Elsevier Editorial System(tm) for CORTEX
Manuscript Draft

Manuscript Number: CORTEX-D-18-00417R1

Title: Dark Control: The Default Mode Network as a Reinforcement Learning Agent

Article Type: SI:The Evolution of the Mind

Keywords: systems neuroscience, artificial intelligence, mind-wandering

Corresponding Author: Professor Danilo Bzdok,

Corresponding Author's Institution: Research Center Jülich

First Author: Elvis Dohmatob, Dr

Order of Authors: Elvis Dohmatob, Dr; Guillaume Dumas, Prof. Dr. ; Danilo Bzdok

Abstract: The default mode network (DMN) is believed to subserve the baseline mental activity in humans. Its higher energy consumption compared to other brain networks and its intimate coupling with conscious awareness are both pointing to an unknown overarching function. Many research streams speak in favor of an evolutionarily adaptive role in envisioning experience to anticipate the future. In the present work, we propose a process model that tries to explain how the DMN may implement continuous evaluation and prediction of the environment to guide behavior. The main purpose of DMN activity, we argue, may be described by Markov Decision Processes that optimize action policies via value estimates based through vicarious trial and error. Our formal perspective on DMN function naturally accommodates as special cases previous interpretations based on (1) predictive coding, (2) semantic associations, and (3) a sentinel role. Moreover, this process model for the neural optimization of complex behavior in the DMN offers parsimonious explanations for recent experimental findings in animals and humans.

*Cover Letter

Danilo Bzdok, MD, PhD
Parietal team,
INRIA Saclay-Île-de-France,
Neurospin,
CEA Saclay, Bât 145,
91191 GIF/YVETTE,
FRANCE
E-mail : danilo.bzdok@rwth-aachen.de

Saclay, August 10, 2018

The Editors
Cortex

Cortex special issue "The evolution of the mind and brain":
Revision of manuscript D-18-00417

Dear Professor Zilles, Dear Professor Thiebaut de Schotten,

Please find attached our revised manuscript entitled "Dark Control: The Default Mode Network as a Reinforcement Learning Agent," which we would like to submit to *Cortex*.

We are thankful to the reviewers, whose comments have helped us to considerably improve the manuscript. A point-by-point response is listed below, with changes in the manuscript marked as red.

While looking forward to hearing from you, we thank you very much for your time and consideration.

Yours sincerely,

Elvis Dohmatob, Guillaume Dumas, Danilo Bzdok

CENTRE DE RECHERCHE
SACLAY - ÎLE-DE-FRANCE
Parc Orsay Université
4, rue Jacques Monod
91893 Orsay cedex
Tél. : +33 (0)1 72 92 59 00
Fax : +33 (0)1 74 85 42 42
www.inria.fr

*Detailed Response to Reviewers

Reviewer #1: Dohmatob and colleagues aim to address the difficult question of the function of the Default Mode Network (DMN). In this theoretical article, the authors propose that the DMN function is to perform behavioral performance optimization based on Markov Decision Processes. This optimization relies on a continuous implementation of choices of next actions guided by outcomes of really happened, hypothetically imagined, and expected futures. The proposed model could be a timely contribution on an important issue: while many studies focused on how the DMN is differentially involved in brain diseases or altered states of consciousness, few studies addressed the hard question of the functions and mechanisms of this functional network. However, there are several concerns that should be addressed.

1) In a first part, the authors review the literature on the functional role of several regions thought to belong to the DMN. Although this section is exhaustive and well documented, it often lacks precision and focus regarding both the anatomical definition of the regions of the DMN and the justification of the functions attributed to them.

A misunderstanding may account for a part of this comment. Our manuscript did not attempt to describe the human default mode network based on the finest-possible anatomical nomenclature. Rather, we wanted to refer to the major nodes of the DMN by primarily denoting *functional zones* in the aim to facilitate across-species comparison and discussion.

The referred review section of the manuscript focuses on juxtaposing experimental findings across humans, monkeys, rats, and other animals. Importantly, however, the exact homology of specific anatomical areas in the highly associative DMN is still subject to debate. For instance, the precise homologue of the human “temporoparietal junction” in primates is still uncertain (Geschwind, 1965; Seghier, 2013; Zilles and Palomero-Gallagher, 2001). However, seminal resting-state connectivity analyses identified a similar set of functional nodes in monkeys, including bilateral nodes in the inferior parietal cortex, which bear close resemblance with the spatial appearance of the human DMN (Mantini et al., 2011).
For these reasons, we opted for a broader descriptive language that can denote the major constituent nodes of the DMN in both higher and simpler mammals. This is probably all the more important because the animal literature is an essential source of strong neurobiological insight through the possibility of invasive experiments, such as locally circumscribed lesion experiments and in-vivo single-cell recordings.

Regarding precision of functional descriptions, the overall ambition of the paper is to go new ways and try an engineering approach towards a possible *description* of DMN function, rather than the traditional and commonly tried attempt by imaging cognitive functional manipulations through experimental design. That is, the large majority of previous papers attempting to get at functional conclusions based on cognitive theory used to justify precise experimental stimuli and task sets. Our conceptual approach is fundamentally different in *trying to get at functional precision by means of a mathematical formalization* (i.e., Markov Decision Processes, MDP) that is a generally accepted and empirically successful model approach in various engineering domains (e.g., such as to quantify and predict stock market changes and bird migration behavior) to formally outline what each region of the DMN may be doing in a general reinforcement learning (RL) framework.

As such, our review of putative functional roles of DMN nodes was conducted in with the very specific focus on previously proposed evidence (mainly based on animals) that appear coherent to subserve components of an RL agent. Additionally, a model-based approach to

CENTRE DE RECHERCHE
SACLAY - ÎLE-DE-FRANCE
Parc Orsay Université
4, rue Jacques Monod
91893 Orsay cedex
Tél. : +33 (0)1 72 92 59 00
Fax : +33 (0)1 74 85 42 42
www.inria.fr

speculating about DMN function allows to propose a way of reasoning *how* the DMN does what it is doing, rather than *what* it is doing.

In response to this helpful reviewer comment, we have added new sentences that better explain our intention of emphasizing broad functional role and consciously restraining the interpretational focus and literature review to the goal of specific perspective of our paper.

Section with literature review (Page 3): "We begin by a neurobiological deconstruction of the DMN based on integrating experimental findings in the neuroscience literature from different species. This walkthrough across main functional zones of the DMN (i.e., deemphasizing their precise anatomical properties) will outline the individual functional profiles with the goal of paving the way for their algorithmic interpretation in our formal account. As our focus will be on major *functional* zones of the DMN, please see elsewhere for excellent surveys on their *anatomical* boundaries and connectivity patterns (e.g., Buckner et al., 2008; Binder et al., 2009; Seghier et al., 2013)."

Please note that we now provide several paper references with rigorous treatment of the anatomy of the DMN.

Conclusion (Page 23): "Which brain function could be important enough for the existence and survival of the human species to justify constantly high energy costs? While existing experiments on the DMN frequently set out to investigate *what* its subserved function may be, we have proposed a way of reasoning *how* this major network may do what it is doing."

Binder, J.R., Desai, R.H., Graves, W.W., Conant, L.L., 2009. Where is the semantic system? A critical review and meta-analysis of 120 functional neuroimaging studies. Cereb Cortex 19, 2767-2796.

Buckner, R.L., Andrews-Hanna, J.R., Schacter, D.L., 2008. The brain's default network: anatomy, function, and relevance to disease. Ann N Y Acad Sci 1124, 1-38.

Geschwind, N., 1965. Disconnexion syndromes in animals and man. I. Brain 88, 237-294.

Mantini, D., Gerits, A., Nelissen, K., Durand, J.B., Joly, O., Simone, L., Sawamura, H., Wardak, C., Orban, G.A., Buckner, R.L., Vanduffel, W., 2011. Default mode of brain function in monkeys. J Neurosci 31, 12954-12962.

Seghier, M.L., 2013. The Angular Gyrus: Multiple Functions and Multiple Subdivisions. Neuroscientist, 43-61.

Zilles, K., Palomero-Gallagher, N., 2001. Cyto-, myelo-, and receptor architectonics of the human parietal cortex. Neuroimage 14, S8-20.

a. As examples of anatomical lack of precision, the authors consider the PMC as a unitary region with a specific role, while it is a large and heterogeneous region encompassing subregions that probably have distinct roles (the cited literature relates to the precuneus, medial posterior parietal, posterior cingulate and retrosplenial

CENTRE DE RECHERCHE
SACLAY - ÎLE-DE-FRANCE
Parc Orsay Université
4, rue Jacques Monod
91893 Orsay cedex
Tél. : +33 (0)1 72 92 59 00
Fax : +33 (0)1 74 85 42 42
www.inria.fr

regions). Similarly, the TPJ is a large region with imprecise anatomical limits and which also may encompass distinct functional subregions. Related to the TPJ, the authors discuss the literature on the role of the angular gyrus and supramarginal gyrus, which are themselves also heterogeneous functionally (see for instance Seghier et al 2010). The dmPFC is also ill defined and the cited literature seems to relate to several distinct prefrontal areas. The middle temporal gyrus is not mentioned in the paper although it is typical component of the DMN that may also play a role in semantic aspects of the model. In addition, instead of hippocampus, the authors should consider using 'hippocampus/parahippocampal cortex'.

In fact, we refer to the posterior midline of the DMN as “PMC” when we indicate this part of the cortex as a functional node as observed in many functional neuroimaging studies. However, different parts of the manuscript go into more precise topographical detail when explicitly referring to anatomical portions of this functional zones, of which we give examples here:

“Moreover, the retrosplenial portion of the PMC could support representation of action possibilities and evaluation of reward outcomes by integrating information from memory recall and different perspective frames.”

“Regarding perspective frames, the retrosplenial subregion of the PMC has been proposed to mediate between...”

“Specifically, among all parts of the PMC, the ventral posterior cingulate cortex was most connected to the laterobasal nuclei group of the amygdala.”

Regarding the TPJ, we have made explicit our intention to denote the “TPJ” as a functional zone of the DMN, given that its anatomical properties are still subject to debate.

Page 8: “The TPJ in the right hemisphere (RTPJ) denotes a broad functional zone with varying anatomical nomenclature (Mars et al., 2011; Seghier et al., 2010; Seghier et al., 2013) that has been shown to be closely related to multi-sensory prediction and prediction error signaling.”

Mars, R.B., Jbabdi, S., Sallet, J., O'Reilly, J.X., Croxson, P.L., Olivier, E., Noonan, M.P., Bergmann, C., Mitchell, A.S., Baxter, M.G., Behrens, T.E., Johansen-Berg, H., Tomassini, V., Miller, K.L., Rushworth, M.F., 2011. Diffusion-weighted imaging tractography-based parcellation of the human parietal cortex and comparison with human and macaque resting-state functional connectivity. J Neurosci 31, 4087-4100.

Seghier, M.L., Fagan, E., Price, C.J., 2010. Functional subdivisions in the left angular gyrus where the semantic system meets and diverges from the default network. J Neurosci 30, 16809-16817.

Seghier, M.L., 2013. The Angular Gyrus: Multiple Functions and Multiple Subdivisions. Neuroscientist, 43-61.

CENTRE DE RECHERCHE
SACLAY - ÎLE-DE-FRANCE
Parc Orsay Université
4, rue Jacques Monod
91893 Orsay cedex
Tél. : +33 (0)1 72 92 59 00
Fax : +33 (0)1 74 85 42 42
www.inria.fr

Regarding the prefrontal cortex, the vmPFC and dmPFC are well known to be distinct based on neurological lesion studies, functional and structural connectivity analyses as well as many other neuroanatomical and neurophysiological arguments based neuroimaging and non-neuroimaging studies. To allow for the more ventral versus more dorsal distinction in the medial prefrontal cortex but still be able to relate our functional zones in findings from other species, such as monkeys and rodents, we opted for a broader division into vmPFC, mostly related to BA10, and dmPFC, mostly related to BA9, as a most parsimonious way to distinguish functional compartments in the medial prefrontal cortex (see changes to Figure 1 below).

Regarding the middle temporal gyrus, we note that, while some general DMN mapping studies have reported the MTG, there is also a series of quantitative neuroimaging investigations on DMN-related cognitive processes that have revealed negligible neural activity changes in the MTG. This was for instance the case for studies on semantic processing (Binder et al., 2009), guessing social traits from faces (Mende-Siedlecki et al., 2013), autobiographical memory retrieval (Spreng et al., 2009), spatial navigation (Spreng et al., 2009), imagining the future (Spreng et al., 2009), putting oneself into others' shoes (Spreng et al., 2009), as well as empathy and moral cognition (Bzdok et al., 2012).

Binder, J.R., Desai, R.H., Graves, W.W., Conant, L.L., 2009. Where is the semantic system? A critical review and meta-analysis of 120 functional neuroimaging studies. Cereb Cortex 19, 2767-2796.

Bzdok, D., Schilbach, L., Vogeley, K., Schneider, K., Laird, A.R., Langner, R., Eickhoff, S.B., 2012. Parsing the neural correlates of moral cognition: ALE meta-analysis on morality, theory of mind, and empathy. Brain Struct Funct 217, 783-796.

Mende-Siedlecki, P., Said, C.P., Todorov, A., 2013. The social evaluation of faces: a meta-analysis of functional neuroimaging studies. Soc Cogn Affect Neurosci 8, 285-299.

Spreng, R.N., Mar, R.A., Kim, A.S., 2009. The common neural basis of autobiographical memory, prospection, navigation, theory of mind, and the default mode: a quantitative meta-analysis. J Cogn Neurosci 21, 489-510.

Regarding the HC, we agree that neural activity changes in the hippocampus are often coinciding with the parahippocampal parts in the reviewed studies, which we now make explicit in the enhanced paper manuscript at first mention.

Page 6: "The DMN midline has close functional links with the HC (henceforth implying to include also parahippocampal regions) in the medial temporal lobe..."

b. The justification of the conclusions regarding the specific function supported by each region is not always clear and well-argued, with sometimes conceptual leaps. For instance, the arguments on the role of the right TPJ alternates between its role shown previously in sensory attention, encoding of action, attention reallocation,

CENTRE DE RECHERCHE
SACLAY - ÎLE-DE-FRANCE
Parc Orsay Université
4, rue Jacques Monod
91893 Orsay cedex
Tél. : +33 (0)1 72 92 59 00
Fax : +33 (0)1 74 85 42 42
www.inria.fr

prediction of outcomes, without a clear explanation on how these findings explain its role as a potential switch between externally-oriented and externally-oriented mind sets (lines 37-43). The fact that the left TPJ is anatomically close to Wernicke area is not a strong argument toward its role in semantics. The Balint syndrome in patients with parietal lesions is not a strong argument for a role of the PMC in information integration as the anatomical bases of the Balint syndrome are not well understood (this syndrome usually results from several bilateral lateral parieto-occipital lesions). It is also unclear how the Balint syndrome can be interpreted as a "high level impairment in exploration and exploitation" and how it argues for the role of the PMC in "abstract" integration (page 4). On page 15, lines 12-16, the meaning of this sentence and how it relates to the previous paragraph is unclear, the authors may want to clarify this.
The authors may want to focus their hypotheses on the precisely defined DMN regions involved in the model and on their exact role or place in the processes.

Rather than referring to isolated neuroimaging findings, we tried to provide support for our functional explanations by widely accepted neurological lesion symptoms. We now provide further detail on these arguments and adapt the manuscript accordingly:

Regarding the TPJs, Wernicke's areas and TPJ are used in our paper as functional terms, not meant to imply a certain anatomically circumscribed cortical area. This is a view that we share with anatomists and cognitive neurologists such as the exemplary quotes in the following:

Hal Blumenfeld wrote in his book "Neuroanatomy through clinical cases" (page 832): "Wernicke's aphasia is usually caused by a lesion of Wernicke's area and adjacent structures in the dominant temporoparietal lobes."

Marcel Mesulam's textbook "Principles of Behavioral and Cognitive Neurology" (pages 29-30): "The traditional literature gives the impression that Wernicke's area is confined to auditory association cortex in the posterior third of the superior temporal gyrus. This is quite unlikely since the deficit in Wernicke's aphasia is multimodal and impairs language comprehension in all modalities of input. A more likely possibility is that Wernicke's area extends into heteromodal cortical areas and that the posterior third of the superior temporal gyrus constitutes only one of its components. [...] Wernicke's area has no universally accepted boundary. It is usually defined as "the region which causes Wernicke's aphasia when damaged."

Page 8: "The left TPJ of the DMN (LTPJ), in turn, may have a functional relationship to Wernicke's area involved in semantic processes (Blumenfeld, 2002) and has been described as "a temporoparietal transmodal gateway for language" by some investigators (Mesulam, 2000)."

Regarding the right TPJ, its implication in attention and prediction error on the one hand and potentially related network coupling changes has been described by several previous publications, of which we list a few in the following:

CENTRE DE RECHERCHE
SACLAY - ÎLE-DE-FRANCE
Parc Orsay Université
4, rue Jacques Monod
91893 Orsay cedex
Tél. : +33 (0)1 72 92 59 00
Fax : +33 (0)1 74 85 42 42
www.inria.fr

Downar, J., Crawley, A.P., Mikulis, D.J., Davis, K.D., 2000. A multimodal cortical network for the detection of changes in the sensory environment. Nat Neurosci 3, 277-283.

Geng, J.J., Vossel, S., 2013. Re-evaluating the role of TPJ in attentional control: contextual updating? Neuroscience & Biobehavioral Reviews 37, 2608-2620.

Shulman, G.L., Pope, D.L., Astafiev, S.V., McAvoy, M.P., Snyder, A.Z., Corbetta, M., 2010. Right hemisphere dominance during spatial selective attention and target detection occurs outside the dorsal frontoparietal network. Journal of Neuroscience 30, 3640-3651.

Shulman, G.L., Astafiev, S.V., McAvoy, M.P., d'Avossa, G., Corbetta, M., 2007. Right TPJ deactivation during visual search: functional significance and support for a filter hypothesis. Cereb Cortex 17, 2625-2633.

Vetter, P., Butterworth, B., Bahrami, B., 2011. A candidate for the attentional bottleneck: set-size specific modulation of the right TPJ during attentive enumeration. J Cogn Neurosci 23, 728-736.

Regarding Bálint's syndrome, we fully agree with the reviewer that there is unfortunately insufficient knowledge on how parietal cortex impairments led to this set of dysfunctions in high-level cognition. However, several authoritative sources have linked tissue damage in the posterior partial midline (especially Brodmann area 7) of the right and left hemisphere to Balint's syndrome, of which we list a few here:

Marcel Mesulam's textbook "Principles of Behavioral and Cognitive Neurology" (page 355):

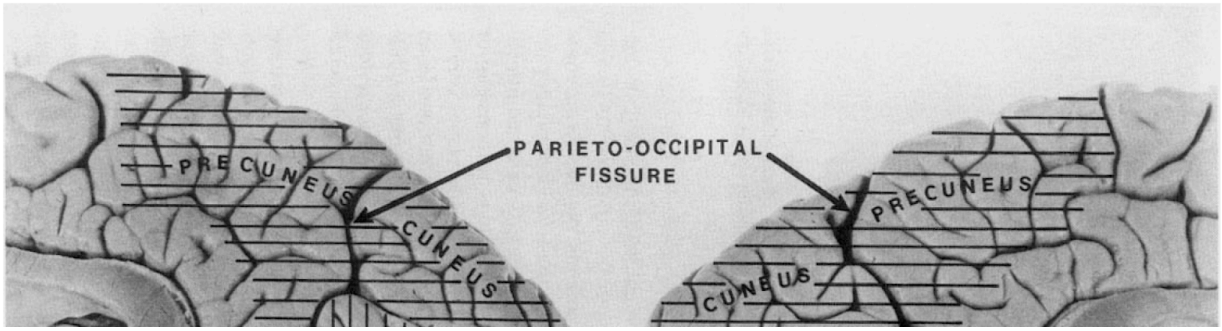


Figure 7-3: Mesial view of the posterior part of a human brain...Patients with Bálint's syndrome have bilateral lesions in the occipitoparietal region (horizontal hatching), which is composed of cuneus and precuneus. [...]

Marcel Mesulam's also emphasizes the relation to BA7 which is what we denote to by "precuneus" in the manuscript (page 356): "When the entire syndrome is present some involvement (either medially or dorsally) of Brodmann's area (BA) 7 is common[...]"

Buckner et al., 2008: "Finally, bilateral lesions that extend across precuneus and cuneus can induce Balint's syndrome. […] Balint's syndrome is characterized by a form of tunnel vision. Patients can only perceive a small portion of the visual world at one time and often fail to notice the appearance of objects outside the immediate focus of attention"

CENTRE DE RECHERCHE
SACLAY - ÎLE-DE-FRANCE
Parc Orsay Université
4, rue Jacques Monod
91893 Orsay cedex
Tél. : +33 (0)1 72 92 59 00
Fax : +33 (0)1 74 85 42 42
www.inria.fr

Hal Blumenfeld in his book “Neuroanatomy through clinical cases” (page 863): “From a clinical perspective, Bálint’s syndrome provides an interesting example of how focal lesions of the parietal-occipital cortex can cause a profound deficit in the ability to bind various individual parts of a visual scene into a single integrated whole.”

In response to the helpful reviewer comment, we have now simplified our statements and provide additional literature references.

Page 4: “For instance, Bálint’s syndrome is a neurological disorder of conscious awareness that can result from tissue damage in the posterior medial cortex (Bálint et al., 1909; Buckner et al., 2008). Such neurological patients are plagued by an inability to bind various individual features of the visual environment into an integrated whole (i.e., simultanagnosia) as well as an inability to direct action towards currently unattended environmental objects (i.e., optic ataxia). Scanning complex scenes is impaired in that static or moving objects in the environment may not be invisible or disappear in the subjective perception of the patient (Mesulam, 2000; Blumenfeld, 2002). This dysfunction can be viewed as a high-level impairment in gathering information about alternative objects (i.e., exploration) as well as using these environmental opportunities towards a behavioral goal (i.e., exploitation).”

As another consequence from this comment, we now more consistently mention that our review of the functional literature is targeted at common denominators through the lens of the reinforcement learning account for all considered major DMN nodes.

Page 5: “Estimated value, found to differs across individuals, might enrich statistical assessment of the environment to map and predict delayed reward opportunities in the future. Viewed from a RL perspective, the PMC may continuously adapt the organism to changes in both the external environment and its internal representation to enable strategic behavior.”

Page 5: “The dmPFC may subserve representation and assessment of one’s own and other individuals’ action considerations - a necessary component of a full-blown RL agent.”

Page 8: “Within a RL framework, the HC could thus orchestrate re-experience of environmental aspects for consolidations based on re-enactment and for integration into rich mental scene construction (Deuker et al., 2016; Bird et al., 2010). As such, the HC may impact ongoing perception of and action on the environment (Zeidman and Maguire, 2016; De Lavilleon et al., 2015).”

Page 8: “Viewed from a RL account, the RTPJ might reflect an important relay that shifts away from the internally directed baseline processes to, instead, deal with unexpected environmental stimuli and events.”

Page 9: “As a candidate component of a RL agent, neural processes in the LTPJ may contribute to the automated predictions of the”

CENTRE DE RECHERCHE
SACLAY - ÎLE-DE-FRANCE
Parc Orsay Université
4, rue Jacques Monod
91893 Orsay cedex
Tél. : +33 (0)1 72 92 59 00
Fax : +33 (0)1 74 85 42 42
www.inria.fr

2) The paper does not discuss how MDP processes could account for the observed DMN deactivation during experimental tasks. For example, if the role of vmPFC is to assign value to stimuli, then the activity of this brain area during a valuation task should be really similar than the activity observed during resting state. However, it has been often observed that the vmPFC activity decreases during such a valuation task and the subjective value is represented on top of this deactivation (see for instance Abitbol et al, 2015). The global dynamics of the DMN during tasks and resting state consequently deserves further discussion.

