Modelling meta-awareness and attentional control with deep parametric active inference : a computational neurophenomenology account to mental action

Romy BEAUTÉ
École Normale Supérieure Paris-Saclay (MVA)
romy.beaute@ens.psl.eu

***Keywords*—** Meta-awareness, Deep active inference, Hierarchical modelisation, Neurophenomenology, Inferencial architecture, Bayesian models, Generative model, Mind-Wandering, Mind-Blanking, Attentional control, ADHD

Code of this project available here

Abstract

The project proposed here is based upon the work published by Sandved-Smith et al. [23] entitled "Towards a computational phenomenology of mental action: modelling meta-awareness and attentional control with deep parametric active inference". It presents a computational approach for the study of the inferential architecture and processes required to get access, control, and evaluate cognitive states at different levels, that are thought to modulate perception and behavior. More specifically, the project focuses on the relationship between meta-awareness abilities (aware of being aware : higher-level cognitive states), deployment of mental actions abilities (policy selection), and perception of lower-level cognitive states (such as attentional and perceptual states). Through a computational implementation of a cognitive agent that is endowed with an hierarchical and inferential architecture, we attempt to 1) simulate and understand the mecanisms underlying the emergence of specific phenomenological components, 2) apply constraints on some parameters to account for an attentional deficit disorder (ADHD) and observe phenomenal, behavioral and architectural variations. Following the work of Sandved-Smith et al., [23], the phenomenal properties that will be simulated and addressed refer indeed to the ability to gain access (opacity) and control (agency) over lower-level cognitive states (attentional and perceptual). Access then controllability over state will in turn allow 'effective self regulation', to obtain the desired outcome. This work present an interesting avenue for providing more explanatory power, understanding and for providing a broader and more reliable access to the underlying content (phenomenology) and structure (cognitive architecture) of meta-awareness and mental actions.

Finally, the objectives of this project are to I) establish a thoughtful reading of the paper of Sandved Smith *et al. [23], II) provide a description of the proposed framework's components by explaining the different principles, their interactions and consequences ; III) reproduce their results ; IV) propose an extension of the work by simulating an agent with and without pathological condition - Attention deficit hyperactivity disorder (ADHD). We will use the latter point to illustrate the use of such approach in practical, clinical context.*

1. Introduction

Understanding the dynamical mechanisms and the phenomenology of perception, **cognitive control** and **cognitive awareness** processes is an important challenge to address for the study of human experiences and cognitive architectures. The need for the understanding of the mechanisms underlying and influencing the emergence of certain phenomenal properties extends beyond the question of the hard (or meta-) problem of consciousness [4] per see, as such knowledge might help addressing important clinical and ethical issues outlining to conceptions of health and pathology [1, 19, 24], especially in the growing field of computational psychiatry [13].

Yet, several shortcomings remain in the understanding of the relationship between causal phenomenal traits of perceptual experience (1st person) and their physiological substrates (3rd person) [32, 12], and such gap in knowledge is often due to methodological limitations.

In this work, we emphasize on the importance to understand the mechanisms underlying the emergence of a specific type of phenomenal properties : the ”opacity” of a cognitive process - referring the ability to access the causes that generated the current observation (latent content)-. The importance of studying the mechanisms underlying emergence of opacity, and their consequences is justified by the fact that opacity enhance controllability over cognitive states, through mental actions. This in turn provides elements of understanding regarding the capacity and decision to engage in mental actions, which in highly relevant notably for the field of clinical neurosiences. Indeed, dysfunctions of attentional control is thought to play a key role in several psychiatric conditions, linked with atypical inferences about states, and erroneous control over associated processes. Therefore, we suggest that providing a realistic computational model accounting for the generation and modifications of internal states could help understanding both phenomenal and architectural manifestations of such dysfunctions.

Following this introduction, we will now describe and analyse the main concepts that are necessary to understand the contributions brought by this article, and the scope of the present project. In a second part, we will give a full description of the model architecture. In a third part, we will explain the simulation done with the present model implemented in this paper.

2. Description of concepts and foundations

Recent decades have brought a paradigm shift in computational and cognitive neurosciences towards a **Bayesian conception of the neural operations** accounting for perception, cognition, action, and consciousness [26, 9]. Before continuing on the analysis and extension of the work of Sandved-Smith et al., [23] we will provide with a more in depth comprehension and analysis the concepts, terminology and mathematical foundations of **active inference** and **computational phenomenology**.

According to a **Bayesian conception of the neural operations** accounting for perception, cognition, action, and consciousness, the brain uses its encoded knowledge (Bayesian beliefs, or priors) about the world to **actively** perform **predictive inferences** about sensory inputs. Such inferences about a given input are shaped through minimizing the prediction errors (or surprise) associated with the experience of this input. Formally speaking, this means that the brain will try to minimize the **"variational free energy"** F, an information-theoretic quantity that quantifies the discrepancy between the observed sensory outcomes (latent) and the outcome that would be predicted/expected from the generative model (hidden). This assumption comes from the fact that the brain is a biological agent and therefore needs to maintains the body in an optimal physiological state (homeostatic regulation) by choosing the most appropriate action (policy selection). This constraint also implies that the brain will need to optimizes its computation for policy selection, in order to minimize energetic cost and memory overload. It is also important to note that the experience of a specific stimuli is **enacted** and **embodied** in a dynamical environment, i.e in which the agent engage actively and directly [30] to make sense of it. The importance of taking into account the influence of enactive and embodied properties on the perceptual outcome is depicted in Fig. 3. Embodiment and enactive properties indeed imply the need for considering and implementing fluctuations of the external world for modelling policy selection, perception, and cognitive processes more generally. In this respect, the present work integrates previous theoretical and empirical work by providing an extension of a Bayesian model based on deep active inference framework [11], taking into account phenomenal, embodied, enacted and inferential properties of cognitive processes. Precisely, this framework suggests that the opacity of a set of states (i.e awareness of these states as being mental states per see) corresponds to a "second-order inference" about a quantity informing about the reliability of an information that is encoded in its associated parameter called **"precision"** [14, 15], all this in an embodied and active agent. In this section we will provide a description of the main concepts along two axes: **Computational (neuro)phenomenology** framework and **Active Inference** framework.

