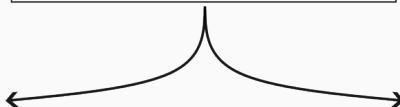
λ = inverse temperature

smaller values make the choice probability distribution more uniform



α = individual response speed

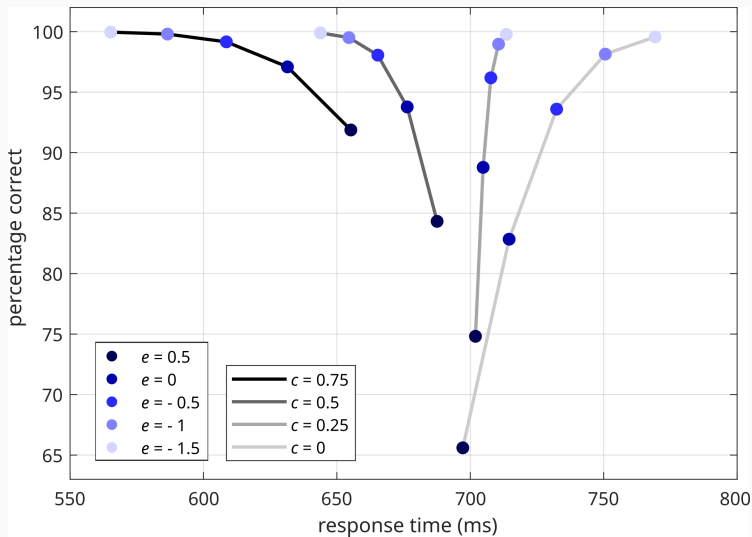
expected RT of the participant when the entropy is zero (no indecision about the choice)



choice

Stimulus	Response
green	red
blue	blue
blue	green
...	...

Reproduction of real-data features: speed-accuracy trade-off



Model inversion (single-subject fitting)

Response choice is sampled from a distribution $\mathbf{u}$ of the expected free energy under different policies $\mathbf{G}(\pi)$:

$$\mathbf{u}_t = \text{softmax}(-\lambda \mathbf{G}(\pi_{t+1})) \quad \lambda \in \mathbb{R}^+$$

Response time depends on the entropy $H_t = -\mathbf{u}_t \cdot \ln \mathbf{u}_t$:

$$r_t \sim \exp(\alpha + H_t + n) \quad \alpha \in \mathbb{R}, n \sim \mathcal{N}(0, \frac{1}{16})$$

As a result, the **log-likelihood** can be written as:

$$\mathcal{L} = \underbrace{\sum_t \ln(o_t \cdot \mathbf{u}_{t-1})}_{\text{choices}} - \underbrace{\sum_t \frac{1}{16} (\ln r_t + \mathbf{u}_t \cdot \ln \mathbf{u}_t - \alpha)^2}_{\text{response times}}$$

Model inversion (single-subject fitting)

Model inversion was performed by Variational Laplace on each individual's data

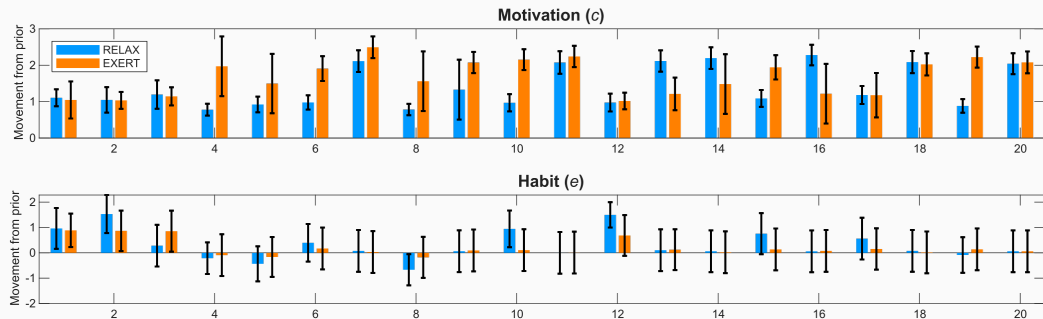
The model was fitted separately for the EXERT and RELAX blocks