Most neuroscientists today probably agree that there may be a trade-off between online performance at cognitive experiments and experiment-independent offline processing subserved by the DMN as measured using resting-state fMRI. In brain-imaging studies, goal-directed, attention-focused task performance has repeatedly been reported to improve when neural activity in default mode regions decreased (Weissman et al., 2006). Additionally, increased DMN activity was frequently linked to more task-independent thoughts (Mason et al., 2007). It is widely acknowledged that the more the attended stimuli are novel and the executed task is unpracticed, the less stimulus-independent thoughts tend to occur (Christoff et al., 2016). Additionally, several parts of the DMN are known to have the highest absolute baseline energy consumption in the human cortex (Raichle et al., 2001). Stimulus and task changes during fMRI experiments usually induce rather small changes of <5% of the resting-state signal level, thus typically preserving >95% of the DMN baseline activity (this is stated in the introduction: "The baseline energy demand is only weakly modulated at the onset of defined psychological tasks (Gusnard and Raichle, 2001)."). As such, we hoped to convey a graded perspective on DMN engagement, including its vmPFC node: The less the stimuli are novel and the less the ongoing task is new, the more DMN activity may be devoted, as unused resource, to MDP-like processes extending beyond the present timescale. Without requiring explicit awareness, such offline processes may contribute to continuously optimizing control of the organism in general.

This perspective concurs with the previously voiced view of the DMN as a neurobiological manifestation of the “stream of consciousness” (James, 1890) as a potential baseline brain function (Buckner et al., 2013). This brain network might have emerged to continuously make predictions and continuously enhance these by means of mental imagery as an evolutionary advantage. In line with this contention, the posteromedial DMN has for instance been proposed to serve a function resembling an inner “mind’s eye” (Cavanna and Trimble, 2006). Consequently, we expect that “housekeeping” computations subserved by the DMN are attenuated in the face of novel cues and unexpected events in the environment to reallocate resources towards basic survival needs, in line with reviewer’s points.

One important candidate mechanism for updating the MDP (improving policy matrix and value function in particular) after event experience and feedback collection is information transfer from the medial temporal limbic system to DMN areas during memory consolidation at sleep. Indeed, “[s]ome investigators have suggested that the vmPFC is the consolidated (remote) memory homologue of the [hippocampus], taking over its function” (Moscovitch et al., 2016). If, at a later point in time, the agent encounters the previously anticipated events, the MDP will be more “prepared”, the context will be evaluated to be less novel and the DMN will deactivate less as during the initial encounter of that event allowing for more environment-independent mind-wandering to occur.

We clarified the Section 3.3. (Summary and hypotheses for future studies) to better discuss those aspects (Page 15): “The DMN is today known to consistently increase in neural activity

CENTRE DE RECHERCHE
SACLAY - ÎLE-DE-FRANCE
Parc Orsay Université
4, rue Jacques Monod
91893 Orsay cedex
Tél. : +33 (0)1 72 92 59 00
Fax : +33 (0)1 74 85 42 42
www.inria.fr

when humans engage in cognitive processes that are relatively detached from the current sensory environment. The more familiar and predictable the current environment, the more brain resources may remain for allocating DMN activity to MDP processes extending beyond the present time and sensory context. This speculation receives quantitative support in that connectional links between nodes of the DMN have been reported to be more consistent and reliable than functional couplings within any other macroscopical networks (Shehzad et al., 2009). As such, random-sampling-related baseline evaluation of action possibilities and their consequences may be subserved by the DMN and get partly suspended when novelty in the external environment is encountered or immediate action is required. In line with this perspective, DMN engagement was shown to heighten and relate to effective behavioral responses in the practiced phase of a demanding cognitive flexibility task, as compared to acquisition phase when participants learned context-specific rules. This involvement in automated decision-making has led the authors to propose an “autopilot” role for the DMN (Vatansever et al., 2017), which may contribute to optimizing intervention of the organism on the world in general. Among all parts of the DMN, the RTPJ is perhaps the most evident candidate for a network-switching relay that calibrates between processing of environment-engaged versus internally generated information (Downar et al., 2000; Golland et al., 2006; Bzdok et al., 2013c)."

Buckner, R.L., Krienen, F.M., 2013. The evolution of distributed association networks in the human brain. Trends Cogn Sci 17, 648-665.

Cavanna, A.E., Trimble, M.R., 2006. The precuneus: a review of its functional anatomy and behavioural correlates. Brain 129, 564-583.

Christoff, K., Irving, Z.C., Fox, K.C.R., Spreng, R.N., Andrews-Hanna, J.R., 2016. Mind-wandering as spontaneous thought: a dynamic framework. Nature Reviews Neuroscience.

James, W., 1890. The Principles of Psychology. Henry Holt and Company, New York.

Mason, M.F., Norton, M.I., Van Horn, J.D., Wegner, D.M., Grafton, S.T., Macrae, C.N., 2007. Wandering minds: the default network and stimulus-independent thought. Science 315, 393-395.

Moscovitch, M., Cabeza, R., Winocur, G., Nadel, L., 2016. Episodic memory and beyond: the hippocampus and neocortex in transformation. Annual review of psychology 67, 105-134.

Raichle, M.E., MacLeod, A.M., Snyder, A.Z., Powers, W.J., Gusnard, D.A., Shulman, G.L., 2001. A default mode of brain function. Proc Natl Acad Sci USA 98, 676-682.

Weissman, D.H., Roberts, K.C., Visscher, K.M., Woldorff, M.G., 2006. The neural bases of momentary lapses in attention. Nat Neurosci 9, 971-978.

3) In the section 'The components of reinforcement learning in the DMN', the PMC and the left and right TPJ are not mentioned in the text and the suggested computations occurring in those regions in Figure 4 are not properly justified in the text. Regarding the TPJ, the overlap between the social brain network and the DMN (Mars et al, 2012)

CENTRE DE RECHERCHE
SACLAY - ÎLE-DE-FRANCE
Parc Orsay Université
4, rue Jacques Monod
91893 Orsay cedex
Tél. : +33 (0)1 72 92 59 00
Fax : +33 (0)1 74 85 42 42
www.inria.fr

could be discussed. The attribution of 'Prediction Error Signaling' function to the right TPJ should be more strongly justified by the authors. They mention that the rTPJ could be a 'network switching relay that calibrates between processing of environment-engaged versus internally generated information' but this does not imply that 'Prediction Error Signaling' is computed in this area.

Regarding the components of the reinforcement learning section, we agree that some parts of the DMN are mentioned more often than others. As part of the revision process, we have increased mention and explicit linkage with particular DMN nodes as well as make explicit when functional aspects are more likely to be a product of the entire DMN as an integrated process.

Regarding the overlap between the DMN and the social brain, we have several paragraphs in the revised manuscript that deal with this important question, including the mentioned references by Mars and colleagues.

Relation to existing accounts (Page 21): “In particular, environmental cues that are especially important for humans are frequently of social nature. This may not be surprising given that the complexity of the social systems is likely to be a human-defining property (Tomasello, 2009; Dunbar and Shultz, 2007). According to the \social brain hypothesis", the human brain has especially been shaped for forming and maintaining increasingly complex social systems, which allows solving ecological problems by means of social relationships (Whiten and Byrne, 1988). In fact, social topics probably amount to roughly two thirds of human everyday communication (Dunbar et al., 1997). Mind-wandering at daytime and dreams during sleep are also rich in stories about people and the complex interactions between them. In line with this, DMN activity was advocated to be specialized in continuous processing of social information as a physiological baseline of human brain function (Schilbach et al., 2008). This view was later challenged by observing analogues of the DMN in monkeys (Mantini et al., 2011), cats (Popa et al., 2009), and rats (Lu et al., 2012), three species with social capacities that can be expected to be less advanced than in humans (Mars et al., 2012). ...”

Regarding the right TPJ, prediction error in the right TPJ is supported by many studies (without a focus or task related that suggests involvement of reward processing). Thanks to the reviewer comment, we have extended the corresponding passage with supportive paper references:

Page 8: “The TPJ in the right hemisphere (RTPJ) has been shown to be closely related to multi-sensory prediction and prediction error signaling (Downar et al., 2000; Shulman et al., 2010, 2007; Vetter et al., 2011).”

Downar, J., Crawley, A.P., Mikulis, D.J., Davis, K.D., 2000. A multimodal cortical network for the detection of changes in the sensory environment. Nat Neurosci 3, 277-283.

Shulman, G.L., Pope, D.L., Astafiev, S.V., McAvoy, M.P., Snyder, A.Z., Corbetta, M., 2010. Right hemisphere dominance during spatial selective attention and target detection occurs outside the dorsal frontoparietal network. Journal of Neuroscience 30, 3640-3651.

CENTRE DE RECHERCHE
SACLAY - ÎLE-DE-FRANCE
Parc Orsay Université
4, rue Jacques Monod
91893 Orsay cedex
Tél. : +33 (0)1 72 92 59 00
Fax : +33 (0)1 74 85 42 42
www.inria.fr

Shulman, G.L., Astafiev, S.V., McAvoy, M.P., d'Avossa, G., Corbetta, M., 2007. Right TPJ deactivation during visual search: functional significance and support for a filter hypothesis. Cereb Cortex 17, 2625-2633.

Vetter, P., Butterworth, B., Bahrami, B., 2011. A candidate for the attentional bottleneck: set-size specific modulation of the right TPJ during attentive enumeration. J Cogn Neurosci 23, 728-736.

It has been speculated previously that a potential consequence of prediction error detection in the right TPJ may be network recruitment shifts, which we have cited in the manuscript:

Bzdok, D., Langner, R., Schilbach, L., Jakobs, O., Roski, C., Caspers, S., Laird, A.R., Fox, P.T., Zilles, K., Eickhoff, S.B., 2013. Characterization of the temporo-parietal junction by combining data-driven parcellation, complementary connectivity analyses, and functional decoding. Neuroimage 81, 381-392.

In general, it is our view that there is probably no DMN node for which the discussion of possible functional roles in the literature would be consensual. It is for this reasons that we have taken a different route and started by an overarching engineering framework that we then align the existing functional interpretations with.

4) For this theoretical article, the possible studies for testing and validating the model are very important and should be more detailed than in box 2, including potential experiments, and focused on the current hypothesis.

We have decompressed Box 1 and 2 and created a whole new part of Section 3.3. (Summary and hypotheses for future studies) with proposed experimental predictions and scenarios to probe them.

Page 16: Our formal account on the DMN readily motivates several empirical predictions for future neuroscience research. Perhaps one of the first experimental venue concerns the neural correlates of the Bellman equation in the DMN. There are already relationship between the decomposition of consecutive action choices by the Bellman equation and neuroscientific insights: specific neural activity in the dorsal prefrontal cortex (BA9) was for instance linked to processing “goal-tree sequences” in human brain-imaging experiments (Koechlin et al., 1999, 2000). Sub-goal exploration may require multi-task switching between cognitive processes as later parts of a solution frequently depend on respective earlier steps in a given solution path, which necessitates storage of expected intermediate outcomes. As such, “cognitive branching” operations for nested processing of behavioral strategies are likely to entail secondary reallocation of attention and working-memory resources. Further brain-imaging experiments corroborated the prefrontal DMN to subserve “processes related to the management and monitoring of sub-goals while maintaining information in working memory” (Braver and Bongiolatti, 2002) and to functionally couple with the hippocampus conditioned by “deep versus shallow planning” (Kaplan et al., 2017). Moreover, neurological patients with lesions in this DMN region were reported to be impaired in aspects of realizing “multiple sub-goal scheduling” (Burgess et al., 2000). Hence, the various advanced human

CENTRE DE RECHERCHE
SACLAY - ÎLE-DE-FRANCE
Parc Orsay Université
4, rue Jacques Monod
91893 Orsay cedex
Tél. : +33 (0)1 72 92 59 00
Fax : +33 (0)1 74 85 42 42
www.inria.fr

abilities subserved by the DMN, such as planning and abstract reasoning, can be viewed to involve some form of action-decision branching to enable higher-order executive control.
We therefore hypothesize in humans a functional dissociation between computations pertaining to action policy versus adapting stimulus-value associations as we expect implementation in different subsystems of the DMN. First, we expect that fMRI signals in the right temporo-parietal junction relate to behavioral changes subsequent to adaptation in the action choice tendencies (policy matrix) involved in non-value-related prediction error. Second, fMRI signals in the ventromedial prefrontal cortex should relate to behavioral changes following adaptation in value estimation (value matrix) due to reward-related stimulus-value association. We further expect that fMRI signals in the posteromedial cortex, as a potential global information integrator, are related to shifts in overt behavior based on previous adaptations in both policy or value estimation.
Our process model of the DMN has also implications for experiments in neuroeconomy; especially for temporal discounting and continuous learning paradigms. More specifically, we hypothesize in humans a functional relationship between the DMN closely associated with the occurrence of stimulus-independent thoughts and the reward circuitry. During an iterative neuroeconomic two-player game, fMRI signals in the DMN could be used to predict reward-related signals in the nucleus accumbens across trials in a multi-step learning paradigm. We expect that the more DMN activity is measured to be increased, supposedly the higher the tendency for stimulus-independent thoughts, the more the fMRI signals in the reward circuits should be independent of the reward context in the current sensory environment. In the case of temporal discounting, we hypothesize in humans that the relevant time horizon is modulated by various factors such as age, acute stress, and time-enduring impulsivity traits (Luksys et al., 2009; Haushofer and Fehr, 2014). Using such a delayed-reward experiment, it can be quantified how the time horizon is affected at the behavioral level and then traced back to its corresponding neural representation. Such experimental investigation can be designed to examine between-group and within-group effects (e.g., impulsive population like chronic gamblers or drug addicts); and brought in context with the participants age, education, IQ, and personality traits.
As another experimental prediction derived from our MDP approach to the DMN, the HC may contribute to generating perturbed action-transition-state-reward samples as batches of pseudo-experience (i.e., recalled, hypothesized, and forecasted scenarios). The small variations in these experience samplings allow searching through a larger space of model parameters and candidate experiences. Taken to its extreme, stochastic recombination of experience building blocks can further optimize the behavior of the RL agent by learning from scenarios in the environment that the agent might encounter only very rarely or never. An explanation is thus offered for experiencing seemingly familiar situations that a human has however never actually encountered (i.e., d´eja` vu effect). While such a situation may not have been experienced in the physical world, the DMN may have previously stochastically generated, evaluated, and adapted to such a randomly synthesized event. Generated representations arguably are “internally manipulable, and can be used for attempting actions internally, before or instead of acting in the external reality, and in diverse goal and sensory contexts, i.e. even outside the context in which they were learned” (Pezzulo, 2011). In the context of scarce environmental input and feedback (e.g., mind-wandering or sleep), mental scene construction allows pseudo-experiencing possible future scenarios and action outcomes.
A possible interplay between memory retrieval and “mind-searching” moreover suggests that experience replay for browsing problem solutions subserved by the DMN contributes to choice behavior in mice. Hippocampal single-cell recordings have shown that neural patterns during experimental choice behavior are reiterated during sleep and before making analogous choices in the future. We hypothesize that, in addition to the hippocampus, there

CENTRE DE RECHERCHE
SACLAY - ÎLE-DE-FRANCE
Parc Orsay Université
4, rue Jacques Monod
91893 Orsay cedex
Tél. : +33 (0)1 72 92 59 00
Fax : +33 (0)1 74 85 42 42
www.inria.fr

is a necessity of cortical DMN regions for “mind-searching” candidate actions during choice behavior. It can be experimentally corroborated by causal disruption of DMN regions, such as by circumscribed brain lesion or optogenetic intervention in the inferior parietal and prefrontal cortices. From the perspective of a RL agent, prediction in the DMN reduces to generalization of policy and value computations from sampled experiences to successful action choices and reward predictions in future states. As such, plasticity in the DMN arises naturally. If an agent behaving optimally in a certain environment moves to new, yet unexperienced environment, reward prediction errors will largely increase. This feedback will lead to adaptation of policy considerations and value estimations until the intelligent system converges to a new steady state of optimal action decisions in a volatile world.
A last experimental prediction for future studies concerns how synaptic epigenesis may shape the policy matrix. Indeed, we did not adress here the additional layer of learning which concerns the addition of new entries in the state and action spaces. Extension of the action repertoire could be biologically realized by synaptic epigenesis (Gisiger et al., 2005). The tuning of synaptic weights through learning can stabilize additional patterns of activity by creating new attractors in the neural dynamics landscape (Takeuchi et al., 2014). Those attractors can then constrain both the number of factors taken into account by decision processes and the possible behaviors of the agent (Wang, 2008). To examine this potential higher-level mechanism, we propose to probe how synaptic epigenesis is related to neural correlates underlying policy matrix updates: in humans the changes of functional connectivity between DMN regions can be investigated following a temporal discounting experiment and in monkeys or rodents anterograde tracing can be used to study how homolog regions of the DMN present increased synaptic changes compare to other parts of the brain.

5) The figures do not pay tribute to the paper. The legends do not describe sufficiently the figures.

* **Figure 1: The Brodmann's areas numbers do not exactly match the name of the brain regions, e.g., BA9 and BA10. The function assigned to each region is sometimes misleading (Prediction error in the right TPJ is not consensual in the literature; see point 3).**
* **Figure 2 describes the results of a study. The methods of this study should be fully described or the reference of the related publication provided. The authors should also clarify how these results bring more information on the link between the reward system and the DMN, and explain the negative coupling with the left TPJ and positive coupling with the right TPJ. The wording of the legend should be checked.**
* **Figure 3: The title is misleading. A schematic representation of MDP with appropriate legend would be appreciated (meaning of the thought cloud, of the origin of the values in the policy matrix…). The figure is quite aesthetically poor, nice examples of model representation can be found in Sutton and Barto (1998).**
* **Figure 4: It could be merged with the figure 1.**

Regarding Fig. 1: We added a horizontal blue dashed line between BA9 and BA10, and paste the updated figure here for convenience:

CENTRE DE RECHERCHE
SACLAY - ÎLE-DE-FRANCE
Parc Orsay Université
4, rue Jacques Monod
91893 Orsay cedex
Tél. : +33 (0)1 72 92 59 00
Fax : +33 (0)1 74 85 42 42
www.inria.fr

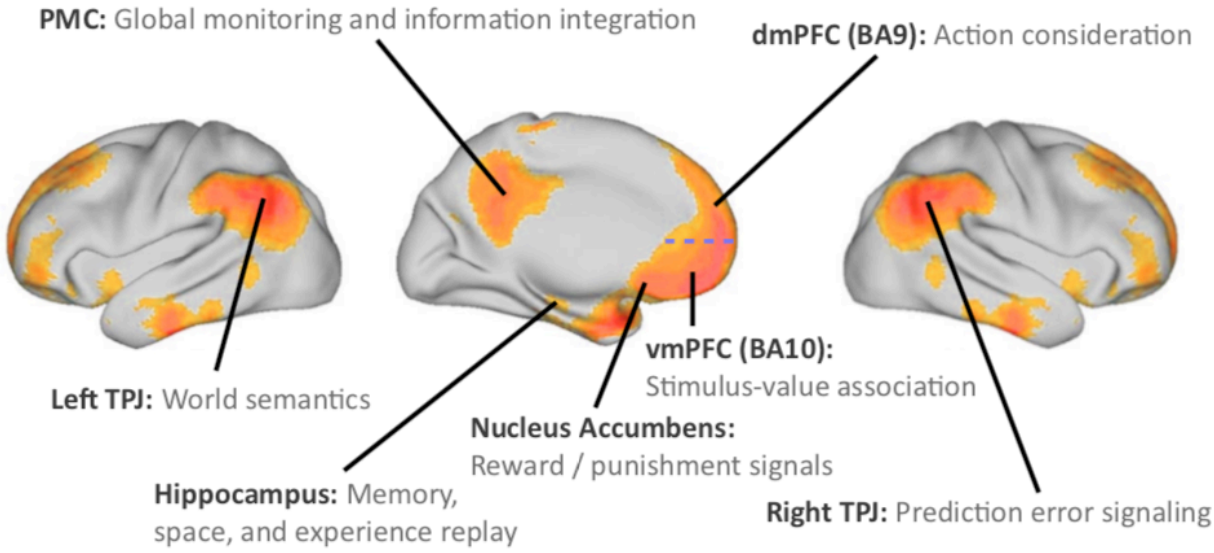


Fig 1. Default mode network: key functions. Neurobiological overview of the DMN with its major constituent parts and the associated functional roles relevant in our functional interpretation. The blue horizontal dashed line indicates the cytoarchitectonic border between the more dorsal BA9 and the more ventral BA10 (Brodmann, 1909).

Brodmann, K., 1909. Vergleichende Lokalisationslehre der Großhirnrinde. Barth, Leipzig.

Regarding Fig. 2: Cite related part of the text & improve caption www.github.com/banilo/darkcontrol_2018.

In response to the reviewer comment, we have clarified and extended the explanation and interpretation of the volume pattern prediction results in the revised manuscript.

"**Fig 2. Predictive structural association between reward system and default mode network.** Reward tasks (e.g., O'Doherty et al., 2015) and neural processing in the DMN (e.g., Buckner et al., 2008), often called "task-negative", have been studied so far in largely separate niches of the neuroscience literature. A currently underappreciated link is however suggested here based on 9,932 human subjects from the UK Biobank, inter-individual differences in left NAc volume ($R2 = 0.11 \pm 0.02$ [standard deviation across cross-validation folds]) and right NAc volume ($R2 = 0.14 \pm 0.02$) could be predicted from (z-scored) volume in the DMN regions. These out-of-sample generalizations reflect the expected performance in yet-to-be-observed individuals and obtained from linear support vector regression applied to normalized region volumes in the DMN in a 10-fold cross-validation procedure (Hastie et al., 2001). Consistent for the left and right reward system, NAc volume in a given subject is positively coupled with the vmPFC and HC. The congruence of our structural association results for both NAc targets speaks to the robustness of our pattern-prediction findings. The opposite relation of the left and right TPJ to the NAc appears to reflect a repeatedly recognized hemispheric asymmetry with respect to functional implications (Seghier, 2013),

CENTRE DE RECHERCHE
SACLAY - ÎLE-DE-FRANCE
Parc Orsay Université
4, rue Jacques Monod
91893 Orsay cedex
Tél. : +33 (0)1 72 92 59 00
Fax : +33 (0)1 74 85 42 42
www.inria.fr

impairments in neurological patients (Corbetta et al., 2000), different types of connectivity (Uddin et al., 2010; Caspers et al., 2011) as well as micro- and macroanatomy (Caspers et al., 2006, 2008). The colors are indicative of the (red = positive, blue = negative) and relative importance (the lighter the higher) of the regression coefficients. The code for reproduction and visualization: www.github.com/ banilo/darkcontrol_2018.

Buckner, R.L., Andrews-Hanna, J.R., Schacter, D.L., 2008. The brain's default network: anatomy, function, and relevance to disease. Ann N Y Acad Sci 1124, 1-38.

Caspers, S., Geyer, S., Schleicher, A., Mohlberg, H., Amunts, K., Zilles, K., 2006. The human inferior parietal cortex: cytoarchitectonic parcellation and interindividual variability. Neuroimage 33, 430-448.

Caspers, S., Eickhoff, S.B., Geyer, S., Scheperjans, F., Mohlberg, H., Zilles, K., Amunts, K., 2008. The human inferior parietal lobule in stereotaxic space. Brain Struct Funct 212, 481-495.

Caspers, S., Eickhoff, S.B., Rick, T., von Kapri, A., Kuhlen, T., Huang, R., Shah, N.J., Zilles, K., 2011. Probabilistic fibre tract analysis of cytoarchitectonically defined human inferior parietal lobule areas reveals similarities to macaques. Neuroimage 58, 362-380.

Corbetta, M., Kincade, J.M., Ollinger, J.M., McAvoy, M.P., Shulman, G.L., 2000. Voluntary orienting is dissociated from target detection in human posterior parietal cortex. Nat Neurosci 3, 292-297.