2.1. Computational phenomenology of mental action & metacognition

As highlighted above, our ability to understand the nature and mechanisms of the processes underlying facets of phenomenal experience, such as cognitive awareness and control, and their variation remains largely limited. This gap in knowledge can be partially attributed to the fact that current computational models of mental processes do not readily account for richness and range of perceptual phenomenology. Obstacles to the study of covert mental processes include the low reliability and availability of subjective first-person reports of experience [20], difficulties in establishing the influences of environment, habits and context on conscious experience [33], as well as the lack of effective ways for studying the relationships between contents and structural features of conscious experience. These challenges then translate into limitations in studying the relationship between experiential (subjective) phenomena that is measured through "first-person" data, and neurobiological (objective) phenomena, measured through "third-person" data [21]. In this work, we are interested in describing and using a generative model of inferential architecture (based on the active inference framework) for the study of meta-awareness and control of attentional states. The claim is that such approach, based on mathematical and modelling tools taken from the active inference framework, could provide a way to reduce methodological challenges. Indeed, the architecture proposed here allow to model higher-level abilities in an simulated agent, that has (mental covert) control abilities over their attentional processes. Such architecture is structured in three nested and hierarchical levels, in which distinct phenomenal (perceptual) and cognitive properties are described. These three levels are perception of the environment (1st, lower-order level); perception of attentional state (2nd level); and perception of meta-awareness of attentional state (3-rd, higher-order level). Through simulation of this architecture, we will be able to model some phenomenological features of states, and their hierarchical relations.

Formal neurophenomenology (NP)

The goal of developping a formal neurophenomenology (NP) approach is to formalise aspects of phenomenal (lived) experience, aspects that are revealed by 1-st person data (obtained from subjective and descriptive reports). By promoting an understanding of first-person experiential dynamics, the NP approach - in Neuroscience of Consciousness research - relocates the focus by studying the content of consciousness - referring to the ”What it is like” [18] of experience - and their associated correlates. The significance of this shift from a more quantitative study of consciousness (access consciousness) is that it may improve our understanding of the fundamental aspects relating to structural and reflexive processes of consciousness and cognitive processes in general by focusing on how process-related changes in the structures of experience are associated with changes in the organism.

By recognizing the importance of phenomenal perspective and tempting to formalize it, Sandved-Smith et al. [23], propose a model of cognitive control, relying on a parametric deep active inference architecture, in which levels of state inference are implemented at different hierarchical levels. Such modelisation of attentional control bring three main contributions :

1. it enables to define attentional and meta-awareness states as ”mechanism of opacity control at different hierarchical levels”;
2. it provides a formal definition of attentional and meta-awareness control via an ”account of the higher-level policy selection”;
3. it formalizes the relationship between meta-awareness and attentional control.

Meta-awareness states & phenomenal opacity

Meta-awareness of a state refers to the ability of being explicitly aware of one’s own awareness of the current content of a state, or of a conscious event. It is a high order cognitive process (awareness about awareness), that generate outcomes (mapping precision, beliefs) at a lower level. Note that we refer to this definition on the context of this project, while being aware that there is still no universal definition of meta-awareness, topic of ongoing discussion. Regarding the study of meta-awareness, one advantage of the model described, replicated and extended through this work is to account for nuances in the definition (eg continuity of meta-awareness state), nuances that are notably supported by the notion of phenomenal descriptions mentioned above. In order to formalise and implement meta-awareness abilities in an agent, we can translate the definition in more computational terms, we can define meta-awareness as a (higher-order) process that enable one to become explicitly aware (opacity) of its current conscious content (content of lower-order). The notion of opacity (vs transparency) of a state will be useful for the fomalization of meta-awareness mechanisms. The concepts of “phenomenal opacity” and “phenomenal transparency” have indeed been theorized by Thomas Metzinger [16] as an instrument for analyzing self-representational conscious content and their relevance in understanding cognitive (reflexive)

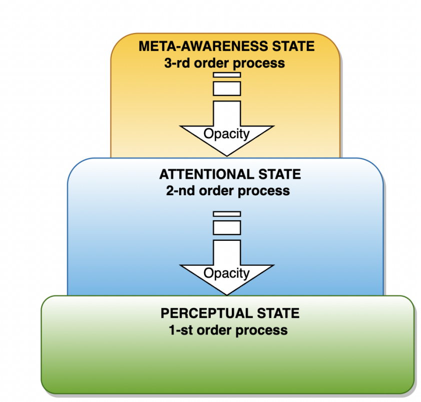

Figure 1. Perceptual, Attentional, and Meta-awareness states represented as hierarchical levels. Higher order levels enable a process for becoming aware of the lower level, making it **’opaque’** to the system.

subjectivity. These notions can be used to characterise (covert) mental states according to their perceptibility as being constructed by one’s mind or not. According to this framework, a state is considered ’opaque’ if the cognitive constructive processes underlying them are perceptible (meta-awareness). Conversely, ’transparent’ states characterise mental states for whom we are not aware of their constructive cognitive processes per see, but only have access to the content they make available. State opacity therefore signifies that knowledge about the experience of this state, and will therefore increase the abilities to control over mental processes or to update beliefs.