The prior means for c and e were kept equal for both blocks¹:

- **prior preference** c_{prior} $\longrightarrow$ the agent assumes being correct is $\exp(2) \approx 7.4$ times more probable than being incorrect
- **prior habit** e_{prior} $\longrightarrow$ the agent expects to respond 65% of the time to the word text and 35% of the time to the font color

¹ $C = \exp(c)(-1, 1)$, with $c = 0$; $E = \exp(e)(-0.3, 0.3)$, with $e = 0$

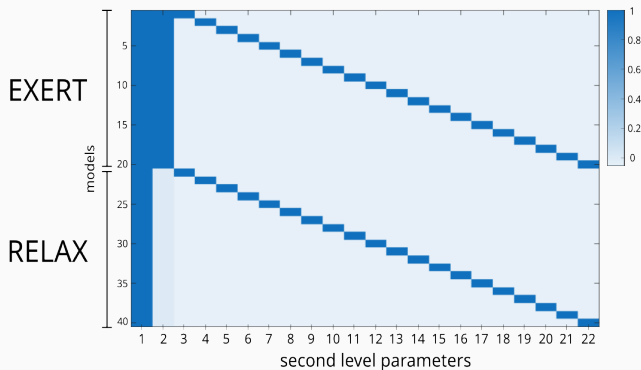
Individual posterior estimates for c and e



Group-level analysis: Parametrical Empirical Bayes (PEB)

To assess whether the effort manipulation systematically affected the estimates of c and e across participants, we used the PEB approach

This involves specifying a design matrix accounting for the sources of variability in the data (parameter estimates) across conditions and participants:

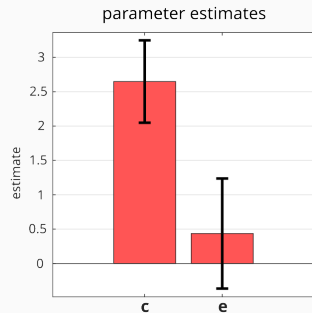
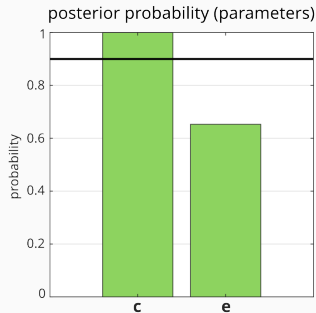
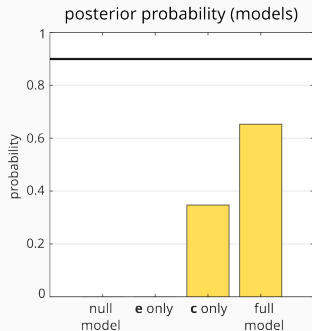


- The first column contains all 1s and represents the intercept
- The second column refers to the effect of the EXERT condition
- Each other column represents a specific participant

Group-level analysis: Bayesian Model Reduction (BMR)

Comparing the evidence for reduced versions of the full model obtained with PEB

Group-level effect of effort (EXERT > RELAX)



⇒ effort has a clear impact only on the **motivation** parameter *c*

Conclusions and future directions

All incongruent stimuli in the Stroop task require a degree of effort not to follow the automatic tendency to read the text

When we are asked to “try harder”, what we actually do may be to **increase endogenously our motivation** — maybe akin to setting up a *virtual* reward

Potential extensions of the approach:

- Clinical: patients with apathy (e.g., Parkinson's Disease)
- Educational: volitional effort engagement and learning trajectories
- Workplace: prediction of burnout vulnerability

Thank you for your attention!

Riccardo Maramotti
Thomas Parr
Daniela Ballotta

Manuela Tondelli
Giovanna Zamboni
Sanjay G. Manohar

Published article: <https://doi.org/10.3758/s13415-026-01419-z>