Hastie, T., Tibshirani, R., Friedman, J., 2001. The Elements of Statistical Learning. Springer Series in Statistics, Heidelberg, Germany.

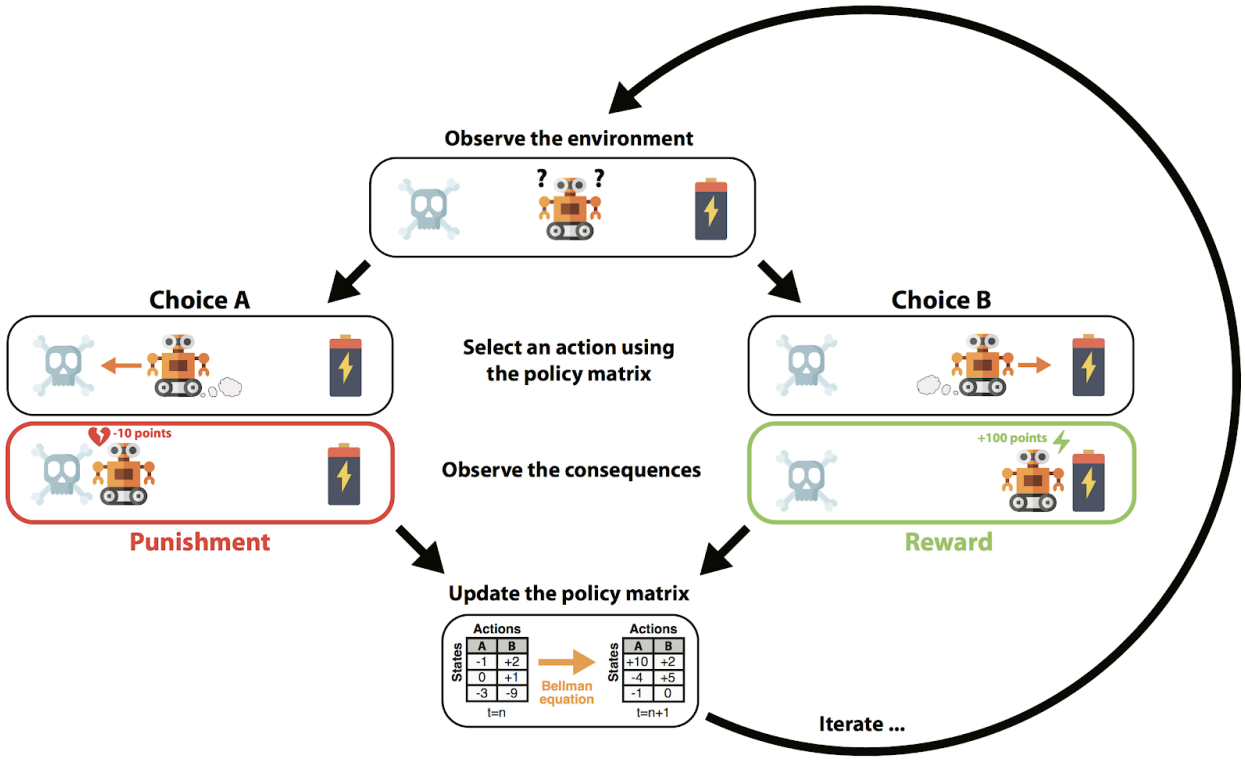
O'Doherty, J.P., Lee, S.W., McNamee, D., 2015. The structure of reinforcement-learning mechanisms in the human brain. Current Opinion in Behavioral Sciences 1, 94-100.

Seghier, M.L., 2013. The Angular Gyrus: Multiple Functions and Multiple Subdivisions. Neuroscientist, 43-61.

Uddin, L.Q., Supekar, K., Amin, H., Rykhlevskaia, E., Nguyen, D.A., Greicius, M.D., Menon, V., 2010. Dissociable connectivity within human angular gyrus and intraparietal sulcus: evidence from functional and structural connectivity. Cereb Cortex 20, 2636-2646.

Regarding Fig. 3: We entirely redesign the figure 3 to make it clearer and more aesthetic. We also changed its title to *"Illustration of Partially Observable Markov Decision Process (POMDP)."* The much enhanced figure follows:

CENTRE DE RECHERCHE
SACLAY - ÎLE-DE-FRANCE
Parc Orsay Université
4, rue Jacques Monod
91893 Orsay cedex
Tél. : +33 (0)1 72 92 59 00
Fax : +33 (0)1 74 85 42 42
www.inria.fr



Regarding Fig. 4: We specifically wanted to keep the Fig.1 and 4 separated to emphasize the global process model approach compare to local functional account. They reflect the narrative starting from biological insight - with the quintessential summaries as candidate functional descriptions" to the MDP account form a formal perspective which then adds the equation components from the middle part to the functional candidate description.

6) The following paragraph, page 1, line 58, might be removed or better related to the model, since the question of difference between electrophysiological signals and BOLD are not further assessed in the manuscript.
"Despite observation of similar large-scale networks of co-varying spontaneous activity in electrophysiological investigations (De Pasquale et al., 2010; Brookes et al., 2011; Baker et al., 2014), the link between the fMRI BOLD signal and population-level neural activity is still unclear. If those frequency-specific electrophysiological correlations are proposed as complementary to those observed with BOLD (Hipp and Siegel, 2015), their role in DMN function remains elusive (Maldjian et al., 2014)."

Thanks to this reviewer comment, we realized that the referred comments on electrophysiological properties of DMN correlates were out of scope in that part of the manuscript. We have therefore deleted the cited part from to "Despite observation...in DMN function remains elusive (Maldjian et al., 2014)." The revised passage reads as followed:

CENTRE DE RECHERCHE
SACLAY - ÎLE-DE-FRANCE
Parc Orsay Université
4, rue Jacques Monod
91893 Orsay cedex
Tél. : +33 (0)1 72 92 59 00
Fax : +33 (0)1 74 85 42 42
www.inria.fr

Page 1-2: “This dark matter of brain physiology (Raichle, 2006) begs the question of the biological purpose underlying DMN activity, which however remains elusive at the electrophysiological level. [deleted] What has early been described as the stream of consciousness" in psychology (James, 1890) found a potential...”

7) It remains unclear in the proposed model whether the value system is part of the DMN or just connected to it. It seems in the manuscript that the vmPFC and NA do not have the same status regarding this point? The link between the DMN and other brain regions or networks for each main function implemented by the MDP may be clarified, for instance in a figure.

Thanks to this reviewer comment, we have made the figure captions more explicit about how reward-related information my enter the DMN based on neuroanatomical knowledge.

Caption of figure 1/4: “Axonal tracing in monkeys and diffusion tractography in humans suggested that the NAc of the reward circuitry has monosynaptic fiber connections to the vmPFC (Haber et al., 1995; Croxson et al., 2005). Evaluation of propagated value information and triggered affective states encoded in the vmPFC may then feed into the functionally connected partner nodes of the DMN, such as the dmPFC and PMC (Andrews-Hanna et al., 2010; Bzdok et al., 2013).”

Andrews-Hanna, J.R., Reidler, J.S., Sepulcre, J., Poulin, R., Buckner, R.L., 2010. Functional-anatomic fractionation of the brain's default network. Neuron 65, 550-562.

Bzdok, D., Langner, R., Schilbach, L., Engemann, D.A., Laird, A.R., Fox, P.T., Eickhoff, S.B., 2013. Segregation of the human medial prefrontal cortex in social cognition. Front Hum Neurosci 7, 232.

8) On page 3 lines 18-22 "we also contribute to the discussion of DMN function by providing some of the first empirical evidence that morphological variability in DMN regions is linked to the reward circuitry". The authors may want to rephrase this sentence as this paper does not provide any empirical evidence (this is an unvalidated theoretical model) and does not describe hypotheses about anatomical variability.

We changed to sentence to *"*We also contribute to the discussion of DMN function by providing tentative evidence that variation of the grey-matter volume in DMN regions is linked to the reward circuitry*"* (Page 3).

9) The DMN can be identified through several ways. As mentioned by the authors, it is a network that responds positively during resting-state and negatively during a cognitive task. Simple contrast analyses can be used to isolate this network. Another way to identify it is to use functional connectivity analysis during resting-state. Each component of the DMN is globally more connected to all the other components compared to other brain regions. The specific pattern of connectivity of this network is not addressed in the manuscript. Basically, the (not so trivial) question that deserves

CENTRE DE RECHERCHE
SACLAY - ÎLE-DE-FRANCE
Parc Orsay Université
4, rue Jacques Monod
91893 Orsay cedex
Tél. : +33 (0)1 72 92 59 00
Fax : +33 (0)1 74 85 42 42
www.inria.fr

a discussion is: why brain regions implementing different types of computation should be more connected to each other?

We are grateful to the reviewer for pointing out a missing aspect in our work. We feel that the paper has quite a trans-disciplinary setup integrating work from invasive animal experiments from various species, high-level cognition in humans to neurophysiological properties of the cerebral cortex. As such, we feel that this already long perspective paper cannot easily bare even more levels of prior evidence and discussed streams of research. To incorporate the reviewer's suggestions, we have now added the following dye-tracing and other anatomical finding from monkeys and provide an explicit reference to exhaustive discussion of the connectivity aspects of DMN.

Page 2: "What has early been described as the "stream of consciousness" in psychology (James, 1890) found a potential neurobiological manifestation in the DMN (Shulman et al., 1997; Raichle et al., 2001). Axonal tracing injection in such parts of the association cortex in monkeys (Buckner & Krienen, 2013) were shown to resemble connectivity links between nodes of the human DMN (see here for details on anatomical connections: Buckner et al., 2008). Additionally, myelination patterns of axon connections were found to finish particularly late in these cortical areas (Flechsig, 1920), often believed to suggest sophistication of subserved neural processes (Sowell et al., 2003; Yakovlev, 1967). We propose that this set of some of the most advanced regions in the association cortex (Mesulam, 1998; Margulies et al., 2016b) are responsible for higher-order control of human behavior. Our functional account follows the notion of…"

Buckner, R.L., Krienen, F.M., 2013. The evolution of distributed association networks in the human brain. Trends Cogn Sci 17, 648-665.

Buckner, R.L., Andrews-Hanna, J.R., Schacter, D.L., 2008. The brain's default network: anatomy, function, and relevance to disease. Ann N Y Acad Sci 1124, 1-38.

Flechsig, P., 1920. Anatomie des menschlichen Gehirns und Rückenmarks auf myelogenetisch Grundlage. Thieme, Leipzig.

Sowell ER, et al. (2003) Mapping cortical change across the human life span. Nat neurosci 6(3):309.

Yakovlev P (1967) The myelogenetic cycles of regional maturation of the brain. Regional development of the brain in early life:3-70.

In agreement with the reviewer comment, we now also emphasize the prominent intra-network functional coupling strengths of the DMN and relate to our ideas revolving around Monte-Carlo sampling of possible events to adapt the organism's behavioral profiles for the benefit of future action outcomes.

Summary (page 15): "The more familiar and predictable the current environment, the more brain resources may remain for allocating DMN activity to MDP processes extending beyond the present time and sensory context. This speculation receives quantitative support in that connectional links between nodes of the DMN have been reported to be more consistent and reliable than functional couplings within any other macroscopical networks (Shehzad et al., 2009). As such, random-sampling-related baseline evaluation of possibilities and events may

CENTRE DE RECHERCHE
SACLAY - ÎLE-DE-FRANCE
Parc Orsay Université
4, rue Jacques Monod
91893 Orsay cedex
Tél. : +33 (0)1 72 92 59 00
Fax : +33 (0)1 74 85 42 42
www.inria.fr

be subserved by the DMN and get partly suspended when novelty in the external environment is encountered or immediate action is required. In line with this perspective, DMN engagement was shown to heighten and relate to effective behavioral responses in the practiced phase of a demanding cognitive flexibility task, as compared to acquisition phase when participants learned context-specific rules."

Shehzad, Z., Kelly, A.M., Reiss, P.T., Gee, D.G., Gotimer, K., Uddin, L.Q., Lee, S.H., Margulies, D.S., Roy, A.K., Biswal, B.B., Petkova, E., Castellanos, F.X., Milham, M.P., 2009. The resting brain: unconstrained yet reliable. Cereb Cortex 19, 2209-2229.

Reviewer #2: The article is well written and includes figures that complement in a concise and clear way the information present in the text. The manuscript integrates different levels of research, from electrophysiological recordings in animals to lesion studies in humans. The hypotheses and proposals for future research are comprehensive and are built based on good evidence.

I recommend the publication of this article with some minor suggestions:

We are thankful to the reviewer for the positive feedback. We have taken into account all their remarks, corrections, and suggestions, and updated the manuscript accordingly. Our narrative is now much improved.

Page 2, line 35 - "The present work adopts the perspective of a human agent faced with the choice of the next actions and guided by outcomes of really happened, hypothetically imagined, and expected futures to optimize behavioral performance.": The sentence is long and not easy to understand at the first time. I suggest its reformulation or its separation in two sentences.

The sentence has been simplified and broken down into two sentences.

"The present work adopts the control-theoretical perspective of a human agent faced with the choice of the next actions guided by outcomes to optimize behavioral performance. These outcomes can be really experienced, hypothetically imagined, or expected in the future."

Page 2E, line 37 - "Formally, we propose reinforcement learning to be a particularly attractive framework to describe, quantify, DMN function and the brain.": I suspect some word is missing in this sentence (maybe an "and" instead of the last comma).

Indeed this sentence is wrong. It has been rephrased in the manuscript update.

"Formally, we propose reinforcement learning as a particularly attractive framework for describing, containing, and quantifying the unknown function underlying DMN activity."

Page 3E, line 5: The abbreviation "RL" is not defined. It is defined some pages later only (page 8, line 13). It should be stated the first time it appears in the text for an easier reading.

CENTRE DE RECHERCHE
SACLAY - ÎLE-DE-FRANCE
Parc Orsay Université
4, rue Jacques Monod
91893 Orsay cedex
Tél. : +33 (0)1 72 92 59 00
Fax : +33 (0)1 74 85 42 42
www.inria.fr

We define RL (=reinforcement learning) at the first mention of the acronym, at the beginning of section 3.

Page 6E, line 38 "Mental scenes created by neurological patients with HC lesion exposed a lack of spatial integrity, richness in detail, and overall coherence.": A reference supporting this sentence should be added (Hassabis et al, 2007, previously used in the manuscript, for instance)

The Hassabis et al. 2017 reference has been inserted here as advised by the reviewer.

Page 11E, line 6 "This DMN region has direct connections to the NAc, known to be involved in reward evaluation.": A reference supporting this sentence should be added. The word "to" is repeated.

We have cited Carlezon et al. 2009, Haber et al. 2005, and Croxson et al. 2005 to justify the sentence.

Additionally, the word repetition has been removed.

Page 15E: There is no reference to figure 4 in the text. I suggest it to an easier reading.

We added a reference to figure 4 in the text.

Page 16E: There is no reference to figure 5 in the text. I suggest it to an easier reading. References: the references are ordered according to the surname, however the first name is stated first. It makes the finding of the desired reference more difficult. I suggest that the first names should be reduced to initials and appear after the surname.

The missing cross-references to some of the figures (e.g., Fig. 4) was a bug and has been fixed.

CENTRE DE RECHERCHE
SACLAY - ÎLE-DE-FRANCE
Parc Orsay Université
4, rue Jacques Monod
91893 Orsay cedex
Tél. : +33 (0)1 72 92 59 00
Fax : +33 (0)1 74 85 42 42

www.inria.fr

*Title page

Dark Control: The Default Mode Network as a Reinforcement Learning Agent

Elvis Dohmatob[1,2], Guillaume Dumas[5,6,7,8], Danilo Bzdok[1,2,3,4]

1 INRIA, Parietal Team, Saclay, France
2 Neurospin, CEA, Gif-sur-Yvette, France
3 Department of Psychiatry, Psychotherapy and Psychosomatics, RWTH Aachen University, Aachen, Germany
4 JARA-BRAIN, Jülich-Aachen Research Alliance, Germany
5 Institut Pasteur, Human Genetics and Cognitive Functions Unit, Paris, France
6 CNRS UMR 3571 Genes, Synapses and Cognition, Institut Pasteur, Paris, France
7 University Paris Diderot, Sorbonne Paris Cité, Paris, France
8 Centre de Bioinformatique, Biostatistique et Biologie Intégrative, Paris, France

Abstract

The default mode network (DMN) is believed to subserve the baseline mental activity in humans. Its highest energy consumption compared to other brain networks and its intimate coupling with conscious awareness are both pointing to an overarching function. Many research streams speak in favor of an evolutionarily adaptive role in envisioning experience to anticipate the future. In the present work, we propose a *process model* that tries to explain *how* the DMN may implement continuous evaluation and prediction of the environment to guide behavior. The main purpose of the DMN, we argue, may be to perform optimization based on Markov Decision Processes through vicarious trial and error. Our formal account of DMN function naturally accommodates as special cases previous interpretations based on (1) predictive coding, (2) semantic associations, and (3) a sentinel role. Moreover, this process model for the neural optimization of complex behavior in the DMN offers parsimonious explanations for recent experimental findings in animals and humans.

keywords: systems neuroscience, artificial intelligence, reinforcement learning, mind-wandering

Dark Control: The Default Mode Network as a Reinforcement Learning Agent

Elvis Dohmatob[1,2], Guillaume Dumas[5,6,7,8], Danilo Bzdok[1,2,3,4]

1 INRIA, Parietal Team, Saclay, France
2 Neurospin, CEA, Gif-sur-Yvette, France
3 Department of Psychiatry, Psychotherapy and Psychosomatics, RWTH Aachen University, Aachen, Germany
4 JARA-BRAIN, Jülich-Aachen Research Alliance, Germany
5 Institut Pasteur, Human Genetics and Cognitive Functions Unit, Paris, France
6 CNRS UMR 3571 Genes, Synapses and Cognition, Institut Pasteur, Paris, France
7 University Paris Diderot, Sorbonne Paris Cité, Paris, France
8 Centre de Bioinformatique, Biostatistique et Biologie Intégrative, Paris, France

Abstract

The default mode network (DMN) is believed to subserve the baseline mental activity in humans. Its higher energy consumption compared to other brain networks and its intimate coupling with conscious awareness are both pointing to an unknown overarching function. Many research streams speak in favor of an evolutionarily adaptive role in envisioning experience to anticipate the future. In the present work, we propose a *process model* that tries to explain *how* the DMN may implement continuous evaluation and prediction of the environment to guide behavior. The main purpose of DMN activity, we argue, may be described by Markov Decision Processes that optimize action policies via value estimates based through vicarious trial and error. Our formal perspective on DMN function naturally accommodates as special cases previous interpretations based on (1) predictive coding, (2) semantic associations, and (3) a sentinel role. Moreover, this process model for the neural optimization of complex behavior in the DMN offers parsimonious explanations for recent experimental findings in animals and humans.

keywords: systems neuroscience, artificial intelligence, mind-wandering

1 Introduction

In the absence of external stimulation, the human brain is not at rest. At the turn to the 21st century, brain-imaging may have been the first technique to allow for the discovery of a unique brain network that would subserve baseline mental activities (Raichle et al., 2001; Buckner et al., 2008; Bzdok and Eickhoff, 2015). The "default mode network" (DMN) continues to metabolize large quantities of oxygen and glucose energy to maintain neuronal computation during free-ranging thought (Kenet et al., 2003; Fiser et al., 2004). The baseline energy demand is only weakly modulated at the onset of defined psychological tasks (Gusnard and Raichle, 2001). At its opposite, during sleep, the decoupling of brain structures discarded the idea of the DMN being only a passive network resonance and rather supported an important role in sustaining conscious awareness (Horovitz et al., 2009).

This *dark matter of brain physiology* (Raichle, 2006) begs the question of the biological purpose underlying neural activity in the DMN, which however still remains elusive at the electrophysiological level (De Pasquale et al., 2010; Brookes et al., 2011; Baker et al., 2014). What has early been described as the "stream of consciousness" in psychology (James, 1890)

found a potential neurobiological manifestation in the DMN (Shulman et al., 1997; Raichle et al., 2001). Axonal tracing injection in such parts of the association cortex in monkeys were shown to resemble connectivity links between nodes of the human DMN (see here for details on anatomical connections: (Buckner et al., 2008)). Additionally, myelination patterns of axon connections were found to complete particularly late in these cortical areas (Flechsig, 1920), often believed to reflect sophistication of subserved neural processes (Sowell et al., 2003; Yakovlev, 1967). We propose that this set of some of the most advanced regions in the association cortex (Mesulam, 1998; Margulies et al., 2016b) are responsible for higher-order control of human behavior. Our perspective therefore follows the notion of "a hierarchy of brain systems with the DMN at the top and the salience and dorsal attention systems at intermediate levels, above thalamic and unimodal sensory cortex" (Carhart-Harris and Friston, 2010).

2 Towards a formal account of default mode function: higher-order control of the organism

The network nodes that compose the human DMN are hubs of high baseline neural activity, which typically decreases when engaged in well-defined psychological experiments (Gusnard and Raichle, 2001). The standard mode of neural information maintenance and manipulation has been argued to mediate evolutionarily conserved functions (Brown, 1914; Binder et al., 1999; Buzsáki, 2006). Today, many psychologists and neuroscientists believe that the DMN implements some form of probabilistic estimation of past, hypothetical, and future events (Fox et al., 2005; Hassabis et al., 2007; Schacter et al., 2007; Binder et al., 2009; Buckner et al., 2008; Spreng et al., 2009). This brain network might have emerged to continuously predict the environment using mental imagery as an evolutionary advantage (Suddendorf and Corballis, 2007). However, information processing in the DMN has also repeatedly been shown to directly impact human behavior. Goal-directed task performance improved with decreased activity in default mode regions (Weissman et al., 2006) and increased DMN activity was linked to more task-independent, yet sometimes useful thoughts (Mason et al., 2007; Seli et al., 2016). Gaining insight into DMN function is particularly challenging because this brain network appears to simultaneously influence perception-action cycles in the present and to support mental travel across time, space, and content domains (Boyer, 2008).

We aim at proposing an alternative to reasoning about the DMN based on longstanding cognitive theory. The present work adopts the control-theoretical perspective of a human *agent* faced with the choice of the next actions guided by outcomes to optimize behavioral performance. These outcomes can be really experienced, hypothetically imagined, or expected in the future. Formally, we propose reinforcement learning as a particularly attractive framework for describing, containing, and quantifying the unknown function underlying DMN activity. An intelligent agent improves the interaction with the environment by continuously updating its computation of value estimates and action predispositions through integration of feedback outcomes. That is, "[agents], with their actions, modify the environment and in doing so partially determine their next stimuli, in particular stimuli that are necessary for triggering the next action" (Pezzulo, 2011). Agents with other behavioral policies therefore sample different distributions of action-perception trajectories (Ghavamzadeh et al., 2015). Henceforth, *control* refers to the influence that an agent exerts by interacting with the environment to reach preferred states.

At the psychological level, the more the ongoing executed task is unknown and unpracticed, the less stimulus-independent thoughts occur (Filler and Giambra, 1973; Teasdale et al., 1995; Christoff et al., 2016). Conversely, it has been empirically shown that, the more the world is easy to foresee, the more human mental activity becomes detached from the actual sensory environment (Antrobus et al., 1966; Pope and Singer, 1978; Mason et al., 2007; Weissman et al., 2006). Without requiring explicit awareness, these "offline" processes may contribute to optimizing control of the organism in general. We formalize a *policy matrix* to capture the space of possible actions that the agent can perform on the environment given the current state. A *value function* maps environmental objects and events (i.e., states) to expected reward outcomes. Switching between states reduces to a sequential processing model. Informed by outcomes of performed actions, neural computation reflected in DMN dynamics could be

constantly shaped by prediction error through feedback loops. The present computational account of DMN function will be described in the mathematical framework of Markov Decision Processes (MDP). MDPs specifically formalize decision making in stochastic contexts with reward feedback.