2.2. Active inference in the brain

Active Inference Framework

Active inference framework treats cognitive processes and their interactions as interdependent forms of inference. Namely, this framework consider the brain as a Bayesian agent that combines **prior beliefs** about sensory inputs to infer the probability of different states and events in the environment. Beliefs can be understood as conditional expectation about a given sensory input or event, associated with a probabilistic representation and distribution encoded by neuronal activity [14]. In this sense, beliefs correspond to active inference about sensorium or events, constructing prior expectations about precision [15]. The probability distribution of a possible state (i.e. prior belief) for a given observation embedded in the environment can therefore be formalized using the Bayes’ theorem :

$$p(s|o,m) = \frac{p(o|s,m)p(s|m)}{p(o|m)}$$

where s denote the possible ***states*** *of the agent, m the model of the environment, and o the observation or outcome. The likelihood $p(o|s,m)$ specifies how states generate observations o embedded in a model of the world m.*

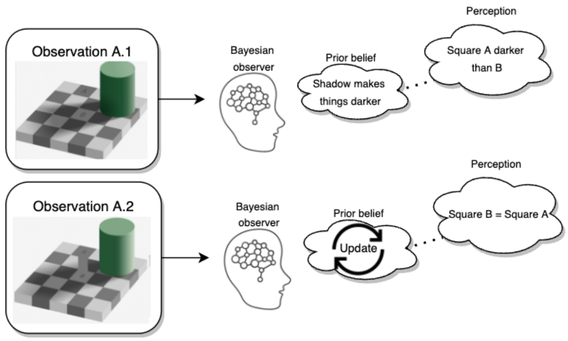

Figure 2. Perception of visual illusion, in light of Bayesian inference framework. This example illustrates how beliefs about the world we live in influence (unconsciously) our perception. For the observation A.1) Here the (unconscious) assumption that there is a shadow behind the green block lead the cognitive agent to perceive the square A as darker than the square B. For the observation A.2) a bar connecting the two squares (new element) leads to an update of the prior belief, which will reshape the final perceptual outcome : the two squares are now perceived as being of the same colour. Visual illusion , and illusion in general are highly robust examples reinforcing the Bayesian inference framework, considering that we do not perceive the ”true” properties of the world, but a statistical representation of it.

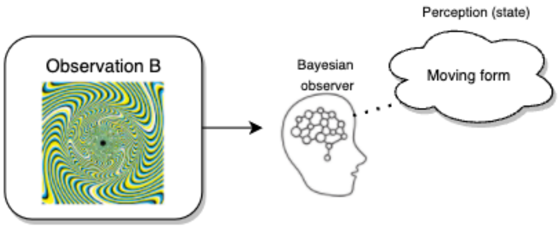

Figure 3. Perception of visual illusion, in light of Bayesian and active inference framework. As our eyes move on and around the coloured form, we perceive the static form as moving. This example illustrates the importance of considering enactive and embodied properties of our cognition, as our cognitive processes and perception are actively constructed in a dynamical world.

Variational inference

The goal of a ”Bayesian observer” is to make inferences about how the states s of the world have changed based on new observations/outcomes o. This corresponds to find the posterior distribution $p(s|o)$, ie trying to infer the most probable cause s of the observation o. The quantification of the relationship between observations o and the states s that caused them is encoded in the likelihood mapping matrix A. However, this implies the calculation of the - generally intractable - marginal likelihood $p(o)$. Variational inference will answer this problem by turning this inference problem into a optimization one, by seeking to optimize the so called ”variational posterior” (auxiliary distribution) $q(o)$ that will help approximating the true posterior $p(s|o)$ using the Kullback Leibler (KL) divergence. KL divergence will serve as a measure of the relative distance between the true posterior distribution $p(s|o)$ and its associated variational posterior distribution $q(o)$ used to approximate it. In this model, we will refer to this quantity as **”Variational Free Energy” (VFE)**, which, by providing computationally tractable quantity, will enable us to approximate inference.

We note that active inference models are thus characterized as **generative** as they seek, by accumulating beliefs about the world, for probabilistic (predictive) mappings from causes (latent or hidden dynamical states of the world) to consequences (observed outcome given the sensory state of the organism and the world it is interacting with and embedded in) [8]. Moreover, we can note that active inference models are an extension of basic generative models in the sense that they account for knowledge about temporal dynamics of states. Inferences (beliefs induction) about state transitions can be modelled as a Markov chain.

In a nutshell, two main assumptions are the conceptualisation foundation of the active inference framework :

- **Enaction & Embodiment** accounting for the active engagement of an agent in a dynamical world.
- **Bayesian inference** accounting for the statistical procedure performed by the brain to update its own beliefs about (both external and internal) world.

Beliefs and Free energy principle

The free-energy principle assumes that everything in the brain should change to minimize the amount of prediction error (i.e minimize the free-energy). This minimisation can be done by adapting perceptual representations so that they approximate a posterior or conditional density on the causes of sensations. In other words, perception reduces free-energy by changing predictions [10], given the prediction-errors. Prediction-errors store the inconsistency between sensory input (i.e true state of the world) and current beliefs about these inputs. The information conveyed by the prediction-errors lead the system to update or create new beliefs. This means an adjusted probability distributions of the beliefs to be more consistent with the true value of sensory input, and therefore minimize error.

Deep generative model A generative model is a robust way of learning the data distribution using unsupervised learning. More specifically, they learn the ”true distribution” of the (training sample) data and generate new data values with some variation. Given the limitation in accessing the true distribution of the data we are interested in, we need to model a distribution which is as akin as possible to the true data distribution.