Such a RL perspective on DMN activity can naturally embed human behavior into the tension between exploitative action with immediate gains and exploratory action with longer-term gratification. We argue that DMN implication in many of the most advanced human capacities can be recast as prediction error minimization informed by internally generated probabilistic simulations - "covert forms of action and perception" (Pezzulo, 2011) -, allowing maximization of action outcomes across different time scales. Such a purposeful optimization objective may be solved by a stochastic approximation based on a brain implementation of Monte Carlo sampling. Even necessarily imperfect memory recall, random day-time mind-wandering, and seemingly arbitrary dreams during sleep may provide randomly sampled blocks of pseudo-experience that are instrumental to iteratively optimize the behavioral agenda of the organism.

Evidence from computational modeling of human behavior (Körding and Wolpert, 2004) and cell recording experiments in ferrets (Fiser et al., 2004) suggest that much of brain activity is dedicated to "the development and maintenance of [a] probabilistic model of anticipated events" (Raichle and Gusnard, 2005). The present paper proposes a process model that satisfies this previously proposed contention. We also contribute to the discussion of DMN function by providing tentative evidence that variation of the grey-matter volume in DMN regions is linked to the reward circuitry (Fig. 2), thus linking two literatures with currently scarce cross-references. Finally, we derive explicit hypotheses that could be tested in targeted neuroscience experiments in the future and we detail how our process model relates to previous cognitive and theoretical accounts of DMN function.

Please appreciate the importance of differentiating which levels of observation are at play in the present account. A process model is not solely intended to capture behavior of the agent, such as cognitive accounts of DMN function, but also the neurocomputational specifics of the agent. Henceforth, we will use "inference" when referring to aspects of the statistical model, "prediction" when referring to the neurobiological implementation, and words like "forecast" or "forsee" when referring to the cognitive behavior of the agent. It is moreover important to note that our account does not claim that neural activity in the DMN in particular or the brain in general are identical with reinforcement learning algorithms. Rather, we advocate feedback-based learning strategies as an attractive alternative perspective to describe, quantify, and interpret research findings on the DMN.

3 Known neurobiological properties of the default mode network

We begin by a neurobiological deconstruction of the DMN based on integrating experimental findings in the neuroscience literature from different species. This walkthrough across main functional zones of the DMN (i.e., de-emphasizing their precise anatomical properties) will outline the individual functional profiles with the goal of paving the way for their algorithmic interpretation in our formal account (section 3). As our focus will be on major *functional* zones of the DMN, please see elsewhere for excellent surveys on their *anatomical* boundaries and connectivity patterns (Buckner et al., 2008; Binder et al., 2009; Seghier, 2013).

3.1 The posteromedial cortex: global monitoring and information integration

The midline structures of the human DMN, including the posteromedial cortex (PMC) and the medial prefrontal cortex (mPFC), are probably responsible for highest turn-overs of energy consumption (Raichle et al., 2001; Gusnard and Raichle, 2001). These metabolic characteristics go hand-in-hand with brain-imaging findings that suggested the PMC and mPFC to potentially represent the functional core of the DMN (Andrews-Hanna et al., 2010; Hagmann et al., 2008).

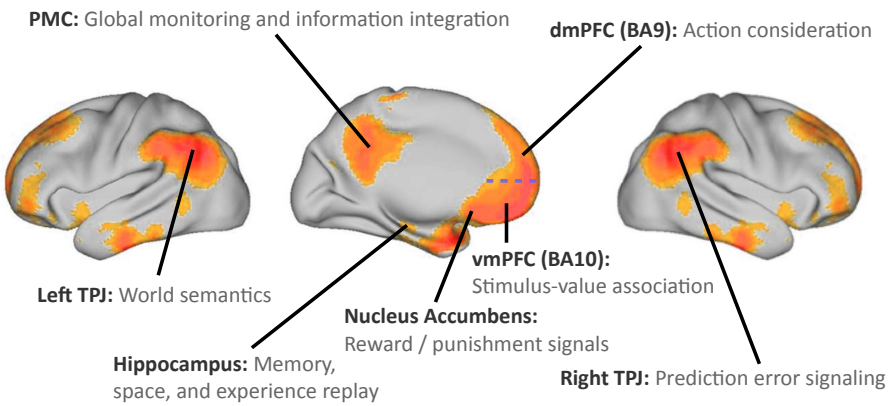


Fig 1. Default mode network: key functions. Neurobiological overview of the DMN with its major constituent parts and the associated functional roles relevant in our functional interpretation. The blue horizontal dashed line indicates the cytoarchitectonic border between the more dorsal BA9 and the more ventral BA10 (Brodmann, 1909). Axonal tracing in monkeys and diffusion tractography in humans suggested that the NAc of the reward circuitry has monosynaptic fiber connections to the vmPFC (Haber et al., 1995b; Croxson et al., 2005). Evaluation of propagated value information and triggered affective states encoded in the vmPFC may then feed into the functionally connected partner nodes of the DMN, such as the dmPFC and PMC (Andrews-Hanna et al., 2010; Bzdok et al., 2013b).

Normal and disturbed metabolic fluctuations in the human PMC have been closely related to changes of conscious awareness (Cavanna and Trimble, 2006; Leech and Sharp, 2014). Indeed, the PMC matures relatively late (i.e., myelination) during postnatal development in monkeys (Goldman-Rakic, 1987), which is generally considered to be a sign of evolutionary sophistication. This DMN region has long been speculated to reflect constant computation of environmental statistics and its internal representation as an inner "mind's eye" (Cavanna and Trimble, 2006; Leech and Sharp, 2014). For instance, Bálint's syndrome is a neurological disorder of conscious awareness that can result from tissue damage in the posterior medial cortex (Bálint et al., 1909; Buckner et al., 2008). Such neurological patients are plagued by an inability to bind various individual features of the visual environment into an integrated whole (i.e., simultanagnosia) as well as an inability to direct action towards currently unattended environmental objects (i.e., optic ataxia). Scanning complex scenes is impaired in that statistic or moving objects in the environment may be invisible or disappear in the subject perception of the patient (Mesulam, 2000; Blumenfeld, 2002). This dysfunction can be viewed as a high-level impairment in gathering information about alternative objects (i.e., exploration) as well as using these environmental opportunities towards a behavioral goal (i.e., exploitation). Congruently, the human PMC was coupled in two different functional connectivity analyses (Bzdok et al., 2015) with the amygdala, involved in significance evaluation, and the nucleus accumbens (NAc), involved in reward evaluation. Specifically, among all parts of the PMC, the ventral posterior cingulate cortex was most connected to the laterobasal nuclei group of the amygdala (Bzdok et al., 2015). This amygdalar subregion has been proposed to continuously scan environmental input for biological relevance assessment (Bzdok et al., 2013a; Ghods-Sharifi et al., 2009; Baxter and Murray, 2002).

The putative role of the PMC in continuous abstract integration of environmental relevance and ensuing top-level guidance of action on the environment is supported by many neuroscience experiments. Electrophysiological recordings in animals implicated PMC neurons in strategic decision making (Pearson et al., 2009), risk assessment (McCoy and Platt, 2005), outcome-dependent behavioral modulation (Hayden et al., 2009), as well as approach-avoidance behavior (Vann et al., 2009). Neuron spiking activity in the PMC allowed distinguishing

whether a monkey would pursue an exploratory or exploitative behavioral strategy during food foraging (Pearson et al., 2009). Further, single-cell recordings in the monkey PMC demonstrated this brain region's sensitivity to subjective target utility (McCoy and Platt, 2005) and integration across individual decision-making instances (Pearson et al., 2009). This DMN region encoded the preference for or aversion to options with uncertain reward outcomes and its neural spiking activity was more associated with subjectively perceived relevance of a chosen object than by its actual value, based on an "internal currency of value" (McCoy and Platt, 2005). In fact, direct stimulation of PMC neurons in monkeys promoted exploratory actions, which would otherwise be shunned (Hayden et al., 2008). Graded changes in firing rates of PMC neurons indicated changes in upcoming choice trials, while their neural patterns were distinct from neuronal spike firings that indicated choosing either option. Similarly in humans, the DMN has been shown to gather and integrate information over different parts of auditory narratives in an fMRI study (Simony et al., 2016).

Moreover, the retrosplenial portion of the PMC could support representation of action possibilities and evaluation of reward outcomes by integrating information from memory recall and different perspective frames. Regarding memory recall, retrosplenial damage has been consistently associated with anterograde and retrograde memory impairments of various kinds of sensory information in animals and humans (Vann et al., 2009). Regarding perspective frames, the retrosplenial subregion of the PMC has been proposed to mediate between the organism's egocentric (i.e., focused on external sensory environment) and allocentric (i.e., focused on internal world knowledge) viewpoints in animals and humans (Epstein, 2008; Burgess, 2008; Valiquette and McNamara, 2007).

Consequently, the PMC may contribute to overall DMN function by monitoring the subjective outcomes of possible actions and integrating that information with memory and perspective frames into short- and longer-term behavioral agendas. Estimated value, found to differs across individuals, might enrich statistical assessment of the environment to map and predict delayed reward opportunities in the future. Viewed from a RL perspective, the PMC may continuously adapt the organism to changes in both the external environment and its internal representation to enable strategic behavior.

3.2 The prefrontal cortex: action consideration and stimulus-value association

Analogous to the PMC, the dorsomedial PFC (dmPFC, related to BA9) of the DMN is believed to subserve multi-sensory processes across time, space, and content domains to exert top-level control on behavior. Comparing to the PMC, however, dmPFC function may be closer to a "mental sketchpad" (Goldman-Rakic et al., 1996). This DMN part potentially subserves the de-novo construction and manipulation of meaning representations instructed by stored semantics and memories (Bzdok et al., 2013b). The dmPFC may subserve representation and assessment of one's own and other individuals' action considerations - a necessary component of a full-blown RL agent. Generally, neurological patients with tissue damage in the prefrontal cortex are known to struggle with adaptation to new stimuli and events (Stuss and Benson, 1986). Specifically, neural activity in the human dmPFC reflected expectations about other peoples' actions and outcomes of these predictions. Neural activity in the dmPFC indeed explained the performance decline of inferring other peoples' thoughts in aging humans (Moran et al., 2012). Certain dmPFC neurons in macaque monkeys exhibited a preference for processing others', rather than own, action with fine-grained adjustment of contextual aspects (Yoshida et al., 2010).

Comparing to the dmPFC, the ventromedial PFC (vmPFC, related to BA10) is probably more specifically devoted to subjective value evaluation and risk estimation of relevant environmental stimuli (Fig. 1 and 2). The ventromedial prefrontal DMN may subserve adaptive behavior by bottom-up-driven processing of what matters now, drawing on sophisticated value representations (Kringelbach and Rolls, 2004; O'Doherty et al., 2015). Quantitative lesion findings across 344 human individuals confirmed a substantial impairment in value-based action choice (Gläscher et al., 2012). Indeed, this DMN region is preferentially connected with reward-related and limbic regions. The vmPFC is well known to have direct connections with the NAc in axonal tracing studies in monkeys (Haber et al., 1995a). Congruently, the

gray-matter volume of the vmPFC and NAc correlated with indices of value-guided behavior and reward attitudes in humans (Lebreton et al., 2009). NAc activity is further thought to reflect reward prediction signals from dopaminergic neurotransmitter pathways (Schultz, 1998) that not only channel action towards basic survival needs but also enable more abstract reward processings, and thus perhaps RL, in humans (O'Doherty et al., 2015).

Consistently, diffusion MRI tractography in humans and monkeys (Croxson et al., 2005) quantified the NAc to be more connected to the vmPFC than dmPFC in both species. Two different functional connectivity analyses in humans also revealed strong vmPFC connections with the NAc, hippocampus (HC), and PMC (Bzdok et al., 2015). In line with these connectivity findings in animals and humans, the vmPFC is often proposed to represent triggered emotional and motivational states (Damasio et al., 1996). Such real or imagined arousal states could be mapped in the vmPFC as a bioregulatory disposition influencing cognition and decision making. In neuroeconomic studies of human decision making, the vmPFC consistently reflects an individuals subjective value predictions (Behrens et al., 2008). This finding may also explain why performance within and across participants was reported to relate to state encoding in the vmPFC (Schuck et al., 2016). Such a "cognitive map" of the action space - an integral part of a RL agent - was argued to encode the current task state even when states are unobservable from the sensory environment.

3.3 The hippocampus: memory, space, and experience replay

The DMN midline has close functional links with the HC (henceforth implying to include also parahippocampal regions) in the medial temporal lobe (Vincent et al., 2006; Shannon et al., 2013) —a region long known to be involved in memory operations and spatial navigation in animals and humans. While the HC is traditionally believed to allow recalling past experience, there is now increasing evidence for an important role in constructing mental models in general (Zeidman and Maguire, 2016; Schacter et al., 2007; Gelbard-Sagiv et al., 2008; Javadi et al., 2017; Boyer, 2008). Its recursive anatomical architecture may be specifically designed to allow reconstructing entire sequences of experience from memory fragments. Indeed, hippocampal damage was not only associated with an impairment in re-experiencing the past (i.e., amnesia), but also forecasting of one's own future and imagination of experiences more broadly (Hassabis et al., 2007).

Mental scenes created by neurological patients with HC lesion exposed a lack of spatial integrity, richness in detail, and overall coherence (c.f (Hassabis et al., 2007)). Single-cell recordings in the animal HC revealed constantly active neuronal populations whose firing coincided with specific locations in space during environmental navigation. Indeed, when an animal is choosing between alternative paths, the corresponding neuronal populations in the HC spike one after another (Johnson and Redish, 2007). Such neuronal patterns in the HC appear to directly indicate upcoming behavior, such as in planning navigational trajectories (Pfeiffer and Foster, 2013) and memory consolidation of choice relevance (De Lavilléon et al., 2015). Congruently, London taxi drivers, humans with high performance in forecasting spatial navigation, were shown to exhibit increased gray-matter volume in the HC (Maguire et al., 2000).

There is hence increasing evidence that HC function extends beyond simple forms of encoding and reconstruction of memory and space information. Based on spike recordings of hippocampal neuronal populations, complex spiking patterns can be followed across extended periods including their modification of input-free self-generated patterns after environmental events (Buzsáki, 2004). Specific spiking sequences, which were elicited by experimental task design, have been shown to be re-enacted spontaneously during quiet wakefulness and sleep (Hartley et al., 2014; ONeill et al., 2010). Moreover, neuronal spike sequences measured in hippocampal place cells of rats featured re-occurrence directly after experimental trials as well as directly before (prediction of) upcoming experimental trials (Diba and Buzsáki, 2007). Similar spiking patterns in hippocampal neurons during rest and sleep have been proposed to be critical in communicating local information to the neocortex for long-term storage, potentially including DMN regions. Moreover, in mice, invasively triggering spatial experience recall in the HC during sleep has been demonstrated to subsequently alter action choice during wakefulness (De Lavilléon et al., 2015). These HC-subserved mechanisms conceivably

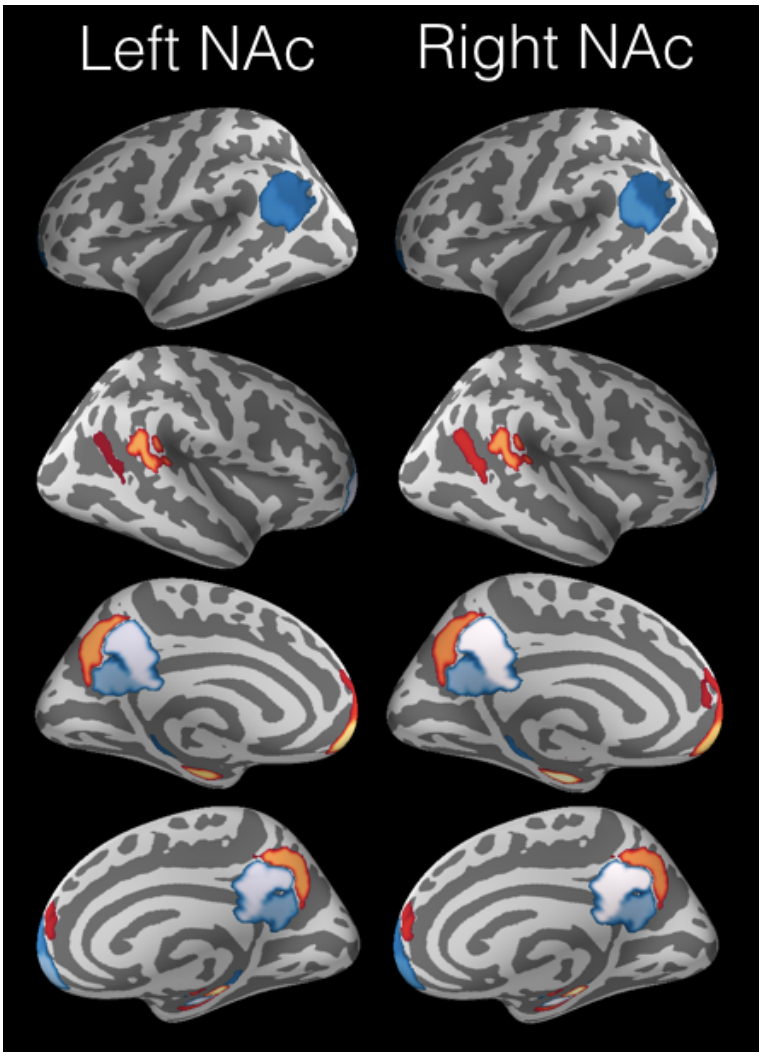


Fig 2. Predictive structural association between reward system and DMN nodes. Reward tasks (O'Doherty et al., 2015) and neural processing in the DMN (Buckner et al., 2008), often considered "task-negative", have been studied so far in largely separate niches of the neuroscience literature. A currently underappreciated link is however suggested here based on 9,932 human subjects from the UK Biobank, inter-individual differences in left NAc volume ($R^2 = 0.11 \pm 0.02$ [standard deviation across cross-validation folds]) and right NAc volume ($R^2 = 0.14 \pm 0.02$) could be predicted from (z-scored) volume in the DMN regions. These out-of-sample generalizations reflect the expected performance in yet-to-be observed individuals obtained from linear support vector regression applied to normalized region volumes in the DMN in a 10-fold cross-validation procedure (Hastie et al., 2011). Consistent for the left and right reward system, NAc volume in a given subject is positively coupled with the vmPFC and HC. The congruence of our structural association results for both NAc targets speaks to the robustness of our pattern-prediction findings. The opposite relation of the left and right TPJ to the NAc appears to reflect a repeatedly recognized hemispheric asymmetry with respect to functional implications (Seghier, 2013), impairments in neurological patients (Corbetta et al., 2000), different types of connectivity (Uddin et al., 2010; Caspers et al., 2011) as well as micro- and macroanatomy (Caspers et al., 2006, 2008). The colors are indicative of the (red = positive, blue = negative) and relative importance (the lighter the higher) of the regression coefficients. The code for reproduction and visualization: `www.github.com/banilo/darkcontrol_2018`.

contribute to advanced cognitive processes that require re-experiencing or newly constructed mental scenarios, such as in recalling autobiographical memory episodes (Hassabis et al., 2007). Within a RL framework, the HC could thus orchestrate re-experience of environmental aspects for consolidations based on re-enactment and for integration into rich mental scene construction (Deuker et al., 2016; Bird et al., 2010). In this way, the HC may impact ongoing perception of and action on the environment (Zeidman and Maguire, 2016; De Lavilléon et al., 2015).

3.4 The right and left TPJ: prediction error signaling and world semantics

The DMN emerges with its midline structures early in human development (Doria et al., 2010), while the right and left TPJs may become fully functionally integrated into this macroscopical network only after birth. The TPJs are known to exhibit hemispheric differences based on microanatomical properties and cortical gyrification patterns (Seghier, 2013). In general, neuroscientific investigations on hemispheric functional specialization have highlighted the right cerebral hemisphere as more dominant for attentional functions and the left side more for semantic functions (Seghier, 2013; Bzdok et al., 2013c, 2016a; Stephan et al., 2007).

The TPJ in the right hemisphere (RTPJ) denotes a broad functional zone with varying anatomical nomenclature (Mars et al., 2011; Seghier et al., 2010; Seghier, 2013) that has been shown to be closely related to multi-sensory event representation and prediction error signaling (Downar et al., 2000; Vetter et al., 2011; Shulman et al., 2010, 2007). This DMN region is probably central for action initiation during goal-directed psychological tasks and for sensorimotor behavior by integrating multi-sensory attention (Corbetta and Shulman, 2002). Its involvement was repeatedly reported in monitoring multi-step action execution (Hartmann et al., 2005), visuo-proprioceptive conflict (Balslev et al., 2005), spatial re-orientation (Corbetta et al., 2000), and detection of environmental changes across visual, auditory, or tactile stimulation (Downar et al., 2000). Direct electrical stimulation of the human RTPJ during neurosurgery was associated with altered perception and stimulus awareness (Blanke et al., 2002). It was argued that the RTPJ encodes actions and predicted outcomes, without necessarily relating these neural processes to value estimation (Liljeholm et al., 2013; Hamilton and Grafton, 2008; Jakobs et al., 2009). More specifically, neural activity in the RTPJ has been proposed to reflect stimulus-driven attentional reallocation to self-relevant and unexpected sources of information as a circuit breaker that recalibrates functional control of brain networks (Bzdok et al., 2013c; Corbetta et al., 2008). In the face of large discrepancies between actual and previously predicted environmental events, the RTPJ acts as a potential switch between externally-oriented mind sets focussed on the sensory environment and internally-oriented mind sets focussed on mental scene construction. For instance, temporally induced RTPJ damage in humans diminished the impact of predicted intentions of other individuals (Young et al., 2010), a capacity believed to be enabled by the DMN. Viewed from a RL perspective, the RTPJ might reflect an important relay that shifts away from the internally directed baseline processes to, instead, deal with unexpected environmental cues and events.

The left TPJ of the DMN (LTPJ), in turn, may have a functional relationship to Wernicke's area involved in semantic processes (Blumenfeld, 2002) and has been described as "a temporoparietal transmodal gateway for language" by some investigators (Mesulam, 2000). Neurological patients with damage in this region have a major impairment of language comprehension when listening to others or reading a book. Patient speech preserves natural rhythm and normal syntax, yet the voiced sentences lack meaning (i.e., aphasia). Abstracting from speech interpretations in linguistics and neuropsychology, the LTPJ appears to mediate access to and binding of world knowledge, such as required during action considerations (Binder and Desai, 2011; Seghier, 2013). Consistent with this view, LTPJ damage in humans also entailed problems in recognizing others' pantomimed action towards objects without obvious relation to processing explicit language content (Varney and Damasio, 1987). Inner speech also hinges on knowledge recall about the physical and social world. Indeed, the internal production of verbalized thought ("language of the mind") was closely related to the LTPJ in a pattern analysis of brain volume (Geva et al., 2011). Further, episodic memory recall and mental imagery to forecast future events strongly draw on re-assembling world knowledge. Isolated building blocks of world structure get rebuilt in internally constructed mental scenarios that

guide present action choice, weigh hypothetical possibilities, and forecast event outcomes. As a candidate component of a RL agent, neural processes in the LTPJ may contribute to the automated predictions of the environment by incorporating experience-derived building blocks of world regularities into ongoing action, planning, and problem solving.

4 Reinforcement learning control: a process model for DMN function

We argue the outlined neurobiological properties of the DMN regions to be sufficient for implementing all components of a full-fledged reinforcement learning (RL) system. Recalling past experience, considering candidate actions, random sampling of possible experiences, as well as estimation of instantaneous and delayed reward outcomes are key components of intelligent RL agents that are plausible to functionally intersect in the DMN.

RL is an area of machine learning concerned with searching optimal behavioral strategies through interactions with an *environment* with the goal to maximize the *cumulative reward* over time (Sutton and Barto, 1998). Optimal behavior typically takes the future into account as certain rewards could be *delayed*. Through repeated action on and feedback from the environment, the agent learns how to reach goals and continuously improve the collection of reward signals in a trial-and-error fashion (Fig. 3). At a given moment, each taken *action* a triggers a change in the *state* of the environment $s \to s'$, accompanied by environmental feedback signals as *reward* $r = r(s, a, s')$ obtained by the agent. If the collected reward outcome yields a negative value it can be more naturally interpreted as *punishment*. The environment can be partly controlled by the action of the agent and the reward can be thought of as satisfaction —or aversion —that accompany the execution of a particular action.

The environment is assumed to be *stochastic*, that is, changing in random ways. In addition, the environment is only *partially observable* in the sense that only limited aspects of the environment's state are accessible to the agent's sensory perception (Starkweather et al., 2017). We assume that volatility of the environment is realistic in a computational model which sets out to explain DMN functions of the human brain. We argue that an abstract description of DMN activity based on RL can naturally embed human behavior in the tension between exploitative action with immediate gains and explorative action with longer-term reward outcomes (Dayan and Daw, 2008). In short, DMN implication in a diversity of particularly sophisticated human behaviors can be parsimoniously explained as instantiating probabilistic simulations of experience coupled with prediction error minimization to calibrate action trajectories for reward outcome maximization at different time scales. Such a purposeful optimization objective may be subserved by a stochastic approximation based on a brain implementation for Monte Carlo sampling of events and outcomes.

4.1 Markov Decision Processes

In artificial intelligence and machine learning, a popular computational model for multi-step decision processes are MDPs (Sutton and Barto, 1998). An MDP operationalizes a sequential decision process in which it is assumed that environment dynamics are determined by a Markov process, but the agent cannot directly observe the underlying state. Instead, the agent tries to optimize a *subjective* reward signal (i.e., likely to be different for another agent in the same state and possibly driven by neural processing in the vmPFC) by maintaining probability distributions over actions (possibly represented in the dmPFC) according to their expected utility. This is a minimal set of assumptions that can be made about an environment faced by an agent engaged in interactive learning.

Definition. *Mathematically, an MDP involves a quadruple* $(\mathcal{S}, \mathcal{A}, r, p)$ *where*

- $\mathcal{S}$ *is the set of states, such as* $\mathcal{S} = \{happy, sad, puzzeled\}$.
- $\mathcal{A}$ *is the set of actions, such as* $\mathcal{A} = \{read, run, laugh, sympathize, empathize\}$.
- $r : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ *is the reward function, so that* $r(s, a, s')$ *is the instant reward for taking action* a *in state* s *followed by a state-transition* $s \to s'$.

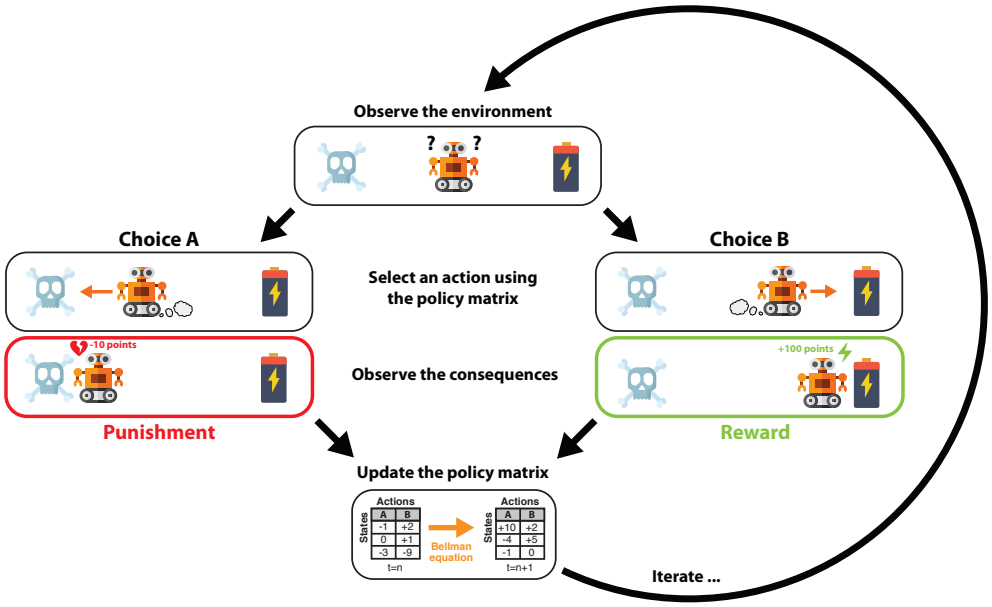


Fig 3. Illustration of a Partially Observable Markov Decision Process (POMDP). Given the current state of the environment, the agent takes an action by following the policy matrix, which is iteratively updated by the Bellman equation. The agent receives a triggered reward and observes the next state. The process goes on until interrupted or a goal state is reached.

- $p : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$, $(s, a, s') \mapsto p(s'|s, a)$, *the probability of moving to state* s' *if action* a *is taken from state* s. *In addition, one requires that such transitions be Markovian. Consequently, the future states are independent of past states and only depend on the present state and action taken.*

The process has *memory* if the subsequent state depends not only on the current state but also on a number of past states. Rational probabilistic planning can thus be reformulated as a standard memoryless Markov process by simply expanding the definition of the state s to include experience episodes of the past. This extension adds the capacity for memory to the model because the next state then depends not only on the current situation but also on previously experienced events, which is the motivation behind Partially Observable MDPs (POMDPs) (Starkweather et al., 2017; O'Reilly and Frank, 2006). Nevertheless, this mathematical property of POMDPs mostly accounts for implicit memory. Since the current paper is concerned with plausibility at the behavioral and neurobiological level, we will address below how our account can accommodate the neurophysiological constraints of the DMN and the explicit memory characteristics of human agents.

Why Markov Decision Processes? One may wonder whether MDP models are applicable to something as complex as human behavior. This class of reinforcement-learning models has had numerous successes in diverse applied domains. For instance, MDPs have been successfully used in financial trading is largely a manifestation of strategic decision-making of interacting human agents. According to how the market responds, the agent incurs gain or loss as environmental feedback of the executed financial actions. Recent research on automatizing market exchanges by algorithmic trading has effectively deployed MDPs as a framework for modeling these elaborate behavioral dynamics (Brázdil et al., 2017; Yang et al., 2015, 2014, 2012; Dempster and Leemans, 2006; Hult and Kiessling, 2010; Abergel et al., 2017). MDPs have also been effective as a behavioral model in robotics (Ng et al., 2004; Abbeel and Ng, 2004) and in challenging multistep strategy games (Mnih et al., 2015; Silver et al., 2016; Pritzel et al., 2017). More recent work has developped an MDP-related way of reasoning about

future behavior of other agents (Rabinowitz et al., 2018). The idea is to use meta-learning (i.e., learning to learn) to build strong priors about the behavior of a population of other agents.

Reinforcement learning in the brain? RL has been argued to be a biologically plausible mechanism in the human brain (O'Doherty et al., 2015; Daw and Dayan, 2014). Indeed, previous authors have proposed (Gershman et al., 2015) that a core property of human intelligence is the improvement of expected utility outcomes as a strategy for action choice in uncertain environments, a view captured by the formalism of MDPs. It has also long been proposed (Dayan and Daw, 2008) that there can be a mapping between algorithmic aspects underlying model-free and model-based RL and neurobiological aspects underlying decision-making, which involves parts of the DMN. The neurotransmitter dopamine could serve as a "teaching signal" to guide estimation of value associations and action policies by modulating synaptic plasticity in the reward-processing circuitry, including the NAc. In contrast, model-based RL would start off with some mechanistic assumptions about the dynamics of the world. These assumptions could relate to the physical laws governing the agent's environment, constraints on the state space, transition probabilities between states, or reward contingencies. An agent might represent such knowledge about the world as follows:

- $r(s, \text{"stand still"}) = 0$ if s does not correspond to a location offering relevant resources.
- $p(s'|s, \text{"stand still"}) = 1$ if $s' = s$ and 0 otherwise.
- etc.

Such knowledge can be partly extracted from the environment: the agent infers a model of the world while learning to take optimal decisions based on the current representation of the environment. These methods learn what the effect is going to be of taking a particular action in a particular state. The result is an estimate of the underlying MDP which can then be either solved exactly or approximately, depending on the setting and what is feasible.

Accumulated rewards and policies The behavior of the agent is governed by a *policy*, which maps states of the world to probability distributions over candidate actions (potentially represented in the dmPFC). Starting at time $t = 0$, following a policy π generates a trajectory of action choices:

$$\begin{aligned}
\textbf{choose action:}\quad & a_0 \sim \pi(a|s_0) \\
\textbf{observe transition:}\quad & s_1 \sim p(s|s_0, a_0) \textbf{ and collect reward } R_0 = r(s_0, a_0, s_1) \\
\textbf{choose action:}\quad & a_1 \sim \pi(a|s_1) \\
\textbf{observe transition:}\quad & s_2 \sim p(s|s_1, a_1), \textbf{ and collect reward } R_1 = r(s_1, a_1, s_2) \\
& \vdots \\
\textbf{choose action:}\quad & a_t \sim \pi(a|s_t) \\
\textbf{observe transition:}\quad & s_{t+1} \sim p(s|s_t, a_t), \textbf{ and collect reward } R_t = r(s_t, a_t, s_{t+1}) \\
& \vdots
\end{aligned}$$

We assume time invariance in that we expect the dynamics of the process to be equivalent over sufficiently long time windows of equal length (i.e., stationarity). Since an action executed in the present moment might have repercussions in the far future. It turns out that the quantity to optimize is not the instantaneous rewards $r(s, a)$, but a *cumulative reward* estimate which takes into account expected reward from action choices in the future. A common approach to modeling this gathered outcome, which is likely to involve extended parts of the DMN, is the time-discounted cumulative reward

$$G^\pi = \sum_{t=0}^{\infty} \gamma^t R_t = R_0 + \gamma R_1 + \gamma^2 R_2 + \ldots + \gamma^t R_t + \ldots \tag{1}$$

This random variable measures the cumulative reward of following an action policy π. The reward outcome is random because it depends both on the environment's dynamics and the

policy π being executed. The exponential delay discounting function used here refers to the usual formulation in the field of reinforcement learning, although psychological experiments may also reveal other discounting regimes (Green and Myerson, 2004). Note that value buffering may be realized in the vmPFC by virtue of this region's connections to the NAc of the reward system (Carlezon and Thomas, 2009; Haber et al., 1995b; Croxson et al., 2005).

The goal of the RL agent is then to successively update this action policy (perhaps most closely related to the PMC) in order to maximize G^π on average (cf. below). In (1), the definition of cumulative reward G^π, the constant γ ($0 \leq \gamma < 1$) is the *reward discount factor*, viewed to be characteristic trait for a certain agent. On the one hand, setting $\gamma = 0$ yields perfectly hedonistic behavior. An agent with such a shortsighted time horizon is exclusively concerned with immediate rewards. This is however not compatible with coordinated planning of longer-term agendas that is potentially subserved by neural activity in the DMN. On the other hand, setting $0 < \gamma < 1$ allows a learning process to arise. A positive γ can be seen as calibrating the risk-seeking trait of the intelligent agent, that is, the behavioral predispositions related to trading longer delays for higher reward outcomes. Such an agent puts relatively more emphasis on rewards expected in a more distant future. Concretely, rewards that are not expected to occur only within a very large number of time steps from the present point are ignored. The complexity reduction by time discounting alleviates the variance of expected rewards accumulated across considered action cascades by limiting the depth of the search tree. Given that there is more uncertainty in the far future, it is important to appreciate that a stochastic policy estimation is more advantageous in many RL settings.

4.2 The components of reinforcement learning in the DMN

Given only the limited information available from an MDP, at a state s the average utility of choosing an action a under a policy π can be captured by the single quantity

$$Q^\pi(s,a) = \mathbb{E}[G^\pi | s_0 = s, a_0 = a], \tag{2}$$

called the Q-value for the state-action pair (s,a). In other words, $Q^\pi(s,a)$ corresponds to the expected reward over all considered action trajectories, in which the agent sets out in the environment in state s, chooses action a, and then follows the policy π to select future actions. For the brain, $Q^\pi(s,a)$ defined in (2) provides the subjective utility of executing a specific action. In this way, we can answer the question "What is the expected utility of choosing action a, and its ramifications, in this situation?". $Q^\pi(s,a)$ offers a formalization of optimal behavior that may well capture processing aspects such as subserved by the DMN in human agents.

4.2.1 Optimal behavior and the Bellman equation

Optimal behavior of the agent corresponds to a strategy π^* for choosing actions such that, for every state, the chosen action guarantees the best possible reward on average. Formally,

$$\pi^*(s) := \operatorname{argmax}_{a \in \mathcal{A}} Q^*(s,a), \text{ where } Q^*(s,a) := \max_\pi Q^\pi(s,a). \tag{3}$$

The learning goal is to approach the ideal policy π^* as close as possible, that is, to solve the MDP. Note that (3) presents merely a definition and does not lend itself as a candidate schema for fully computing MDPs with even moderately sized action and state spaces (i.e., computational intractability). Fortunately, the *Bellman equation* (Sutton and Barto, 1998) provides a fixed-point relation which defines Q^* implicitly via a sampling procedure, without querying the entire space of policies, with the form

$$Q^* = \mathrm{Bel}(Q^*), \tag{4}$$

where the so-called Bellman transform $\mathrm{Bel}(Q)$ of an arbitrary Q-value function $Q : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is another Q-value function defined by

$$\begin{aligned} \mathrm{Bel}(Q)(s,a) &:= \mathbb{E}_{s' \sim p(s'|s,a)}[r(s,a) + \gamma \max_{a' \in \mathcal{A}} Q(s',a')] \\ &= r(s,a) + \gamma \mathbb{E}_{s' \sim p(s'|s,a)}[\max_{a' \in \mathcal{A}} Q(s',a')] \\ &= \text{instantaneous reward} + \text{expected reward for acting greedily thereafter} \end{aligned} \tag{5}$$

The Bellman equation (4) is a temporal consistency equation which provides a dynamic decomposition of optimal behavior by dividing the Q-value function into the immediate reward component and the discounted reward component of the upcoming states. The optimal Q-value operator Q^* is a fixed point for this equation. As a consequence of this outcome stratification, the complicated dynamic programming problem (3) is broken down into simpler sub-problems at different time points. Indeed, exploitation of hierarchical structure in action considerations has previously been related to the medial prefrontal part of the DMN (Koechlin et al., 1999; Braver and Bongiolatti, 2002). Using the Bellman equation, each state can be associated with a certain value to guide action towards a preferred state, thus improving on the current action policy of the agent. Note that in (4) the random sampling is performed only over quantities which depend on the environment. This aspect of the learning process can unroll off-policy by observing state transitions triggered by another (possibly stochastic) behavioral policy.

4.2.2 Value approximation and the policy matrix

As already mentioned in the previous section, Q-learning (Watkins and Dayan, 1992) optimizes over the class of deterministic policies of the form (3). State spaces may be extremely large and tracking all possible states and actions may require prohibitively excessive computation and memory resources, perhaps reflect in the especially high metabolic turn-over of the posterior medial DMN (i.e., PMC). The need of maintaining an explicit table of states can be eliminated by instead using of an approximate Q-value function $\tilde{Q}(s, a|\theta)$ by keeping track of an approximating parameter θ of much lower dimension than the number of states. At a given time step, the world is in a state $s \in \mathcal{S}$, and the agent takes an action which it expects to be the most valuable on average, namely

$$\pi^{\text{hard-max}}(s) = \text{argmax}_{a \in \mathcal{A}} \tilde{Q}(s, a|\theta). \tag{6}$$

This defines a mapping from states directly to actions. For instance, a simple linear model with a kernel ϕ would be of the form $\tilde{Q}(s, a|\theta) = \phi(s, a)^T\theta$, where $\phi(s, a)$ would represent a high-level representation of the state-action pairs (s, a), as was previously proposed (Song et al., 2016), or artificial neural-network models as demonstrated in seminal machine-learning models (Mnih et al., 2015; Silver et al., 2016) for playing complex games (atari, Go, etc.) at super-human levels. In the DMN, the dmPFC is conceivable to implement such a hard-max lookup over the action space. The model parameters θ would correspond to synaptic weights and connection strengths within and between brain regions. It is a time-varying neuronal program which dictates how to move from world states s to actions a via the hard-max policy (6). The approximating Q-value function $\tilde{Q}(s, a|\theta)$ would inform the DMN with the (expected) usefulness of choosing an action a in state s. The DMN, and in particular its dmPFC part, could then contribute to the choice, at a given state s, of an action a which maximizes the approximate Q-values. This mapping from states to actions that is conventionally called *policy matrix* (Mnih et al., 2015; Silver et al., 2016). Learning consists in starting from a given table and updating it during action choices, potentially reflected in neural processing in the PMC, which take the agent to different table entries.

4.2.3 Self-training and the loss function

Successful learning in brains and computer algorithms may not be possible without a defined optimization goal —the *loss function*. The action a chosen in state s according to the policy matrix defined in (6) yields a reward r collected by the agent, after which the environment transitions to a new state $s' \in \mathcal{S}$. One such cycle yields a new *experience* $e = (s, a, r, s')$. Each cycle represents a behavior unit of the agent and is recorded in replay memory buffer —which we hypothesize to involve especially the HC —, possibly discarding the oldest entries to make space: $\mathcal{D} \leftarrow \text{append}(\mathcal{D}, e)$. At time step k, the agent seeks an update $\theta_k \leftarrow \theta_{k-1} + \delta\theta_k$ of the parameters for its approximate model of the Q-value function. Step-by-step model parameter updates warrant a learning process and definition of a loss function. The Bellman equation (4) provides a way to obtain such a loss function (9) as we outline in the following. Experience replay consists in sampling batches of experiences e $(s, a, r, s') \sim \mathcal{D}$ from the replay memory $\mathcal{D}$. The agent then tries to approximate the would-be Q-value for the state-action pair (s, a) as

predicted by the Bellman equation (4), namely

$$y_k := y_k(s, a, s') = r + \gamma \max_{a'} \tilde{Q}(s', a'|\theta_{k-1}), \tag{7}$$

with the estimation of a parametrized regression model $(s, a) \mapsto \tilde{Q}(s, a|\theta_{k-1})$. From a neurobiological perspective, experience replay can be manifested as the re-occurrence of neuron spiking sequences that have also been measured during specific prior actions or environmental states. The HC is a strong candidate for contributing to such neural reinstantiation of behavioral episodes as neuroscience experiments have repeatedly indicated in rats, mice, cats, rabbits, songbirds, and monkeys (Buhry et al., 2011; Nokia et al., 2010; Dave and Margoliash, 2000; Skaggs et al., 2007). Importantly, neural encoding of abstract representations of space and meaning may extent to several parts of the DMN (Constantinescu et al., 2016a) (see Fig. 4).

At the current step k, computing an optimal parameter update then corresponds to finding the model parameters θ_k which minimize the following mean-squared optimization loss

$$\mathcal{L}(\theta_k^Q) = \mathbb{E}_{(s,a,r,s')\sim\mathcal{D}}\left[\frac{1}{2}(\tilde{Q}(s, a|\theta_k) - y_k)^2\right], \tag{8}$$

where y_k is obtained from (4). A recently proposed, practically successful alternative approach is to estimate the representation using an artificial deep neural-network model. This approach leads to the so-called *deep Q-learning* (Mnih et al., 2015; Silver et al., 2016) - a family of methods which is the current state-of-the-art in RL research. The set of model parameters θ that instantiate the non-linear interactions between layers of the artificial neural network may find a neurobiological correspondence in the adaptive strengths of axonal connections between neurons from the different levels of the neural processing hierarchy (Mesulam, 1998; Taylor et al., 2015).

A note on bias in self-training. Some bias may be introduced by self-training due to information shortage caused by the absence of external stimulation. One way to address this issue is using importance sampling to replay especially those state-transitions from which there is more to learn for the agent (Schaul et al., 2015; Hessel et al., 2017). New transitions are inserted into the replay buffer with maximum priority, thus shifting emphasis to more recent transitions. Such insertion strategy would help counterbalance the bias introduced by the information shortage incurred by absent external input. Other authors noticed (Hessel et al., 2017) that such prioritized replay reduces the data complexity and the agent shows faster increases in learning performance.

4.2.4 Optimal control via stochastic gradient descent

Efficient learning of the entire set of model parameters can effectively be achieved via stochastic *gradient descent*, a universal algorithm for finding local minima based on the first derivative of the optimization objective. Stochastic here means that the gradient is estimated from batches of training samples, which here corresponds to blocks of experience from the replay memory:

$$\delta = -\alpha_k \nabla_{\theta_k} \mathcal{L}(\theta_k) = -\alpha_k \mathbb{E}_{(s,a,r,s')\sim\mathcal{D}}[\underbrace{(\tilde{Q}(s, a|\theta_k) - y_k)}_{\text{prediction error}} \underbrace{\nabla_{\theta_k} \tilde{Q}(s, a|\theta_k)}_{\text{aversion}}], \tag{9}$$

where the positive constants $\alpha_1, \alpha_2, \ldots$ are learning rates. Thus, the subsequent action is taken to drive reward prediction errors to percolate from lower to higher processing layers to modulate the choice of future actions. It is known that under special conditions on the learning rates α_k –namely that the learning rates are neither too large nor too small, or more precisely that the sum $\sum_{k=0}^{\infty} \alpha_k$ diverges while $\sum_{k=0}^{\infty} \alpha_k^2$– the thus generated approximating sequence of Q-value functions

$$\tilde{Q}(.,.|\theta_0) \to \tilde{Q}(.,.|\theta_1) \to \tilde{Q}(.,.|\theta_2) \to \ldots$$

are attracted and absorbed by the optimal Q-value function Q^* defined implicitly by the Bellman equation (4).

4.2.5 Does the hippocampus subserve Monte Carlo sampling?

In RL, Monte Carlo simulation is a common means to update the agent's belief state based on stochastic sampling of environmental states and possible transitions (Daw and Dayan, 2014; Silver and Veness, 2010). Monte Carlo simulation provides a simple method for evaluating the value of a state. This inference procedure provides an effective mechanism both for tree search of the considered action trajectories and for belief state updates, breaking the curse of dimensionality and allowing much greater scalability than a RL agent without stochastic resampling procedures. Such methods scale as a function of available data (i.e., sample complexity) that is determined only by the underlying difficulty of the MDP, rather than the size of the state space or observation space, which can be prohibitively large.

In the human brain, the HC could contribute to synthesizing imagined sequences of world states, actions and rewards (Aronov et al., 2017; Chao et al., 2017; Boyer, 2008). These stochastic simulations of experience batches re-assembled from memory would be used to update the value function, without ever looking inside the black box describing the model's dynamics. A brain-imaging experiment in humans for instance identified hippocampal signals that specifically preceded upcoming choice performance in prospective planning in new environments (Kaplan et al., 2017). It would be a simple strategy to evaluate all legal actions and selecting the action with highest expected cumulative rewards. In MDPs, MC simulation provides an effective mechanism both for tree search and for belief-based state updates, breaking the curse of dimensionality and allowing much greater scalability than has previously been possible (Silver et al., 2016). This is because expected consequences of action choices can be well evaluated although only a subset of the states are actually considered (Daw and Dayan, 2014).