3. Probabilistic Graphical Model of mental action : Active Inference architecture

In this work, we show the replication of the deep active inference architecture proposed in [23], in order to model mental action, meta-awareness and attentional control. We will also present an extension of this model, by adding constraints on the architecture in order to simulate an agent that present an attentional deficit disorder (ADHD). Structural and reflexive properties of cognitive control and awareness are modeled with respect to an extended version of the traditional active inference framework. This approach also extends a more general utilisation of Bayesian frameworks (eg. modeling perception), in the sense that it models (dynamical) inferences of optimal actions that should be elicited to achieve one specific goal. Within the active inference framework, a generative model implements policies π, corresponding to a possible sequence of actions u that are susceptible to be selected. When a specific policy is selected, generative process will be updated to change the ”true state” of world through action. Including policy selection in the generative model implies that the generative model will be extended as follow :

$$p(o, s, \pi) = p(o|s, \pi)p(s|\pi)p(\pi)$$

where $p(\pi)$ represent the prior over a specific policy (preference, or bias towards a specific action to choose) ; $p(s|\pi)$ encodes the beliefs (priors) about the true state (s) of the world given the selected policy (π) ; and $p(o|s, \pi)$ the observations expected given the joint distribution of policies and state priors. Policy values are specified in a probability distribution, corresponding to the probability of a given policy of being selected. This probability is directly related to the **’Expected Free-Energy’ (EFE)**, corresponding to beliefs about how likely it is that a policy will generate preferred outcome.

3.1. Hierarchical architecture

We now describe the architecture of the model, that consists in three nested levels enabling the modelisation of different levels of mental action. Such architecture will help to investigate the role of meta-awareness (enabling opcity of lower-order states) on policy selection, update beliefs, and such repercution on lower-order states. A full pictural representation of this architecture, is given in Fig. 4 for clarity purposes.

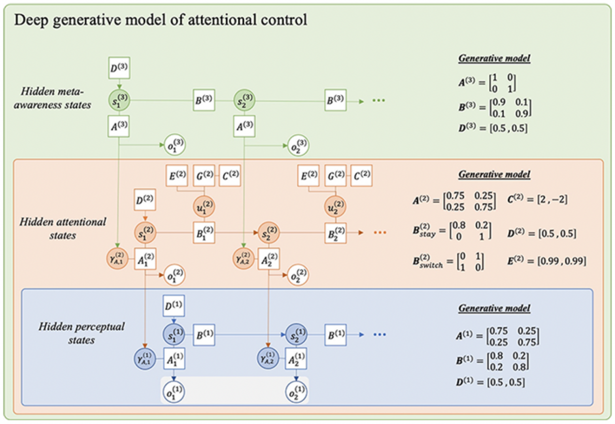

Figure 4. Deep generative model of attentional control, with the three hierarchical levels nested in a probabilistic graphical model. *Extracted from [23]*

3.1.1 First-order level : perception of the external environment

The first-order level, noted $s^{(1)}$ simply refer to the lowest-order cognitive process : sensory observation. In the simulation (see 4), it corresponds to the observation of the stimuli as being either ”deviant” or ”standard”.

3.1.2 Second-order level : perception of internal attentional states

In light of the model presented here, internal attentional (second-order) states, noted $s^{(2)}$ enable the ’opacity’ of perceptual (first-order) states described above. Attentional states are implemented as a second level state inference $s^{(2)}$, as they will directly affect and modulate the confidence in the lower-level state (here sensory observations). As $s^{(2)}$ conditions the precision of the lower-order likelihood mapping (here $A^{(1)}$), this allow us to simulate how changing parameters of this second-order $s^{(2)}$level will affect the confidence and (latent) perception of sensory observations. In the simulation (see 4), it corresponds to the observation of the attentional state as being either ”focus” or ”distracted”. With respect to the NP framework, we note that such implementation enables distinguishing different phenomenal stages of agent’s attentional states. These different states are the following :

- ***State of focus***
  Corresponds to a higher order attentional state, in which the agent is focus

- ***State of distraction (mind-wandering)***
  Corresponds to a state were the agent is unaware of being distracted. Unawareness can be associated with a lack of higher order observations (low precision evidence) collected by the agent, that do not perform mental actions at the second level (low perceptual demands).

- ***State of awareness of distraction***
  State in which the agent is distracted, but becomes aware of such distraction. This suggests the incitement of a policy selection in favour of a return to focus state. This bias can be formalised and implemented in the architecture as a (prior) preference over a policy in favor of a focused state. These belie (corresponding to the $C^{(2)}$ matrix). More generally, $C^{(2)}$ encodes the policy preference (bias) towards a given (internal or external) instruction.

- ***State of redirection of attention (transient stage)***
  This state often follows the state of awareness of distraction, as it corresponds to the transient state of redirection of attention.

The second (hierarchical) level therefore extends the previously described architecture by enabling the implementation of a (directed) mental action depending on the attentional state. The interaction between the different stages of sustained attention described above are summarised in Fig. 5.

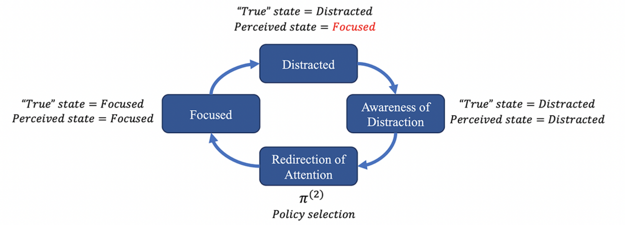
“True” state = Distracted
Perceived state = Focused
Distracted
“True” state = Focused
Perceived state = Focused
Focused
Awareness of Distraction
“True” state = Distracted
Perceived state = Distracted
Redirection of Attention
$\pi^{(2)}$
Policy selection

Figure 5. Formalisation of the phenomenal cycle of sustained attention. *Extracted from Sandved-Smith et al., 2021* [23]

3.1.3 Third-order level : perception of meta-awareness states

For an agent to voluntarily control its attentional state, a necessary (but not sufficient) condition is to be explicitly aware of it. This type of perception refers to states of meta-awareness. This constraint justify the implementation of a third-order level, noted $s^{(3)}$, in the deep generative model of meta-awareness and control (volition) of mental action. In light of the model presented here, we can define meta-awareness states as third-order states that enable the 'opacity' of internal attentional states of second-order described above. By using three-levels, this architecture provides a modelisation of higher-level perceptual processes by giving the cognitive agent control over its own attentional processes. In the simulation (see 4), it corresponds to the observation of a high awareness of the attentional state, or to a low awareness.