A note on implicit and explicit memory. While Markov processes are usually memoryless, it is mathematically feasible to incorporate a set of previous states of such model into the current state. This extension may partly account for implicit memory at the behavioral level, but may not explain the underlying neurobiological implementation or accommodate explicit memory. Implicit memory-based processing arises in our MDP account of DMN function in several different forms: successive updates of a) the action policy and the value function, both being products of the past, as well as b) the deep non-linear relationships within the hierarchical connections of biological neural networks (especially in the association cortex). The brain's adaptive synaptic connections can be viewed as a deep artificial neural-network architecture affording an implicit form of information compression of life experience. Such memory traces are stored in the neural machinery and can be implicitly retrieved as a form of knowledge during simulation of action rather than accessed as a stored explicit representation (Pezzulo, 2011). c) Certain neural processes in the hippocampus can be seen as some type of Monte Carlo sampling for memory recall, which can also be a basis for probabilistic simulations across time scales (Schacter et al., 2007; Axelrod et al., 2017).

4.3 Summary and hypotheses for future studies

The DMN is today known to consistently increase in neural activity when humans engage in cognitive processes that are relatively detached from the current sensory environment. The more familiar and predictable the current environment, the more brain resources may remain for allocating DMN activity to MDP processes extending beyond the present time and sensory context. This speculation receives quantitative support in that connectional links between nodes of the DMN have been reported to be more consistent and reliable than functional couplings within any other macroscopical networks (Shehzad et al., 2009). As such, random-sampling-related baseline evaluation of action possibilities and their consequences may be subserved by the DMN and get partly suspended when novelty in the external environment is encountered or immediate action is required. In line with this perspective, DMN engagement was shown to heighten and relate to effective behavioral responses in the practiced phase of a demanding cognitive flexibility task, as compared to acquisition phase when participants learned context-specific rules. This involvement in automated decision-making has led the authors to propose an "autopilot" role for the DMN (Vatansever et al., 2017), which may contribute to optimizing intervention of the organism on the world in general. Among all parts

of the DMN, the RTPJ is perhaps the most evident candidate for a network-switching relay that calibrates between processing of environment-engaged versus internally generated information (Downar et al., 2000; Golland et al., 2006; Bzdok et al., 2013c).

Additionally, the DMN was proposed to be situated at the top of the brain network hierarchy, with the subordinate salience and dorsal attention network in the middle and the primary sensory cortices at the bottom (Carhart-Harris and Friston, 2010; Margulies et al., 2016b). Its putative involvement in thinking about hypothetical experiences and future outcomes appears to tie in with the implicit computation of action and state cascades as a function of experienced events and collected feedback from the past. A policy matrix encapsulates the choice probabilities of possible actions on the world given a current situation (i.e., state). The DMN may subserve constant exploration of candidate action trajectories and nested estimation of their cumulative reward outcomes. Implicit computation of future choices provides a potential explanation for the evolutionary emergence and practical usefulness of mind-wandering at day-time and dreams during sleep in humans.

Our formal account on the DMN readily motivates several empirical predictions for future neuroscience research. Perhaps one of the first experimental venue concerns the neural correlates of the Bellman equation in the DMN. There are already relationship between the decomposition of consecutive action choices by the Bellman equation and neuroscientific insights: specific neural activity in the dorsal prefrontal cortex (BA9) was for instance linked to processing "goal-tree sequences" in human brain-imaging experiments (Koechlin et al., 1999, 2000). Sub-goal exploration may require multi-task switching between cognitive processes as later parts of a solution frequently depend on respective earlier steps in a given solution path, which necessitates storage of expected intermediate outcomes. As such, "cognitive branching" operations for nested processing of behavioral strategies are likely to entail secondary reallocation of attention and working-memory resources. Further brain-imaging experiments corroborated the prefrontal DMN to subserve "processes related to the management and monitoring of sub-goals while maintaining information in working memory" (Braver and Bongiolatti, 2002) and to functionally couple with the hippocampus conditioned by "deep versus shallow planning" (Kaplan et al., 2017). Moreover, neurological patients with lesions in this DMN region were reported to be impaired in aspects of realizing "multiple sub-goal scheduling" (Burgess et al., 2000). Hence, the various advanced human abilities subserved by the DMN, such as planning and abstract reasoning, can be viewed to involve some form of action-decision branching to enable higher-order executive control.

We therefore hypothesize in humans a functional dissociation between computations pertaining to action policy versus adapting stimulus-value associations as we expect implementation in different subsystems of the DMN. First, we expect that fMRI signals in the right temporo-parietal junction relate to behavioral changes subsequent to adaptation in the action choice tendencies (policy matrix) involved in non-value-related prediction error. Second, fMRI signals in the ventromedial prefrontal cortex should relate to behavioral changes following adaptation in value estimation (value matrix) due to reward-related stimulus-value association. We further expect that fMRI signals in the posteromedial cortex, as a potential global information integrator, are related to shifts in overt behavior based on previous adaptations in both policy or value estimation.

Our process model of the DMN has also implications for experiments in neuroeconomy; especially for temporal discounting and continuous learning paradigms. More specifically, we hypothesize in humans a functional relationship between the DMN closely associated with the occurrence of stimulus-independent thoughts and the reward circuitry. During an iterative neuroeconomic two-player game, fMRI signals in the DMN could be used to predict reward-related signals in the nucleus accumbens across trials in a multi-step learning paradigm. We expect that the more DMN activity is measured to be increased, supposedly the higher the tendency for stimulus-independent thoughts, the more the fMRI signals in the reward circuits should be independent of the reward context in the current sensory environment. In the case of temporal discounting, we hypothesize in humans that the relevant time horizon is modulated by various factors such as age, acute stress, and time-enduring impulsivity traits (Luksys et al., 2009; Haushofer and Fehr, 2014). Using such a delayed-reward experiment, it can be quantified how the time horizon is affected at the behavioral level and then traced back to its corresponding neural representation. Such experimental investigation can be designed to

examine between-group and within-group effects (e.g., impulsive population like chronic gamblers or drug addicts); and brought in context with the participants age, education, IQ, and personality traits.

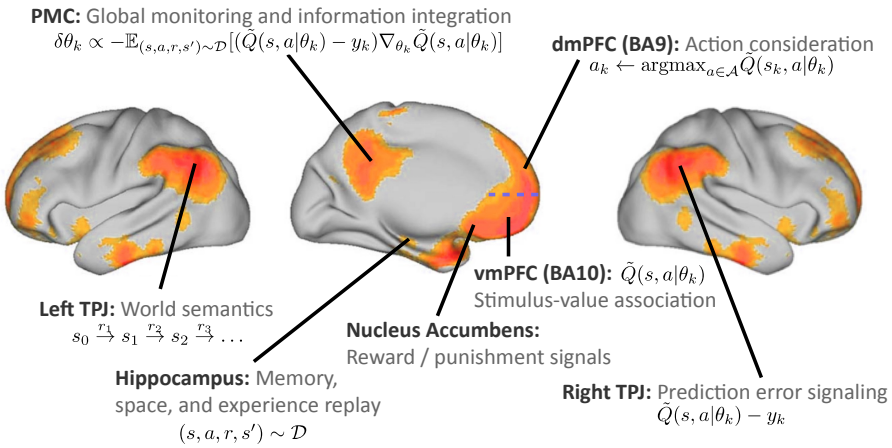


Fig 4. Default mode network: possible neurobiological implementation of reinforcement learning. Overview of how the constituent regions of the DMN (refer to section 3; blue horizontal dashed line indicates the border between BA9 and BA10) may map onto computational components necessary for a RL agent. Axonal tracing in monkeys and diffusion tractography in humans suggested that the NAc of the reward circuitry has monosynaptic fiber connections to the vmPFC (Haber et al., 1995b; Croxson et al., 2005). Evaluation of propagated value information and triggered affective states encoded in the vmPFC may then feed into the functionally connected partner nodes of the DMN, such as the dmPFC and PMC (Andrews-Hanna et al., 2010; Bzdok et al., 2013b).

As another experimental prediction derived from our MDP approach to the DMN, the HC may contribute to generating perturbed action-transition-state-reward samples as batches of pseudo-experience (i.e., recalled, hypothesized, and forecasted scenarios). The small variations in these experience samplings allow searching through a larger space of model parameters and candidate experiences. Taken to its extreme, stochastic recombination of experience building blocks can further optimize the behavior of the RL agent by learning from scenarios in the environment that the agent might encounter only very rarely or never. An explanation is thus offered for experiencing seemingly familiar situations that a human has however never actually encountered (i.e., déjà vu effect). While such a situation may not have been experienced in the physical world, the DMN may have previously stochastically generated, evaluated, and adapted to such a randomly synthesized event. Generated representations arguably are "internally manipulable, and can be used for attempting actions internally, before or instead of acting in the external reality, and in diverse goal and sensory contexts, i.e. even outside the context in which they were learned" (Pezzulo, 2011). In the context of scarce environmental input and feedback (e.g., mind-wandering or sleep), mental scene construction allows pseudo-experiencing possible future scenarios and action outcomes.

A possible interplay between memory retrieval and "mind-searching" moreover suggests that experience replay for browsing problem solutions subserved by the DMN contributes to choice behavior in mice. Hippocampal single-cell recordings have shown that neural patterns during experimental choice behavior are reiterated during sleep and before making analogous choices in the future. We hypothesize that, in addition to the hippocampus, there is a necessity of cortical DMN regions for "mind-searching" candidate actions during choice behavior. It can be experimentally corroborated by causal disruption of DMN regions, such as by circumscribed brain lesion or optogenetic intervention in the inferior parietal and prefrontal cortices. From the perspective of a RL agent, prediction in the DMN reduces to generalization of policy and value computations from sampled experiences to successful action choices and reward

predictions in future states. As such, plasticity in the DMN arises naturally. If an agent behaving optimally in a certain environment moves to new, yet unexperienced environment, reward prediction errors will largely increase. This feedback will lead to adaptation of policy considerations and value estimations until the intelligent system converges to a new steady state of optimal action decisions in a volatile world.

A last experimental prediction for future studies concerns how synaptic epigenesis may shape the policy matrix. Indeed, we did not adress here the additional layer of learning which concerns the addition of new entries in the state and action spaces. Extension of the action repertoire could be biologically realized by synaptic epigenesis (Gisiger et al., 2005). The tuning of synaptic weights through learning can stabilize additional patterns of activity by creating new attractors in the neural dynamics landscape (Takeuchi et al., 2014). Those attractors can then constrain both the number of factors taken into account by decision processes and the possible behaviors of the agent (Wang, 2008). To examine this potential higher-level mechanism, we propose to probe how synaptic epigenesis is related to neural correlates underlying policy matrix updates: in humans the changes of functional connectivity between DMN regions can be investigated following a temporal discounting experiment and in monkeys or rodents anterograde tracing can be used to study how homolog regions of the DMN present increased synaptic changes compare to other parts of the brain.

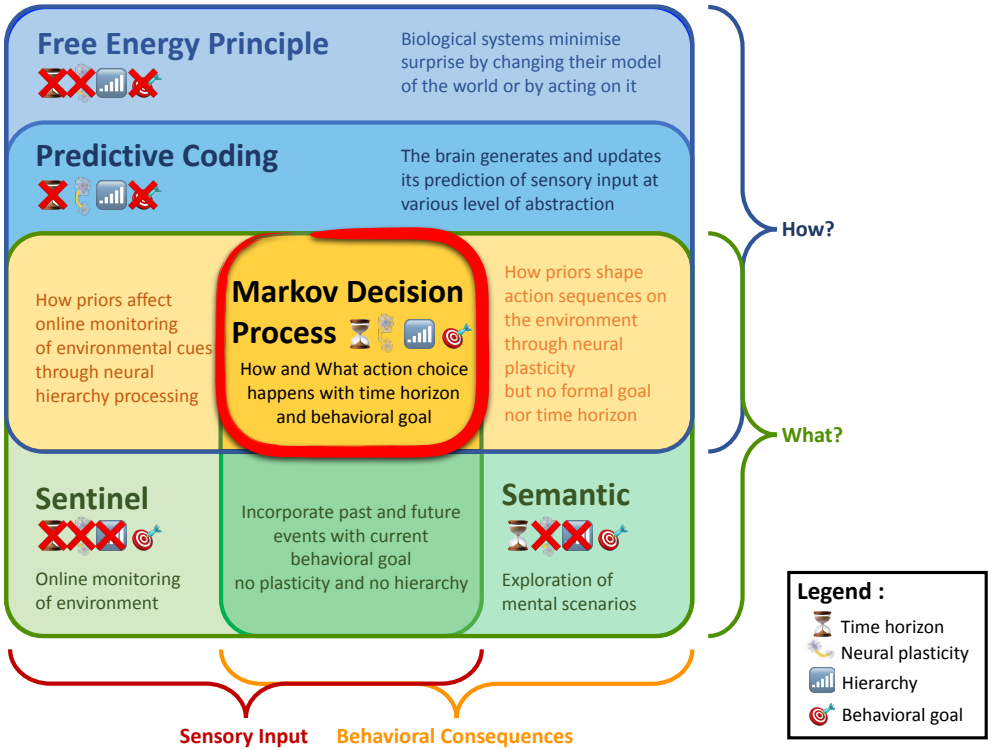


Fig 5. Situating Markov Decision Processes among other accounts of default mode function. The Venn diagram summarizes the relationship between four previously proposed explanations for the functional role of the DMN and our present account. Viewing empirical findings in the DMN from the MDP viewpoint incorporates important aspects of the free energy principle, predictive coding, sentinal hypothesis, and semantic hypothesis. The MDP account may reconcile several strengths of these functional accounts in a process model that simultaneously acknowledges environmental input and behavioral choices as well as the computational and algorithmic properties (How? and What?) underlying higher-order control of the organism.

5 Relation to existing accounts

5.1 Predictive coding

Predictive coding mechanisms (Clark, 2013; Friston, 2008) are a frequently evoked idea in the context of default mode function (Bar et al., 2007). Cortical responses are explained as emerging from continuous functional interaction between higher and lower levels of the neural processing hierarchy. Feed-forward sensory processing is constantly calibrated by top-down modulation from more multi-sensory and associative brain regions further away from primary sensory cortical regions. The dynamic interplay between cortical processing levels may enable learning about aspects of the world by reconciling gaps between fresh sensory input and predictions computed based on stored prior information. At each stage of neural processing, an internally generated expectation of aspects of environmental sensations is directly compared against the actual environmental input. A prediction error at one of the processing levels induces plasticity changes of neuronal projections to allow for gradually improved future prediction of the environment. In this way, the predictive coding hypothesis offers explanations for the constructive, non-deterministic nature of sensory perception (Friston, 2010; Buzsáki, 2006) and the intimate relation of motor movement to sensory expectations (Wolpert et al., 1995; Körding and Wolpert, 2004). Contextual integration of sensorimotor perception-action cycles may be maintained by top-down modulation using internally generated information about the environment.

In short, predictive coding processes conceptualize updates of the internal representation of the environment to best accommodate and prepare the organism for processing the constant influx of sensory stimuli and performing action on the environment (Fig. 5). There are hence a number of common properties between the predictive coding account and the proposed formal account of DMN function based on MDPs. Importantly, a generative model of how perceived sensory cues arise in the world would be incorporated into the current neuronal wiring. Further, both functional accounts are supported by neuroscientific evidence that suggest the human brain to be a "statistical organ" (Friston et al., 2014) with the biological purpose to generalize from the past to new experiences. Neuroanatomically, axonal back projections indeed outnumber by far the axonal connections mediating feedforward input processing in the monkey brain and probably also in humans (Salin and Bullier, 1995). These many and diverse top-down modulations from higher onto downstream cortical areas can inject prior knowledge at every stage of processing environmental information. Moreover, both accounts provide a parsimonious explanation for why the human brain's processing load devoted to incoming information decreases when the environment becomes predictable. This is because the internal generative model only requires updates after discrepancies have occurred between environmental reality and its internally reinstantiated representation. Increased computation resources are however allocated when unknown stimuli or unexpected events are encountered by the organism. The predictive coding and MDP account hence naturally evoke a mechanism of brain plasticity in that neuronal wiring gets increasingly adapted when faced by unanticipated environmental challenges.

While sensory experience is a constructive process from both views, the predictive coding account frames sensory perception of the external world as a generative experience due to the modulatory top-down influence at various stages of sensory input processing. This generative top-down design is replaced in our MDP view of the DMN by a sequential decision-making framework. Further, the hierarchical processing aspect from predictive coding is re-expressed in our account in the form of nested prediction of probable upcoming actions, states, and outcomes. While both accounts capture the consequences of action, the predictive coding account is typically explained without explicit parameterization of the agent's time horizon and has a tendency to be presented as emphasizing prediction about the immediate future. In the present account, the horizon of that look into the future is made explicit in the γ parameter of the Bellman equation. Finally, the process of adapting the neuronal connections for improved top-down modulation takes the concrete form of stochastic gradient computation and back-propagation in our MDP implementation. It is however important to note that the neurobiological plausibility of the back-propagation procedure is controversial (Goodfellow et al., 2016).

In sum, recasting DMN function in terms of MDPs therefore naturally incorporates a majority of aspects from the prediction coding hypothesis. The present MDP account of DMN function may therefore serve as a concrete implementation of many predictive coding ideas. MDPs have the advantage of exposing an explicit mechanisms for modulating the horizon of future considerations and for how the internal representation of the world is updated, as well as why certain predictions may be more relevant to the agent than others.

5.2 The semantic account

This frequently proposed cognitive account to explain DMN function revolves around forming logical associations and abstract analogies between experiences and conceptual knowledge derived from past behavior (Bar, 2007; Binder et al., 1999; Constantinescu et al., 2016b). Analogies might naturally tie incoming new sensory stimuli to explicit world knowledge (i.e., semantics, Fig. 5) (Bar, 2009). The encoding of complex environmental features could thus be facilitated by association to known similar states. Going beyond isolated meaning and concepts extracted from the world, semantic building blocks may need to get recombined to enable mental imagery to (fore)see never-experienced scenarios. As such, semantic knowledge would be an important ingredient for optimizing behavior by constantly simulating possible future scenarios (Boyer, 2008; Binder and Desai, 2011). Such cognitive processes can afford the internal construction and elaboration of necessary information that is not presented in the immediate sensory environment by recombining building blocks of concept knowledge and episodic memories (Hassabis and Maguire, 2009). Indeed, in aging humans, remembering the past and imagining the future equally decreased in the level of detail and were associated with concurrent deficits in forming and integrating relationships between items (Addis et al., 2008; Spreng and Levine, 2006).

Further, episodic memory, language, problem solving, planning, estimating others' thoughts, and spatial navigation represent neural processes that are likely to build on abstract world knowledge and logical associations for integrating the constituent elements in rich and coherent mental scenes (Schacter et al., 2007). "[Foresight] and simulations are not only automatically elicited by external events but can be endogenously generated when needed. [...] The mechanism of access via simulation could be a widespread method for accessing and producing knowledge, and represents a valid alternative to the traditional idea of storage and retrieval" (Pezzulo, 2011). Such mental scene-construction processes could contribute to interpreting the present and foreseeing the future. Further, mental scene imagery has been proposed to imply a distinction between engagement in the sensory environment and internally generated mind-wandering (Buckner and Carroll, 2007). These investigators stated that "A computational model [...] will probably require a form of regulation by which perception of the current world is suppressed while simulation of possible alternatives are constructed, followed by a return to perception of the present".

In comparison, both the semantic hypothesis and the present formal account based on MDPs expose mechanisms of how action considerations could be explored. In both accounts, there is also little reason to assume that contemplating alternative realities of various levels of complexity, abstraction, time scale, and purpose rely on mechanisms that are qualitatively different. This interpretation concurs with DMN activity increases across time, space, and content domains demonstrated in many brain-imaging studies (Spreng et al., 2009; Laird et al., 2009; Bzdok et al., 2012; Binder et al., 2009). Further, the semantic hypothesis and MDP account offer explanations why HC damage does not only impair recalling past events, but also imagining hypothetical and future scenarios (Hassabis et al., 2007). While both semantic hypothesis and our formal account propose memory-enabled, internally generated information for probabilistic representation of action outcomes, MDPs render explicit the grounds on which an action is eventually chosen, namely, the estimated cumulative reward. In contrast to many versions of the semantic hypothesis, the MDPs naturally integrate the egocentric view (more related to current action, state, and reward) and the world view (more related to past and future actions, states, and rewards) on the world in a same optimization problem. Finally, the semantic account of DMN function does not provide suffcient explanation of *how* explicit world knowledge and logical analogies thereof lead to foresight of future actions and states. The semantic hypothesis does also not fully explain why memory recall for scene construction in

humans is typically fragmentary and noisy instead of accurate and reliable. In contrast to existing accounts on semantics and mental scene construction, the random and creative aspects of DMN function are explained in MDPs by the advantages of stochastic optimization. Our MDP account provides an algorithmic explanation in that stochasticity of the parameter space exploration by Monte Carlo approximation achieves better fine-tuning of the action policies and inference of expected reward outcomes. That is, the purposeful stochasticity of policy and value updates in MDPs provides a candidate explanation for why humans may have evolved imperfect noisy memories as the more advantageous adaptation. In sum, mental scene construction according to the semantic account is lacking an explicit time and incentive structure, both of which are integral parts of the MDP interpretation of DMN function.

5.3 The sentinel account

Regions of the DMN have been proposed to process the experienced or expected relevance of environment cues (Montague et al., 2006). Processing self-relevant information was perhaps the first functional account that was proposed for the DMN (Gusnard et al., 2001; Raichle et al., 2001). Since then, many investigators have speculated that neural activity in the DMN may reflect the brain's continuous tracking of relevance in the environment, such as spotting predators, as an advantageous evolutionary adaptation (Buckner et al., 2008; Hahn et al., 2007). According to this cognitive account, the human brain's baseline maintains a "radar" function to detect subjectively relevant cues and unexpected events in the environment (Fig. 5). Propositions of a sentinel function to underlie DMN activity have however seldom detailed the mechanisms of how attention and memory resources are exactly reallocated when encountering a self-relevant environmental stimulus. Instead, in the present MDP account, promising action trajectories are recursively explored by the human DMN. Conversely, certain branches of candidate action trajectories are detected to be less worthy to get explored. This mechanism, expressed by the Bellman equation, directly implies stratified allocation of attention and working memory load over relevant cues and events in the environment.

Further, our account provides a parsimonious explanation for the consistently observed DMN implication in certain goal-directed experimental tasks and in task-unconstrained mind-wandering (Smith et al., 2009; Bzdok et al., 2016b). Both environment-detached and environment-engaged cognitive processes may entail DMN recruitment if real or imagined experience is processed, manipulated, and used in service of organism control. During active engagement in tasks, the policy and value estimates may be updated to optimize especially short-term action. At passive rest, these parameter updates may improve especially mid- and long-term action. This horizon of the agent is expressed in the γ parameter in the MDP account. We thus provide answers for the currently unsettled question why the involvement of the same neurobiological brain circuit (i.e., DMN) has been documented for specific task performances and baseline 'house-keeping' functions.

In particular, environmental cues that are especially important for humans are frequently of social nature. This may not be surprising given that the complexity of the social systems is likely to be a human-defining property (Tomasello, 2009; Dunbar and Shultz, 2007). According to the "social brain hypothesis", the human brain has especially been shaped for forming and maintaining increasingly complex social systems, which allows solving ecological problems by means of social relationships (Whiten and Byrne, 1988). In fact, social topics probably amount to roughly two thirds of human everyday communication (Dunbar et al., 1997).
Mind-wandering at daytime and dreams during sleep are also rich in stories about people and the complex interactions between them. In line with this, DMN activity was advocated to be specialized in continuous processing of social information as a physiological baseline of human brain function (Schilbach et al., 2008). This view was later challenged by observing analogues of the DMN in monkeys (Mantini et al., 2011), cats (Popa et al., 2009), and rats (Lu et al., 2012), three species with social capacities that can be expected to be less advanced than in humans (Mars et al., 2012).