4. Experiment : replication & extension

This section is dedicated to the description and analysis of the model's experimentation (inferential architecture), through a simulation of belief updating during an attentional task. More specifically, it presents the results regarding the emergence of opacity-transparency phenomenological properties in the model. After replicating the results shown in the paper, we will present our extension of the model, simulating an agent with or without Attentional Deficit Disorder (ADHD) condition.

4.1. Replication

We first replicated the results presented in [23] (subsection 4.1), with the same parameters, in order to be more confident in our interpretation when further changing other parameters for the simulation of ADHD condition (subsection 4.2). In the second part (subsection 4.2), the parameters of the generative model will indeed be changed in order to model adaptive behavior in a agent doted with different physiological abilities (attentional deficit).

4.1.1 Influence of attentional states on lower-order (perceptual) states - non ADHD condition

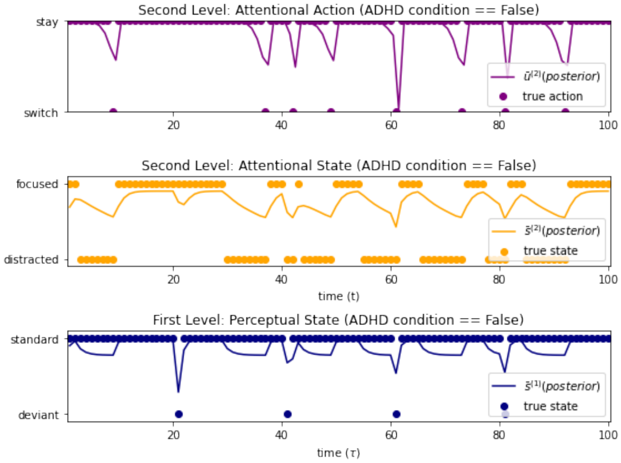

Figure 6. Replication in non ADHD state. Attentional states during the first and second part of the experiment have been respectively set to "focus" and "distracted".

We first replicated the figure 10 of the original paper [23], i.e simulation of attentional cycle on normal (non ADHD) condition. At each time step t, the active inference agent is inferring its own current perceptual (stimuli deviant *vs* standard) and attentional (focus *vs* distracted) states.

This enabled us to look at the influences of the attentional state (6, middle figure) on the action decision to stay or switch in the opposite state (top figure), and on the confidence to detect deviant vs standard stimuli (6, bottom figure). We specifically note that when in focus state, the agent tends to stay focus longer (more stability of focus state, less switching to distracted state). When focus, the agent also infer with a higher probability the state of the stimuli (more confident). However, when the attentional state of the agent goes to distracted, we can observe I) that the agent tries to switch to a attentive state, and II) that the precision decrease regarding the detection of deviant stimuli. The change back to focus state is efficient when the agent infer a distracted state with a high probability.

4.1.2 Influence of attentional states on lower-order (perceptual) states - non ADHD condition

We also replicated the implementation of the 3-rd order level : meta-awareness cognitive process. This enable to observe the influence of the meta-awareness ability on lower-level states (sensory and attentional). Figure 7 clearly shows the impact of meta-awareness on lower-order processes. We notice than when meta-awareness is high, the agent maintains focus for a very long time, and have a high confidence over the sensory perception of the deviant stimuli. However, we see a clear decreases in lower-states confidence when the meta-awareness is low : deviant stimuli detected with less confidence (or not even detected), and the agent has more trouble returning in a focus state. We interpret this difficulty in returning to a focus state coming from the lack of awareness of being distracted, and therefore an absence of mental action that would modify the current attentional state. In other words, as long as the agent is not conscious of being distracted, it is less probable that it will elicit a new policy selection to change the current state and therefore obtain the desired outcome. With low meta-awareness, the accumulation of state evidence is indeed longer as the agent is not consciously perceiving the current attentional state. This in turn results in a longer state of mind-wandering.

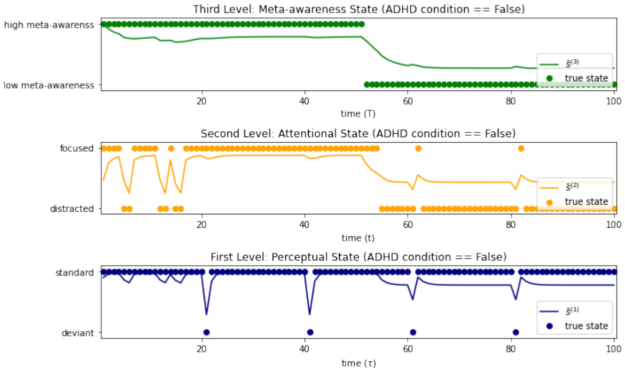

Figure 7. Replication in non ADHD state. Implementation of the 3-rd state level (meta-awareness level)

4.2. Simulating Attentional Deficit condition

The replication presented above (in. subsection 4.1) implies that the expected precision of the action model will shape agent's inference about current and future states, at different hierarchical levels. We recall that active inference model enable the inclusion of factors that modulate decision-making regarding a choice of possible actions (ie policies, notes π), namely policy selection. This opens the door for implementing agents with different behaviors, behaviors that are enhanced by habits (plasticity) and/or phenotypical traits. Attentional deficit-hyperactivity disorder (ADHD) is a psychiatric condition characterized by difficulties in staying in a focus state, inattention of the environment, often resulting in impaired perception, increased distractibility and behaviour [3]. In this part, we will introduce constraints on the architecture of the agent, to account for attentional deficit in a agent presenting a attentional deficit disorder (ADHD). In a second time, we will simulate how such agent performs in the same oddball paradigm task we previously reproduced, to observe the consequences of attentional constraints on lower-order observations.

4.2.1 ADHD architecture : Parameters changes

In ADHD, we note an twofold influence on habits (higher probability of selecting a specific attentional scheme, ie higher prior towards distracted state); and on phenotype (lower ability to select preferred policy C, ie maintain focus state).

Simulating an ADHD agent therefore implies modifying different parametets, as the attentional deficit will have an impact on different structures. Here, we will focus on the impact of two variables : 1) Parameter encoding **prior beliefs about policies**, to account for the phenotypical bias towards an unfocused state, resulting in a increased habit on such attentional state; 2) Parameter encoding for precision; to account for a decrease precision if the agent did not paid attention, or is used to not pay attention. To go further, we note that we could also modulate the parameter encoding for preferred outcome (C), to account for motivational factors regarding the task. To formalize and manipulate preferences, we encode them in a prior probability distribution, namely ‘prior preference distribution’ $p(o|C)$. Importantly, and following literature assumptions on active inference, prior preferences can be interpreted as encoding observations that are implicitly ‘expected’ by the agent with respect to its phenotype (eg. evolutionary preferences).

- ***C : Preferred outcomes***
  In C, we encode preferred outcome towards ”focus” states. Such preferences account for the state that is favored regarding the phenotype. Although there is not a clear understanding between the mapping of psychological and bayesian preferrences, we could formulate the hypothesis that C (bayesian preferences) could encode some factors to be interpreted as motivational (psychological preferences), as one could suppose that motivation is a component known to be in line with phenotypical preferences, and an important element in evolution. In ADHD, patients are associated with brain dysfunctions in fronto-cortical and fronto-subcortical systems, mediating the control of cognition and motivation [5]. Indeed, motivation deficits have been largely associated as a core component underlying the disorder. However, it is still unclear how motivation deficits occur and how they change over time [29]. Here, we could expect the possibility of interpreting and using the C matrix to model intrinsic motivation of the agent. Varying the values of the C parameters could be a way to quantify the consequences of a hyper of hypo motivation. Indeed, observations that have the higher probabilities will be considered as more rewarding [27].

- ***E : Prior beliefs about policies***
  To model the influence of phenotypical habit (bias towards distraction), we chose to encode prior beliefs about policies $p(\pi)$ in a vector noted E, each row corresponding to one policy. This will result in a constraint towards the realisation of the preferred policy (i.e preference towards ’focus’ state, encoded in C), as this will increase the probability of being distracted. We purposely chose to encode this bias towards distraction in a different parameter (E) than the one encoding preference towards a specific policy (C) in order to account for the internal dissociation encountered by a ADHD agent that actively and hardly tries to focus, but that is slowed down because of phenotypical constraints.

By modifying the parameters, we simulate an agent with an ADHD-like (phenomenal and phenotypical) priors, ie priors in favour of distracted states. Here, we observe an influence of the

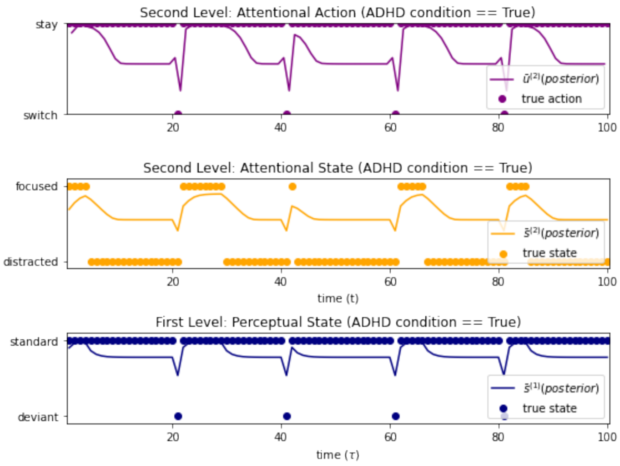

Figure 8. Simulation of ADHD agent with parameter's modifications, and effects on attentional (S^2) and preceptual (S^1) states.

ADHD-like priors at different levels. At the first level (perceptual), we see an overall decreased in the confidence over the perceptual observations (deviant vs standard). At the second level (attentional), we see a higher time latence between distracted state and focused state : it takes more time to go back to a focus state, and the maintenance in distracted state is longer. Conversely, the agent has trouble maintaining a focus state, and is rapidly attracted in the focus state.

5. Discussion

The work presented here proposed a description of the use of active inference framework used in [23], and described how this can be an interesting model to formalize specific phenomenal properties (here transparency and opacity of cognitive states, with respect to attention, and meta-cognition states). We also went further, by simulating an agent with and without an attentional deficit (ADHD) condition, to observe how the 'Balyesian beliefs', formalised as encoded in probability distributions, attentional state, and meta-awareness modulate both mental (covert) action (policy selection) and perceptual outcome (confidence during experimental task).

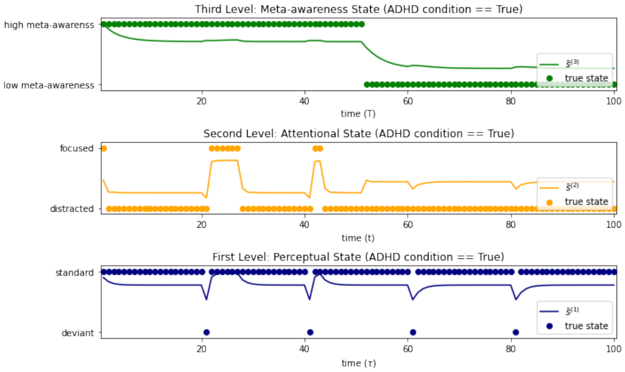

Figure 9. Implementation of the 3-rd state level (meta-awareness level). Simulation of ADHD agent with changing levels of meta-awareness and with parameter's modifications. Observation of the effects of meta-awareness level ($S^{(3)}$) on attentional ($S^{(2)}$) and preceptual ($S^{(1)}$) states.