Moreover, the principal connectivity gradient in the cortex appears to be greatly expanded in humans compared to monkeys, suggesting a phylogenetically conserved axis of cortical expansion with the DMN emerging at the extreme end in humans (Margulies et al., 2016a). Computational models of dyadic whole-brain dynamics demonstrated how the human

connectivity topology, on top of facilitating processing at the intra-individual level, can explain our propensity to coordinate through sensorimotor loops with others at the inter-individual level (Dumas et al., 2012). The DMN is moreover largely overlapping with neural networks associated with higher-level social processes (Schilbach et al., 2012). For instance, the vmPFC, PMC, and RTPJ together may play a key role in bridging the gap between self and other by integrating low-level embodied processes within higher level inference-based mentalizing (Lombardo et al., 2009; Alcal-Lpez et al., 2017).

Rather than functional specificity for processing social information in particular, the present MDP account can parsimoniously incorporate the dominance of social content in human mental activity as high value function estimates for information about humans (Baker et al., 2009; Kampe et al., 2001; Krienen et al., 2010). The DMN may thus modulate reward processing in the human agent in a way that prioritizes appraisal of and action towards social contexts, without excluding relevance of environmental cues of the physical world. In sum, our account on the DMN directly implies its previously proposed "sentinel" function of monitoring the environment for self-relevant information in general and inherently accommodates the importance of social environmental cues as a special case.

5.4 A note on the free-energy principle and active inference

According the *free-energy principle* (FEP) and theories of *active inference* (Friston, 2010; Friston et al., 2009; Dayan et al., 1995), the brain corresponds to a biomechanical reasoning engine. Much of neural computation is dedicated to minimizing the long-term average of surprise: the log-likelihood of the observed sensory input –more precisely, an upper bound thereof– relative to the expectations about the external world derived from internal representations. The brain would continuously generate hypothetical explanations of the world and predict its sensory input $\mathbf{x}$ (analogous to the state-action (s, a) pair in an MDP framework). However, surprise is challenging to optimize numerically because we need to solve the intractable problems of summing over all hidden causes $\mathbf{z}$ of the sensations (an intractable problem). Instead, FEP therefore minimizes an upper-bound on surprise given by

$$\begin{aligned} \text{generative surprise} &:= -\log(p_G(\mathbf{x})) = F_G(\mathbf{x}) \\ &= \underbrace{F_G^R(\mathbf{x})}_{\text{accuracy}} - \underbrace{\mathrm{KL}(p_R(\mathbf{z}|\mathbf{x}) \| p_G(\mathbf{z}|\mathbf{x}))}_{\text{complexity}} \\ &\leq F_G^R(\mathbf{x}), \text{ with equality if } p_R(\mathbf{z}|\mathbf{x}) = p_G(\mathbf{z}|\mathbf{x}) \text{ for all } \mathbf{z}. \end{aligned} \tag{10}$$

where

$$F_G^R(\mathbf{x}) := \langle -\log(p_G(\mathbf{z}, \mathbf{x})) \rangle_{p_R(\mathbf{z}|\mathbf{x})} - \mathcal{H}(p_R(\mathbf{z}|\mathbf{x})) \tag{11}$$

is the *free energy*. Here, the angular brackets denote the *expectation* of the joint negative log-likelihood $-\log(p_G(\mathbf{z}, \mathbf{x}))$ w.r.t the recognition density $p_R(\mathbf{z}|\mathbf{x})$, $\mathcal{H}$ is the *entropy* function defined by $\mathcal{H}(p) := -\sum_{\mathbf{z}} p(\mathbf{z}) \log(p(\mathbf{z})$, while KL(.‖.) is the usual *Kullback-Leibler (KL) divergence* (also known as *relative entropy*) defined by $\mathrm{KL}(p\|q) := \sum_z p(\mathbf{z}) \log(p(\mathbf{z})/q(\mathbf{z})) \geq 0$, which is a measure of difference between two probability distributions. In this framework, the goal of the agent is to iteratively refine the generative model p_G and the recognition model p_R so as to minimize the free energy $F_G^R(\mathbf{x})$ over sensory input $\mathbf{x}$.

Importantly, $F_G^R(x)$ gets low in the following cases:

- $p_R(\mathbf{z}|\mathbf{x})$ puts a lot of mass on configurations $(\mathbf{z}, \mathbf{x})$ which are p_G-likely
- $p_R(\mathbf{z}|\mathbf{x})$ is as uniform as possible (i.e., have high entropy), so as not to concentrate all its mass on a small subset of possible causes for the sensation $\mathbf{x}$

Despite its popularity, criticism against the FEP has been voiced repeatedly, which we allude to in the following. The main algorithm for minimizing free energy $F_G^R(\mathbf{x})$ is the *wake-sleep algorithm* (Dayan et al., 1995). As these authors noted, a crucial drawback of the wake-sleep algorithm (and therefore of theories like the FEP (Friston, 2010)) is that it involves a pair of forward (generation) and backward (recognition) models p_G and p_R that together does not correspond to optimization of a bound of the marginal likelihood because KL divergence is not symmetric in its arguments.

These considerations render the brain less likely to implement a variant of the wake-sleep algorithm. More recently, *variational auto-encoders* (Kingma and Welling, 2013) emerged that may provide an efficient alternative to the wake-sleep algorithm. Such compression-and-reconstruction models overcome a number of the technical limits of the wake-sleep algorithm by using a reparametrization maneuver, which makes it possible to do differential calculus on random sampling procedures without exploding variance. As a result, unlike the wake-sleep algorithm for minimizing free energy, variational auto-encoders can be efficiently trained via back-propagation of prediction errors.

The difference between the FEP and the MDP account may be further clarified by a thought experiment. Since theories based on the FEP (Friston, 2010; Friston et al., 2009) conceptualize ongoing behavior in an organism to be geared towards the surprise-minimizing goal. Hence, an organism entering a dark room would remain trapped in this location because its sensory inputs are perfectly predictable given the environmental state (Friston et al., 2012). However, such a behavior is seldom observed in humans in the real world. In a dark room, the intelligent agents would search for light sources to explore the surroundings or aim to exit the room. One may object that, for the FEP agent, a dark room would paradoxically correspond to a state of particularly high relevance. Driven by the surprise-minimization objective, the FEP agent would eventually not bootstrap itself out of such saddle points to explore more interesting parts of the environment. In contrast, an organism operating under our RL-based theory would inevitably identify the sensory-stimulus-deprived room as a local minimum. Indeed, hippocampal experience replay (see 4.2.3) could serve to sample memories or fantasies of alternative situations with reward structure. Such artificially generated *internal* sensory input, potentially subserved by the DMN, could then entice the organism to explore the room, for instance by looking for and using the light switch or finding the room exit.

We finally note that FEP and active inference can be reframed in terms of our RL framework. This is possible by recasting the Q-value function (i.e., expected long-term reward) maximized by the DMN to correspond to negative surprise, that is, the log-likehood of current sensory priors the agent has about the world. More explicitly, this formulation corresponds to using free-energy as a Q-value approximator for the MDP in the following way:

$$-Q \approx \underbrace{F_G^R(\mathbf{x})}_{\text{negative free energy}} \approx \underbrace{-\log(p_G)}_{\text{FEP generative surprise}} .$$

Such a surprise-guided RL scheme has previously been advocated under the equivalent framework of energy-based RL (Sallans and Hinton, 2004; Elfwing et al., 2016) and information compression (Schmidhuber, 2010; Mohamed and Rezende, 2015). Nevertheless, minimization of surprise quantities alone may be insufficient to explain the diversity of behaviors that humans and other intelligent animals can perform.

6 Conclusion

Which brain function could be important enough for the existence and survival of the human species to justify constantly high energy costs? While previous experiments on the DMN frequently set out to investigate *what* its subserved function may be, we have proposed a way of reasoning *how* this major brain network may do what it is doing. MDPs motivate an attractive formal account of how the human association cortex can be thought to implement multi-sensory representation and high-level decision-making to optimize the organism's behavioral strategies. This idealized process model accommodates a number of previous observations from neuroscience studies on the DMN by simple but non-trivial mechanisms. Viewed as a Markovian sequential decision process, human behavior unfolds by inferring expected reward outcomes from hypothetical action cascades and extrapolation from past experience to upcoming events for guiding behavior in the present. MDPs also provide a formalism how opportunity in the environment can be deconstructed, evaluated, and exploited when an agent is confronted with challenging interdependent decisions. This abstract process interpretation may well be compatible with the DMN's poorly understood involvement across autobiographical memory recall, problem solving, abstract reasoning, social cognition, as well

as delay discounting and self-prospection into the future. For instance, improvement of the internal world representation by injecting stochasticity into the recall of past actions and inference of action outcomes may explain why highly accurate memories have been disfavored in human evolution and why human creativity may be adaptive.

A major hurdle in guessing DMN function from cognitive brain-imaging studies has been its similar neural engagement in different time scales: thinking about the past (e.g., autobiographical memory retrieval), imagining hypothetical presents (e.g., daytime mind-wandering), and anticipating scenarios yet to come (e.g., delay discounting). The MDP account of DMN activity offers a natural integration of a-priori diverging cognitive processes into a common framework. It is an important advantage of the proposed artificial intelligence perspective on DMN biology that it is practically computable and readily motivates neuroscientific hypotheses that can be put to the test in future research. We encourage neuroscience experiments on the DMN to operationalize the set of action, value, and state variables that govern the behavior of intelligent RL agents. At the least, we propose an alternative vocabulary to describe, contextualize, and interpret experimental findings in neuroscience studies on higher-level cognition. Ultimately, neural processes in the DMN may realize a brain-wide information integration ranging from real experience over purposeful dreams to predicted futures to continuously refine the organism's intervention on the world.

References

P. Abbeel and A. Y. Ng. Apprenticeship learning via inverse reinforcement learning. In Proceedings of the Twenty-first International Conference on Machine Learning, ICML '04, pages 1–, New York, NY, USA, 2004. ACM.

F. Abergel, C. Huré, and H. Pham. Algorithmic trading in a microstructural limit order book model. Preprint, May 2017.

D. R. Addis, A. T. Wong, and D. L. Schacter. Age-related changes in the episodic simulation of future events. Psychological science, 19(1):33–41, 2008.

D. Alcal-Lpez, J. Smallwood, E. Jefferies, F. Van Overwalle, K. Vogeley, R. B. Mars, B. I. Turetsky, A. R. Laird, P. T. Fox, S. B. Eickhoff, and D. Bzdok. Computing the social brain connectome across systems and states. Cerebral Cortex, pages 1–26, 2017.

J. R. Andrews-Hanna, J. S. Reidler, J. Sepulcre, R. Poulin, and R. L. Buckner. Functional-anatomic fractionation of the brain's default network. Neuron, 65(4):550–62, 2010.

J. S. Antrobus, J. L. Singer, and S. Greenberg. Studies in the stream of consciousness: experimental enhancement and suppression of spontaneous cognitive processes. Perceptual and Motor Skills, 1966.

D. Aronov, R. Nevers, and D. W. Tank. Mapping of a non-spatial dimension by the hippocampalentorhinal circuit. Nature, 543(7647):719–722, 2017.

V. Axelrod, G. Rees, and M. Bar. The default network and the combination of cognitive processes that mediate self-generated thought. Nat. Hum. Behav., 1(12):896–910, 2017.

A. P. Baker, M. J. Brookes, I. A. Rezek, S. M. Smith, T. Behrens, P. J. P. Smith, and M. Woolrich. Fast transient networks in spontaneous human brain activity. Elife, 3:e01867, 2014.

C. L. Baker, R. Saxe, and J. B. Tenenbaum. Action understanding as inverse planning. Cognition, 113(3):329–349, 2009.

D. Bálint et al. Seelenlähmung des schauens, optische ataxie, räumliche störung der aufmerksamkeit. pp. 51–66. European Neurology, 25(1):51–66, 1909.

D. Balslev, F. A. Nielsen, O. B. Paulson, and I. Law. Right temporoparietal cortex activation during visuo-proprioceptive conflict. Cereb Cortex, 15(2):166–9, 2005.

M. Bar. The proactive brain: using analogies and associations to generate predictions. Trends in cognitive sciences, 11(7):280–289, 2007.

M. Bar. The proactive brain: memory for predictions. Philosophical Transactions of the Royal Society of London B: Biological Sciences, 364(1521):1235–1243, 2009.

M. Bar, E. Aminoff, M. Mason, and M. Fenske. The units of thought. Hippocampus, 2007.

M. G. Baxter and E. A. Murray. The amygdala and reward. Nature reviews neuroscience, 3(7): 563–573, 2002.

T. E. Behrens, L. T. Hunt, M. W. Woolrich, and M. F. Rushworth. Associative learning of social value. Nature, 456(7219):245–249, 2008.

J. R. Binder and R. H. Desai. The neurobiology of semantic memory. Trends in cognitive sciences, 15(11):527–536, 2011.

J. R. Binder, J. A. Frost, T. A. Hammeke, P. S. F. Bellgowan, S. M. Rao, and R. W. Cox. Conceptual processing during the conscious resting state: a functional mri study. Journal of cognitive neuroscience, 11(1):80–93, 1999.

J. R. Binder, R. H. Desai, W. W. Graves, and L. L. Conant. Where is the semantic system? a critical review and meta-analysis of 120 functional neuroimaging studies. Cereb Cortex, 19 (12):2767–96, 2009.

C. M. Bird, C. Capponi, J. A. King, C. F. Doeller, and N. Burgess. Establishing the boundaries: the hippocampal contribution to imagining scenes. Journal of Neuroscience, 30 (35):11688–11695, 2010.

O. Blanke, S. Ortigue, T. Landis, and M. Seeck. Neuropsychology: Stimulating illusory own-body perceptions. Nature, 419(6904):269–270, 2002.

H. Blumenfeld. Neuroanatomy Through Clinical Cases. Sinauer Associates, 2002.

P. Boyer. Evolutionary economics of mental time travel? Trends in cognitive sciences, 12(6): 219–224, 2008.

T. S. Braver and S. R. Bongiolatti. The role of frontopolar cortex in subgoal processing during working memory. Neuroimage, 15(3):523–536, 2002.

T. Brázdil, K. Chatterjee, V. Forejt, and A. Kucera. Trading performance for stability in markov decision processes. J. Comput. Syst. Sci., 84:144–170, 2017.

K. Brodmann. Vergleichende Lokalisationslehre der Grosshirnrinde in ihren Prinzipien dargestellt auf Grund des Zellenbaues. Barth, 1909.

M. J. Brookes, M. Woolrich, H. Luckhoo, D. Price, J. R. Hale, M. C. Stephenson, G. R. Barnes, S. M. Smith, and P. G. Morris. Investigating the electrophysiological basis of resting state networks using magnetoencephalography. Proceedings of the National Academy of Sciences, 108(40):16783–16788, 2011.

T. G. Brown. On the nature of the fundamental activity of the nervous centres; together with an analysis of the conditioning of rhythmic activity in progression, and a theory of the evolution of function in the nervous system. The Journal of physiology, 48(1):18–46, 1914.

R. L. Buckner and D. C. Carroll. Self-projection and the brain. Trends in cognitive sciences, 11(2):49–57, 2007.

R. L. Buckner, J. R. Andrews-Hanna, and D. L. Schacter. The brain's default network: anatomy, function, and relevance to disease. Ann N Y Acad Sci, 1124:1–38, 2008.

L. Buhry, A. H. Azizi, and S. Cheng. Reactivation, replay, and preplay: how it might all fit together. Neural Plast., 2011:203462, 2011.

N. Burgess. Spatial cognition and the brain. Annals of the New York Academy of Sciences, 1124(1):77–97, 2008.

P. W. Burgess, E. Veitch, A. de Lacy Costello, and T. Shallice. The cognitive and neuroanatomical correlates of multitasking. Neuropsychologia, 38(6):848–863, 2000.

G. Buzsáki. Large-scale recording of neuronal ensembles. Nature neuroscience, 7(5):446–451, 2004.

G. Buzsáki. Rhythms of the Brain. Oxford University Press, 2006.

D. Bzdok and S. Eickhoff. The resting-state physiology of the human cerebral cortex. Technical report, Brain mapping: An encyclopedic reference, 2015.

D. Bzdok, L. Schilbach, K. Vogeley, K. Schneider, A. R. Laird, R. Langner, and S. B. Eickhoff. Parsing the neural correlates of moral cognition: Ale meta-analysis on morality, theory of mind, and empathy. Brain Struct Funct, 217(4):783–796, 2012.

D. Bzdok, A. R. Laird, K. Zilles, P. T. Fox, and S. B. Eickhoff. An investigation of the structural, connectional, and functional subspecialization in the human amygdala. Hum Brain Mapp, 34(12):3247–66, 2013a.

D. Bzdok, R. Langner, L. Schilbach, D. A. Engemann, A. R. Laird, P. T. Fox, and S. Eickhoff. Segregation of the human medial prefrontal cortex in social cognition. Frontiers in human neuroscience, 7:232, 2013b.

D. Bzdok, R. Langner, L. Schilbach, O. Jakobs, C. Roski, S. Caspers, A. R. Laird, P. Fox, K. Zilles, and S. B. Eickhoff. Characterization of the temporo-parietal junction by combining data-driven parcellation, complementary connectivity analyses, and functional decoding. Neuroimage, 81:381392, 2013c.

D. Bzdok, A. Heeger, R. Langner, A. R. Laird, P. T. Fox, N. Palomero-Gallagher, B. A. Vogt, K. Zilles, and S. B. Eickhoff. Subspecialization in the human posterior medial cortex. Neuroimage, 106:55–71, 2015.

D. Bzdok, G. Hartwigsen, A. Reid, A. R. Laird, P. T. Fox, and S. B. Eickhoff. Left inferior parietal lobe engagement in social cognition and language. Neuroscience & Biobehavioral Reviews, 68:319–334, 2016a.

D. Bzdok, G. Varoquaux, O. Grisel, M. Eickenberg, C. Poupon, and B. Thirion. Formal models of the network co-occurrence underlying mental operations. PLoS Comput Biol, 12(6): e1004994, 2016b.

R. L. Carhart-Harris and K. J. Friston. The default-mode, ego-functions and free-energy: a neurobiological account of freudian ideas. Brain, page awq010, 2010.

W. A. Carlezon and M. J. Thomas. Biological substrates of reward and aversion: A nucleus accumbens activity hypothesis. Neuropharmacology, 56:122 – 132, 2009. ISSN 0028-3908. Frontiers in Addiction Research: Celebrating the 35th Anniversary of the National Institute on Drug Abuse.

S. Caspers, S. Geyer, A. Schleicher, H. Mohlberg, K. Amunts, and K. Zilles. The human inferior parietal cortex: cytoarchitectonic parcellation and interindividual variability. Neuroimage, 33:430–448, 2006.

S. Caspers, S. B. Eickhoff, S. Geyer, F. Scheperjans, H. Mohlberg, K. Zilles, and K. Amunts. The human inferior parietal lobule in stereotaxic space. Brain Struct Funct, 212(6):481–95, 2008.

S. Caspers, S. Eickhoff, T. Rick, A. von Kapri, T. Kuhlen, R. Huang, N. J. Shah, and K. Zilles. Probabilistic bre tract analysis of cytoarchitectonically dened human inferior parietal lobule areas reveals similarities to macaques. Neuroimage, 58(2):362–380, 2011.

A. E. Cavanna and M. R. Trimble. The precuneus: a review of its functional anatomy and behavioural correlates. Brain, 129(3):564–583, 2006.

O. Y. Chao, S. Nikolaus, M. L. Brandão, J. P. Huston, and M. A. de Souza Silva. Interaction between the medial prefrontal cortex and hippocampal ca1 area is essential for episodic-like memory in rats. Neurobiology of Learning and Memory, 141:72–77, 2017.

K. Christoff, Z. C. Irving, K. C. Fox, R. N. Spreng, and J. R. Andrews-Hanna. Mind-wandering as spontaneous thought: a dynamic framework. Nature Reviews Neuroscience, 2016.

A. Clark. Whatever next? predictive brains, situated agents, and the future of cognitive science. Behavioral and Brain Sciences, 36(03):181–204, 2013.

A. O. Constantinescu, J. X. OReilly, and T. E. Behrens. Organizing conceptual knowledge in humans with a gridlike code. Science, 352(6292):1464–1468, 2016a.

A. O. Constantinescu, J. X. OReilly, and T. E. Behrens. Organizing conceptual knowledge in humans with a gridlike code. Science, 352(6292):1464–1468, 2016b.

M. Corbetta and G. L. Shulman. Control of goal-directed and stimulus-driven attention in the brain. Nature reviews neuroscience, 3(3):201–215, 2002.

M. Corbetta, J. M. Kincade, J. M. Ollinger, M. P. McAvoy, and G. L. Shulman. Voluntary orienting is dissociated from target detection in human posterior parietal cortex. Nat Neurosci, 3(3):292–7, 2000.

M. Corbetta, G. Patel, and G. L. Shulman. The reorienting system of the human brain: from environment to theory of mind. Neuron, 58(3):306–24, 2008.

P. L. Croxson, H. Johansen-Berg, T. E. Behrens, M. D. Robson, M. A. Pinsk, C. G. Gross, W. Richter, M. C. Richter, S. Kastner, and M. F. Rushworth. Quantitative investigation of connections of the prefrontal cortex in the human and macaque using probabilistic diffusion tractography. The Journal of neuroscience, 25(39):8854–8866, 2005.

A. R. Damasio, B. J. Everitt, and D. Bishop. The somatic marker hypothesis and the possible functions of the prefrontal cortex [and discussion]. Philosophical Transactions of the Royal Society of London B: Biological Sciences, 351(1346):1413–1420, 1996.

A. S. Dave and D. Margoliash. Song replay during sleep and computational rules for sensorimotor vocal learning. Science, 290(5492):812–816, Oct 2000.

N. D. Daw and P. Dayan. The algorithmic anatomy of model-based evaluation. Phil. Trans. R. Soc. B, 369(1655):20130478, 2014.

P. Dayan and N. D. Daw. Decision theory, reinforcement learning, and the brain. Cognitive, Affective, & Behavioral Neuroscience, 8(4):429–453, 2008.

P. Dayan, G. E. Hinton, R. M. Neal, and R. S. Zemel. The helmholtz machine. Neural computation, 7(5):889–904, 1995.

G. De Lavilléon, M. M. Lacroix, L. Rondi-Reig, and K. Benchenane. Explicit memory creation during sleep demonstrates a causal role of place cells in navigation. Nature neuroscience, 18 (4):493–495, 2015.

F. De Pasquale, S. Della Penna, A. Z. Snyder, C. Lewis, D. Mantini, L. Marzetti, P. Belardinelli, L. Ciancetta, V. Pizzella, G. L. Romani, et al. Temporal dynamics of spontaneous meg activity in brain networks. Proceedings of the National Academy of Sciences, 107(13):6040–6045, 2010.

M. A. H. Dempster and V. Leemans. An automated fx trading system using adaptive reinforcement learning. Expert Systems with Applications, 30(3):543–552, Apr. 2006.

L. Deuker, J. L. Bellmund, T. N. Schröder, and C. F. Doeller. An event map of memory space in the hippocampus. eLife, 5:e16534, 2016.

K. Diba and G. Buzsáki. Forward and reverse hippocampal place-cell sequences during ripples. Nature neuroscience, 10(10):1241–1242, 2007.

V. Doria, C. F. Beckmann, T. Arichia, N. Merchanta, M. Groppoa, F. E. Turkheimerb, S. J. Counsella, M. Murgasovad, P. Aljabard, R. G. Nunesa, D. J. Larkmana, G. Reese, and A. D. Edwards. Emergence of resting state networks in the preterm human brain. Proc Natl Acad Sci U S A, 107(46):20015–20020, 2010.

J. Downar, A. P. Crawley, D. J. Mikulis, and K. D. Davis. A multimodal cortical network for the detection of changes in the sensory environment. Nature neuroscience, 3(3):277–283, 2000.

G. Dumas, M. Chavez, J. Nadel, and J. Martinerie. Anatomical Connectivity Influences both Intra- and Inter-Brain Synchronizations. PLoS ONE, 7(5):e36414, May 2012.