5.1. Openings

Mind-Blanking : towards a finer-grained phenomenology of distractive states

The model presented here is a very general framework, and has been notably used to model attentional processes [23], by implementing high-order aspects of perceptual inference (awareness and meta-awareness). Given the generality and adaptability of the model, a natural expansion of this model could be to implement other aspects of mental actions. This would allow to a finer-grained understanding of the different aspects of phenomenology of mental actions in terms of inference. It would therefore be interesting to discuss and attempt to explore other features of mental action that might help expanding the model. Further research on the consequences of implementing a finer-grained phenomenology of mental action, based on this this model would also be interesting [31].

As a first step towards an extension of the presented model, we have been directing our attention on the phenomenology of attentional states. But attentional states could be described in a more complete and rich way, considering the many distinct facets that such states could present. For example, the stage of distraction has been merely described here as a state of mind-wandering, characterized as presenting a distractive ”true state” associated with a focused ”perceived state”. However, distractive states are not simply an all-or-nothing global phenomenon, and an imprecise classification of distractive states could lead to methodological and theorical limitations [7]. Indeed, recent works have pointed out to an other phenomenology of disctracted state, namely ”Mind blanking” [6, 17]. Mind-blanking has been identified as a distinct mental state from mind-wandering, being characterised as a state of distraction associated with the inability to report mental content [2]. Mind-Wandering (MW) have been discussed before, and can refer to a state in which mind seems to go ”somewhere” disconnected from the current environment [25]. In contrast to this state, which still contains ”content”, the state of ”Mind-blanking” (MB) has been characterised as a state empty of content, as if the mind were ”nowhere” from a period of time [34]. Such distinction between these two states are important to consider, as MB seems to present an extreme decoupling of both perception and attention. In other words, in such state, attention do not carry any information into conscious awareness nor meta-awareness. It would therefore be interesting to add this distinction in the implemented model, as these states differ from the MW state and other mental states regarding behavioural outcomes, distinct physiological substrates and cognitive processes, and phenomenological experience. Dominant MB over MW would also suggest impairment in executive functions, as reported in a study suggesting that executive functions in ADHD are also required to sustain an internal train of thought (ie a certain amount of conscious content). Understanding the hierarchical relations between lower-levels (attentional and perceptual) and higher-levels (meta-awareness), while considering different phenomenologies of ”distractive” states could help us better understanding the processes behind states where attention and consciousness seemed to be ”nowhere”. As a final note, we could hypothesize that a way to implement such distinction (ie Mind-Blanking state) would be to suppress the feedback between first-order levels (perceptual states) and higher-order level when in this state. This would constraint the update of beliefs, when agents would report a total ”lack of content”.

5.2. Limitations

By observing the influence of preferences and ‘Bayesian beliefs’ values encoded in the probability distributions, this framework presents as an interesting avenue to shed lights on the influence of priors in behavioral outcome. However, we noted an important limitation, that will be needed to address in future studies, in order to have a better explanatory power on the twofold translation of Bayesian beliefs formalism in terms of psychological description, and vice versa. A better understanding of the mapping between the psychological (psychological beliefs) and mathematical levels (bayesian beliefs) of description could lead to a better manipulation of the parameters, as well as to a usability of active inference models in practice. Indeed, we might expect or suppose that using active inference in practice might potentially provide a finer-grained distinctions of psychological descriptions, taking into account different factors driving accumulation of information (eg. motivation, exploration, reward). A finer-grained description, distinction and understanding of (inferential) mechanisms could in turn provide with a more robust and precise explanatory power over ”quantitative folk-psychological predictions” [28].

6. Conclusion

6.1. Coda - *Modeling in Neuroscience & elsewhere*

The present project report is submitted in fulfillment of the requirements for the *"Modeling in neuroscience - and elsewhere"* course (École Normale Supérieure Paris-Saclay) taught by Jean-Pierre Nadal [1]. By providing insights about forms of cognitive architecture, the work presented has the twofold advantage of displaying how advances in machine learning allow for innovation in the field of (cognitive) neuroscience, and reciprocally how cognitive (neuro)sciences can shed light on the nature of the representations constructed by (deep) artificial networks. In line with the scope of the class, we note that the computational approach of phenomenology presented here could also provides openings to other themes, such as in social complex systems. Such model would be particularly an interesting avenue for areas of research that are interested in studying variational properties of universal and individual features (e.g metacognitive skills) across cultures [22]. These subjects would subsequently go beyond neuroscience modelling and include a list of fields related to anthropology, linguistics, psychology and sociology, to name but a few.

The code for the experiment is available here : `https://github.com/romybeaute/Mod-Neuro-MVA/blob/main/simulating_active_inference_ADHD_agent.ipynb`

[1] http://www.phys.ens.fr/~nadal/Cours/MVA/index_en.html

References

[1]

[2] Thomas Andrillon, Jennifer Windt, Tim Silk, Sean Drummond, Mark A Bellgrove, and Naotsugu Tsuchiya. Does the mind wander when the brain takes a break? local sleep in wakefulness, attentional lapses and mind-wandering. *Frontiers in Neuroscience*, page 949, 2019.

[3] Amy F. T. Arnsten. Fundamentals of attention-deficit/hyperactivity disorder: circuits and pathways. *The Journal of clinical psychiatry*, 67 Suppl 8:7–12, 2006.

[4] David Chalmers. The Meta-Problem of Consciousness. *Journal of Consciousness Studies*, 25(9-10):6–61, 2018. Publisher: Imprint Academic.

[5] Ana Cubillo, Rozmin Halari, Anna Smith, Eric Taylor, and Katya Rubia. A review of fronto-striatal and fronto-cortical brain abnormalities in children and adults with attention deficit hyperactivity disorder (adhd) and new evidence for dysfunction in adults with adhd during motivation and attention. *cortex*, 48(2):194–215, 2012.

[6] Athena Demertzi, Enzo Tagliazucchi, Stanislas Dehaene, Gustavo Deco, Pablo Barttfeld, Federico Raimondo, Charlotte Martial, Davinia Fernández-Espejo, Benjamin Rohaut, HU Voss, et al. Human consciousness is supported by dynamic complex patterns of brain signal coordination. *Science advances*, 5(2):eaat7603, 2019.

[7] Juergen Fell and Leila Chaieb. Commentary: Guesdon et al.(2020) mind-wandering changes in dysphoria. *Frontiers in Psychiatry*, 12:1204, 2021.

[8] Karl Friston, Thomas FitzGerald, Francesco Rigoli, Philipp Schwartenbeck, and Giovanni

Pezzulo. Active inference: a process theory. *Neural computation*, 29(1):1–49, 2017.

[9] Karl Friston, Spyridon Samothrakis, and Read Montague. Active inference and agency: optimal control without cost functions. *Biological cybernetics*, 106(8):523–541, 2012.

[10] Karl J Friston, Jean Daunizeau, James Kilner, and Stefan J Kiebel. Action and behavior: a free-energy formulation. *Biological cybernetics*, 102(3):227–260, 2010.

[11] Casper Hesp, Ryan Smith, Thomas Parr, Micah Allen, Karl J Friston, and Maxwell JD Ramstead. Deeply felt affect: the emergence of valence in deep active inference. *Neural computation*, 33(2):398–446, 2021.

[12] Jakob Hohwy and Anil Seth. Predictive processing as a systematic basis for identifying the neural correlates of consciousness. *Philosophy and the Mind Sciences*, 1(II), Dec. 2020. Number: II.

[13] Françoise Lecaignard, Olivier Bertrand, Anne Caclin, and Jérémie Mattout. Adaptive cortical processing of unattended sounds: neurocomputational underpinnings revealed by simultaneous eeg-meg. 2018.

[14] Jakub Limanowski and Karl Friston. 'seeing the dark': grounding phenomenal transparency and opacity in precision estimation for active inference. *Frontiers in psychology*, 9:643, 2018.

[15] Jakub Limanowski and Karl Friston. Attenuating oneself: An active inference perspective on "selfless" experiences. *Philosophy and the Mind Sciences*, 1(I):1–16, 2020.

[16] Thomas Metzinger. Phenomenal transparency and cognitive self-reference. *Phenomenology and the Cognitive Sciences*, 2(4):353–393, 2003.

[17] Sepehr Mortaheb, Manousos A Klados, Laurens Van Calster, Paradeisios Alexandros Boulakis, Kleio Georgoula, Steve Majerus, and Athena Demertzi. Mind blanking is associated with a rigid spatio-temporal profile in typical wakefulness. *bioRxiv*, 2021.

[18] Thomas Nagel. What is it like to be a bat. *Readings in philosophy of psychology*, 1:159–168, 1974.

[19] Marie-Christine Nizzi, Veronique Blandin, and Athena Demertzi. Attitudes towards personhood in the locked-in syndrome: from third- to first-person perspective and to interpersonal significance. *Neuroethics*, 13(2):193–201, July 2018.

[20] Claire Petitmengin. Describing one's subjective experience in the second person: An interview method for the science of consciousness. *Phenomenology and the Cognitive sciences*, 5(3):229–269, 2006.

[21] Russell A Poldrack and Tal Yarkoni. From brain maps to cognitive ontologies: informatics and the search for mental structure. *Annual review of psychology*, 67:587–612, 2016.

[22] Joëlle Proust and Martin Fortier. *Metacognitive diversity: An interdisciplinary approach*. Oxford University Press, 2018.

[23] Lars Sandved-Smith, Casper Hesp, Jérémie Mattout, Karl Friston, Antoine Lutz, and Maxwell JD Ramstead. Towards a computational phenomenology of mental action: modelling meta-awareness and attentional control with deep parametric active inference. *Neuroscience of consciousness*, 2021(1):niab018, 2021.

[24] J. Savulescu and G. kahane. Brain damage and the moral significance of consciousness. *Journal of Medicine and Philosophy*, Feb. 2009.

[25] Jonathan W Schooler, Erik D Reichle, and David V Halpern. *Zoning out while reading: Ev-*

idence for dissociations between experience and metaconsciousness. MIT press, 2004.

[26] Anil K Seth. *The cybernetic Bayesian brain*. Open MIND. Frankfurt am Main: MIND Group, 2014.

[27] Ryan Smith, Karl J Friston, and Christopher J Whyte. A step-by-step tutorial on active inference and its application to empirical data. *Journal of Mathematical Psychology*, 107:102632, 2022.

[28] Ryan Smith, Maxwell JD Ramstead, and Alex Kiefer. Active inference models do not contradict folk psychology. *Synthese*, 200(2):1–37, 2022.

[29] Zoe R Smith and Joshua M Langberg. Review of the evidence for motivation deficits in youth with adhd and their association with functional outcomes. *Clinical Child and Family Psychology Review*, 21(4):500–526, 2018.

[30] Evan Thompson and Francisco J Varela. Radical embodiment: neural dynamics and consciousness. *Trends in cognitive sciences*, 5(10):418–425, 2001.

[31] Charlotte Van den Driessche, Mikaël Bastian, Hugo Peyre, Coline Stordeur, Éric Acquaviva, Sara Bahadori, Richard Delorme, and Jérôme Sackur. Attentional lapses in attention-deficit/hyperactivity disorder: Blank rather than wandering thoughts. *Psychological science*, 28(10):1375–1386, 2017.

[32] Francisco J. Varela, Evan Thompson, and Eleanor Rosch. *The Embodied Mind: Cognitive Science and Human Experience*. MIT Press, 1991. Google-Books-ID: gzLaDQAAQBAJ.

[33] Samuel PL Veissière, Axel Constant, Maxwell JD Ramstead, Karl J Friston, and Laurence J Kirmayer. Thinking through other minds: A variational approach to cognition and culture. *Behavioral and brain sciences*, 43, 2020.

[34] Adrian Frank Ward and Daniel M Wegner. Mind-blanking: When the mind goes away. *Frontiers in psychology*, 4:650, 2013.