R. I. Dunbar, A. Marriott, and N. D. Duncan. Human conversational behavior. Human Nature, 8(3):231–246, 1997.

R. I. M. Dunbar and S. Shultz. Evolution in the social brain. Science, 317(5843):1344–1347, 2007.

S. Elfwing, E. Uchibe, and K. Doya. From free energy to expected energy: Improving energy-based value function approximation in reinforcement learning. Neural Networks, 84: 17–27, 2016.

R. A. Epstein. Parahippocampal and retrosplenial contributions to human spatial navigation. Trends in cognitive sciences, 12(10):388–396, 2008.

M. S. Filler and L. M. Giambra. Daydreaming as a function of cueing and task difficulty. Perceptual and Motor Skills, 1973.

J. Fiser, C. Chiu, and M. Weliky. Small modulation of ongoing cortical dynamics by sensory input during natural vision. Nature, 431(7008):573–578, 2004.

P. Flechsig. Anatomie des menschlichen Gehirns und Rckenmarks auf myelogenetisch Grundlage. Thieme, Leipzig, 1920.

M. D. Fox, A. Z. Snyder, J. L. Vincent, M. Corbetta, D. C. Van Essen, and M. E. Raichle. The human brain is intrinsically organized into dynamic, anticorrelated functional networks. Proc Natl Acad Sci U S A, 102(27):9673–8, 2005.

K. Friston. Hierarchical models in the brain. PLoS Comput Biol, 4(11):e1000211, 2008.

K. Friston. The free-energy principle: a unified brain theory? Nature Reviews Neuroscience, 11(2):127–138, 2010.

K. Friston, C. Thornton, and A. Clark. Free-energy minimization and the dark-room problem. In Front. Psychology, 2012.

K. J. Friston, J. Daunizeau, and S. J. Kiebel. Reinforcement learning or active inference? PLoS ONE, 4(7):e6421, 2009.

K. J. Friston, K. E. Stephan, R. Montague, and R. J. Dolan. Computational psychiatry: the brain as a phantastic organ. Lancet Psychiatry, 1:148158, 2014.

H. Gelbard-Sagiv, R. Mukamel, M. Harel, R. Malach, and I. Fried. Internally generated reactivation of single neurons in human hippocampus during free recall. Science, 322(5898): 96–101, 2008.

S. J. Gershman, E. J. Horvitz, and J. B. Tenenbaum. Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. Science, 349(6245):273–278, 2015.

S. Geva, P. S. Jones, J. T. Crinion, C. J. Price, J.-C. Baron, and E. A. Warburton. The neural correlates of inner speech defined by voxel-based lesion–symptom mapping. Brain, 134(10): 3071–3082, 2011.

M. Ghavamzadeh, S. Mannor, J. Pineau, A. Tamar, et al. Bayesian reinforcement learning: A survey. Foundations and Trends® in Machine Learning, 8(5-6):359–483, 2015.

S. Ghods-Sharifi, J. R. S. Onge, and S. B. Floresco. Fundamental contribution by the basolateral amygdala to different forms of decision making. Journal of Neuroscience, 29(16): 5251–5259, 2009.

T. Gisiger, M. Kerszberg, and J.-P. Changeux. Acquisition and Performance of Delayed-response Tasks: a Neural Network Model. Cerebral Cortex, 15(5):489–506, May 2005. ISSN 1047-3211, 1460-2199. bibtex: gisiger_acquisition_2005.

J. Gläscher, R. Adolphs, H. Damasio, A. Bechara, D. Rudrauf, M. Calamia, L. K. Paul, and D. Tranel. Lesion mapping of cognitive control and value-based decision making in the prefrontal cortex. Proceedings of the National Academy of Sciences, 109(36):14681–14686, 2012.

P. S. Goldman-Rakic. Development of cortical circuitry and cognitive function. Child development, pages 601–622, 1987.

P. S. Goldman-Rakic, A. Cools, and K. Srivastava. The prefrontal landscape: implications of functional architecture for understanding human mentation and the central executive [and discussion]. Philosophical Transactions of the Royal Society B: Biological Sciences, 351 (1346):1445–1453, 1996.

Y. Golland, S. Bentin, H. Gelbard, Y. Benjamini, R. Heller, Y. Nir, U. Hasson, and R. Malach. Extrinsic and intrinsic systems in the posterior cortex of the human brain revealed during natural sensory stimulation. Cerebral cortex, 17(4):766–777, 2006.

I. Goodfellow, Y. Bengio, and A. Courville. Deep learning. MIT Press, 2016.

L. Green and J. Myerson. A discounting framework for choice with delayed and probabilistic rewards. Psychological bulletin, 130(5):769, 2004.

D. A. Gusnard and M. E. Raichle. Searching for a baseline: functional imaging and the resting human brain. Nat Rev Neurosci, 2(10):685–94, 2001.

D. A. Gusnard, E. Akbudak, G. L. Shulman, and M. E. Raichle. Medial prefrontal cortex and self-referential mental activity: relation to a default mode of brain function. Proceedings of the National Academy of Sciences, 98(7):4259–4264, 2001.

S. Haber, K. Kunishio, M. Mizobuchi, and E. Lynd-Balta. The orbital and medial prefrontal circuit through the primate basal ganglia. The Journal of neuroscience, 15(7):4851–4867, 1995a.

S. N. Haber, K. Kunishio, M. Mizobuchi, and E. Lynd-Balta. The orbital and medial prefrontal circuit through the primate basal ganglia. J Neurosci, 15(7 Pt 1):4851–67, 1995b.

P. Hagmann, L. Cammoun, X. Gigandet, R. Meuli, C. J. Honey, V. J. Wedeen, and O. Sporns. Mapping the structural core of human cerebral cortex. PLoS Biol, 6(7):e159, 2008.

B. Hahn, T. J. Ross, and E. A. Stein. Cingulate activation increases dynamically with response speed under stimulus unpredictability. Cerebral cortex, 17(7):1664–1671, 2007.

A. F. d. C. Hamilton and S. T. Grafton. Action outcomes are represented in human inferior frontoparietal cortex. Cerebral Cortex, 18(5):1160–1168, 2008.

T. Hartley, C. Lever, N. Burgess, and J. O'Keefe. Space in the brain: how the hippocampal formation supports spatial cognition. Phil. Trans. R. Soc. B, 369(1635):20120510, 2014.

K. Hartmann, G. Goldenberg, M. Daumüller, and J. Hermsdörfer. It takes the whole brain to make a cup of coffee: the neuropsychology of naturalistic actions involving technical devices. Neuropsychologia, 43(4):625–637, 2005.

D. Hassabis and E. A. Maguire. The construction system of the brain. Philosophical Transactions of the Royal Society of London B: Biological Sciences, 364(1521):1263–1271, 2009.

D. Hassabis, D. Kumaran, S. D. Vann, and E. A. Maguire. Patients with hippocampal amnesia cannot imagine new experiences. Proceedings of the National Academy of Sciences, 104(5): 1726–1731, 2007.

T. Hastie, R. Tibshirani, and J. Friedman. The Elements of Statistical Learning. Springer Series in Statistics, Heidelberg, Germany, 2011.

J. Haushofer and E. Fehr. On the psychology of poverty. Science, 344(6186):862–867, 2014.

B. Y. Hayden, A. C. Nair, A. N. McCoy, and M. L. Platt. Posterior cingulate cortex mediates outcome-contingent allocation of behavior. Neuron, 60:19–25, 2008.

B. Y. Hayden, D. V. Smith, and M. L. Platt. Electrophysiological correlates of default-mode processing in macaque posterior cingulate cortex. Proceedings of the National Academy of Sciences, 106(14):5948–5953, 2009.

M. Hessel, J. Modayil, H. van Hasselt, T. Schaul, G. Ostrovski, W. Dabney, D. Horgan, B. Piot, M. G. Azar, and D. Silver. Rainbow: Combining improvements in deep reinforcement learning. CoRR, abs/1710.02298, 2017.

S. G. Horovitz, A. R. Braun, W. S. Carr, D. Picchioni, T. J. Balkin, M. Fukunaga, and J. H. Duyn. Decoupling of the brain's default mode network during deep sleep. Proceedings of the National Academy of Sciences, 106(27):11376–11381, 2009.

H. Hult and J. Kiessling. Algorithmic trading with markov chains. 2010.

O. Jakobs, L. E. Wang, M. Dafotakis, C. Grefkes, K. Zilles, and S. B. Eickhoff. Effects of timing and movement uncertainty implicate the temporo-parietal junction in the prediction of forthcoming motor actions. Neuroimage, 47(2):667–677, 2009.

W. James. The principles of psychology. Holt and company, 1890.

A.-H. Javadi, B. Emo, L. R. Howard, F. E. Zisch, Y. Yu, R. Knight, J. P. Silva, and H. J. Spiers. Hippocampal and prefrontal processing of network topology to simulate the future. Nature Communications, 8:14652, 2017.

A. Johnson and A. D. Redish. Neural ensembles in ca3 transiently encode paths forward of the animal at a decision point. Journal of Neuroscience, 27(45):12176–12189, 2007.

K. K. Kampe, C. D. Frith, R. J. Dolan, and U. Frith. Psychology: Reward value of attractiveness and gaze. Nature, 413(6856):589–589, 2001.

R. Kaplan, J. King, R. Koster, W. D. Penny, N. Burgess, and K. J. Friston. The neural representation of prospective choice during spatial planning and decisions. PLoS biology, 15 (1):e1002588, 2017.

T. Kenet, D. Bibitchkov, M. Tsodyks, A. Grinvald, and A. Arieli. Spontaneously emerging cortical representations of visual attributes. Nature, 425(6961):954–956, 2003.

D. P. Kingma and M. Welling. Auto-encoding variational bayes. Proceedings of the 2nd International Conference on Learning Representations (ICLR), (2014), 2013.

E. Koechlin, G. Basso, P. Pietrini, S. Panzer, and J. Grafman. The role of the anterior prefrontal cortex in human cognition. Nature, 399(6732):148–151, 1999.

E. Koechlin, G. Corrado, P. Pietrini, and J. Grafman. Dissociating the role of the medial and lateral anterior prefrontal cortex in human planning. Proceedings of the National Academy of Sciences, 97(13):7651–7656, 2000.

K. P. Körding and D. M. Wolpert. Bayesian integration in sensorimotor learning. Nature, 427 (6971):244–247, 2004.

F. M. Krienen, P.-C. Tu, and R. L. Buckner. Clan mentality: evidence that the medial prefrontal cortex responds to close others. Journal of Neuroscience, 30(41):13906–13915, 2010.

M. L. Kringelbach and E. T. Rolls. The functional neuroanatomy of the human orbitofrontal cortex: evidence from neuroimaging and neuropsychology. Prog Neurobiol, 72(5):341–72, 2004.

A. R. Laird, S. B. Eickhoff, K. Li, D. A. Robin, D. C. Glahn, and P. T. Fox. Investigating the functional heterogeneity of the default mode network using coordinate-based meta-analytic modeling. J Neurosci, 29(46):14496–505, 2009.

M. Lebreton, S. Jorge, V. Michel, B. Thirion, and M. Pessiglione. An automatic valuation system in the human brain: evidence from functional neuroimaging. Neuron, 64(3):431–439, 2009.

R. Leech and D. J. Sharp. The role of the posterior cingulate cortex in cognition and disease. Brain, 137(Pt 1):12–32, 2014.

M. Liljeholm, S. Wang, J. Zhang, and J. P. O'Doherty. Neural correlates of the divergence of instrumental probability distributions. The Journal of Neuroscience, 33(30):12519–12527, 2013.

M. Lombardo, B. Chakrabarti, E. Bullmore, S. Wheelwright, S. Sadek, J. Suckling, and S. Baron-Cohen. Shared neural circuits for mentalizing about the self and others. J Cogn Neurosci, 22(7):1623–1635, 2009.

H. Lu, Q. Zou, H. Gu, M. E. Raichle, E. A. Stein, and Y. Yang. Rat brains also have a default mode network. Proceedings of the National Academy of Sciences, 109(10):3979–3984, 2012.

G. Luksys, W. Gerstner, and C. Sandi. Stress, genotype and norepinephrine in the prediction of mouse behavior using reinforcement learning. Nature neuroscience, 12(9):1180, 2009.

E. A. Maguire, D. G. Gadian, I. S. Johnsrude, C. D. Good, J. Ashburner, R. S. Frackowiak, and C. D. Frith. Navigation-related structural change in the hippocampi of taxi drivers. Proceedings of the National Academy of Sciences, 97(8):4398–4403, 2000.

D. Mantini, A. Gerits, K. Nelissen, J.-B. Durand, O. Joly, L. Simone, H. Sawamura, C. Wardak, G. A. Orban, R. L. Buckner, et al. Default mode of brain function in monkeys. The Journal of Neuroscience, 31(36):12954–12962, 2011.

D. S. Margulies, S. S. Ghosh, A. Goulas, M. Falkiewicz, J. M. Huntenburg, G. Langs, G. Bezgin, S. B. Eickhoff, F. X. Castellanos, M. Petrides, E. Jefferies, and J. Smallwood. Situating the default-mode network along a principal gradient of macroscale cortical organization. Proceedings of the National Academy of Sciences, page 201608282, Oct. 2016a.

D. S. Margulies, S. S. Ghosh, A. Goulas, M. Falkiewicz, J. M. Huntenburg, G. Langs, G. Bezgin, S. B. Eickhoff, F. X. Castellanos, M. Petrides, et al. Situating the default-mode network along a principal gradient of macroscale cortical organization. Proceedings of the National Academy of Sciences, page 201608282, 2016b.

R. B. Mars, S. Jbabdi, J. Sallet, J. X. O'Reilly, P. L. Croxson, E. Olivier, M. P. Noonan, C. Bergmann, A. S. Mitchell, M. G. Baxter, T. E. Behrens, H. Johansen-Berg, V. Tomassini, K. L. Miller, and M. F. Rushworth. Diffusion-weighted imaging tractography-based parcellation of the human parietal cortex and comparison with human and macaque resting-state functional connectivity. J Neurosci, 31(11):4087–100, 2011.

R. B. Mars, F. Neubert, M. P. Noonan, J. Sallet, I. Toni, and M. F. Rushworth. On the relationship between the default mode network and the social brain. Front Hum Neurosci, 6 (Article 189), 2012.

M. F. Mason, M. I. Norton, J. D. Van Horn, D. M. Wegner, S. T. Grafton, and C. N. Macrae. Wandering minds: the default network and stimulus-independent thought. Science, 315: 393–395, 2007.

A. N. McCoy and M. L. Platt. Risk-sensitive neurons in macaque posterior cingulate cortex. Nature neuroscience, 8(9):1220–1227, 2005.

M.-M. Mesulam. From sensation to cognition. Brain, 121(6):1013–1052, 1998.

M.-M. Mesulam. Principles of behavioral and cognitive neurology. Oxford University Press, 2000. ISBN 0198030800.

V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis. Human-level control through deep reinforcement learning. Nature, 518(7540):529–533, Feb 2015. Letter.

S. Mohamed and D. J. Rezende. Variational information maximisation for intrinsically motivated reinforcement learning. In Advances in neural information processing systems, pages 2125–2133, 2015.

P. R. Montague, B. King-Casas, and J. D. Cohen. Imaging valuation models in human choice. Annu. Rev. Neurosci., 29:417–448, 2006.

J. M. Moran, E. Jolly, and J. P. Mitchell. Social-cognitive deficits in normal aging. The Journal of Neuroscience, 32(16):5553–5561, 2012.

A. Ng, A. Coates, M. Diel, V. Ganapathi, J. Schulte, B. Tse, E. Berger, and E. Liang. Autonomous inverted helicopter flight via reinforcement learning. In International Symposium on Experimental Robotics, 2004.

M. S. Nokia, M. Penttonen, and J. Wikgren. Hippocampal ripple-contingent training accelerates trace eyeblink conditioning and retards extinction in rabbits. J. Neurosci., 30 (34):11486–11492, Aug 2010.

J. P. O'Doherty, S. W. Lee, and D. McNamee. The structure of reinforcement-learning mechanisms in the human brain. Current Opinion in Behavioral Sciences, 1:94–100, 2015.

R. C. O'Reilly and M. J. Frank. Making working memory work: a computational model of learning in the prefrontal cortex and basal ganglia. Neural computation, 18(2):283–328, 2006.

J. ONeill, B. Pleydell-Bouverie, D. Dupret, and J. Csicsvari. Play it again: reactivation of waking experience and memory. Trends in neurosciences, 33(5):220–229, 2010.

J. M. Pearson, B. Y. Hayden, S. Raghavachari, and M. L. Platt. Neurons in posterior cingulate cortex signal exploratory decisions in a dynamic multioption choice task. Current biology, 19(18):1532–1537, 2009.

G. Pezzulo. Grounding procedural and declarative knowledge in sensorimotor anticipation. Mind & Language, 26(1):78–114, 2011.

B. E. Pfeiffer and D. J. Foster. Hippocampal place-cell sequences depict future paths to remembered goals. Nature, 497(7447):74–79, 2013.

D. Popa, A. T. Popescu, and D. Paré. Contrasting activity profile of two distributed cortical networks as a function of attentional demands. Journal of Neuroscience, 29(4):1191–1201, 2009.

K. S. Pope and J. L. Singer. Regulation of the stream of consciousness: Toward a theory of ongoing thought. In Consciousness and self-regulation, pages 101–137. Springer, 1978.

A. Pritzel, B. Uria, S. Srinivasan, A. Puigdomènech, O. Vinyals, D. Hassabis, D. Wierstra, and C. Blundell. Neural episodic control. arXiv preprint arXiv:1703.01988, 2017.

N. C. Rabinowitz, F. Perbet, H. F. Song, C. Zhang, S. M. A. Eslami, and M. Botvinick. Machine theory of mind. CoRR, abs/1802.07740, 2018.

M. E. Raichle. The brain's dark energy. Science, 314(5803):1249–1250, 2006.

M. E. Raichle and D. A. Gusnard. Intrinsic brain activity sets the stage for expression of motivated behavior. Journal of Comparative Neurology, 493(1):167–176, 2005.

M. E. Raichle, A. M. MacLeod, A. Z. Snyder, W. J. Powers, D. A. Gusnard, and G. L. Shulman. A default mode of brain function. Proceedings of the National Academy of Sciences of the United States of America, 98(2):676–82, 2001.

P.-A. Salin and J. Bullier. Corticocortical connections in the visual system: structure and function. Physiological reviews, 75(1):107–155, 1995.

B. Sallans and G. E. Hinton. Reinforcement learning with factored states and actions. J. Mach. Learn. Res., 5:1063–1088, Dec. 2004. ISSN 1532-4435.

D. L. Schacter, D. R. Addis, and R. L. Buckner. Remembering the past to imagine the future: the prospective brain. Nature Reviews Neuroscience, 8(9):657–661, 2007.

T. Schaul, J. Quan, I. Antonoglou, and D. Silver. Prioritized experience replay. CoRR, abs/1511.05952, 2015.

L. Schilbach, S. B. Eickhoff, A. Rotarska-Jagiela, G. R. Fink, and K. Vogeley. Minds at rest? social cognition as the default mode of cognizing and its putative relationship to the default system of the brain. Consciousness and cognition, 17(2):457–467, 2008.

L. Schilbach, D. Bzdok, B. Timmermans, P. T. Fox, A. R. Laird, K. Vogeley, and S. B. Eickhoff. Introspective minds: Using ale meta-analyses to study commonalities in the neural correlates of emotional processing, social and unconstrained cognition. PLoS One, 7(2):e30920, 2012.

J. Schmidhuber. Formal theory of creativity, fun, and intrinsic motivation (1990–2010). IEEE Transactions on Autonomous Mental Development, 2(3):230–247, 2010.

N. W. Schuck, M. B. Cai, R. C. Wilson, and Y. Niv. Human orbitofrontal cortex represents a cognitive map of state space. Neuron, 91(6):1402–1412, 2016.

W. Schultz. Predictive reward signal of dopamine neurons. Journal of neurophysiology, 80(1): 1–27, 1998.

M. L. Seghier. The angular gyrus multiple functions and multiple subdivisions. The Neuroscientist, 19(1):43–61, 2013.

M. L. Seghier, E. Fagan, and C. J. Price. Functional subdivisions in the left angular gyrus where the semantic system meets and diverges from the default network. J Neurosci, 30(50): 16809–17, 2010.

P. Seli, E. F. Risko, D. Smilek, and D. L. Schacter. Mind-wandering with and without intention. Trends in Cognitive Sciences, 20(8):605–617, 2016.

B. J. Shannon, R. A. Dosenbach, Y. Su, A. G. Vlassenko, L. J. Larson-Prior, T. S. Nolan, A. Z. Snyder, and M. E. Raichle. Morning-evening variation in human brain metabolism and memory circuits. Journal of neurophysiology, 109(5):1444–1456, 2013.

Z. Shehzad, A. M. Kelly, P. T. Reiss, D. G. Gee, K. Gotimer, L. Q. Uddin, S. H. Lee, D. S. Margulies, A. K. Roy, B. B. Biswal, E. Petkova, F. X. Castellanos, and M. P. Milham. The resting brain: unconstrained yet reliable. Cereb Cortex, 19(10):2209–29, 2009.

G. L. Shulman, J. A. Fiez, M. Corbetta, R. L. Buckner, F. M. Miezin, M. E. Raichle, and S. E. Petersen. Common blood flow changes across visual tasks .2. decreases in cerebral cortex. Journal of Cognitive Neuroscience, 9(5):648–663, 1997.

G. L. Shulman, S. V. Astafiev, M. P. McAvoy, G. d'Avossa, and M. Corbetta. Right tpj deactivation during visual search: functional significance and support for a filter hypothesis. Cereb Cortex, 17(11):2625–33, 2007.

G. L. Shulman, D. L. Pope, S. V. Astafiev, M. P. McAvoy, A. Z. Snyder, and M. Corbetta. Right hemisphere dominance during spatial selective attention and target detection occurs outside the dorsal frontoparietal network. Journal of Neuroscience, 30(10):3640–3651, 2010.

D. Silver and J. Veness. Monte-carlo planning in large pomdps. In Advances in neural information processing systems, pages 2164–2172, 2010.

D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, et al. Mastering the game of go with deep neural networks and tree search. Nature, 529(7587):484–489, 2016.

E. Simony, C. J. Honey, J. Chen, O. Lositsky, Y. Yeshurun, A. Wiesel, and U. Hasson. Dynamic reconfiguration of the default mode network during narrative comprehension. Nature Communications, 7, 2016.

W. E. Skaggs, B. L. McNaughton, M. Permenter, M. Archibeque, J. Vogt, D. G. Amaral, and C. A. Barnes. EEG sharp waves and sparse ensemble unit activity in the macaque hippocampus. J. Neurophysiol., 98(2):898–910, Aug 2007.

S. M. Smith, P. T. Fox, K. L. Miller, D. C. Glahn, P. M. Fox, C. E. Mackay, N. Filippini, K. E. Watkins, R. Toro, A. R. Laird, and C. F. Beckmann. Correspondence of the brain's functional architecture during activation and rest. Proc Natl Acad Sci U S A, 106(31): 13040–5, 2009.

Z. Song, R. E. Parr, X. Liao, and L. Carin. Linear feature encoding for reinforcement learning. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, Advances in Neural Information Processing Systems 29, pages 4224–4232. Curran Associates, Inc., 2016.

E. R. Sowell, B. S. Peterson, P. M. Thompson, S. E. Welcome, A. L. Henkenius, and A. W. Toga. Mapping cortical change across the human life span. Nature neuroscience, 6(3):309, 2003.

R. N. Spreng and B. Levine. The temporal distribution of past and future autobiographical events across the lifespan. Memory & cognition, 34(8):1644–1651, 2006.

R. N. Spreng, R. A. Mar, and A. S. Kim. The common neural basis of autobiographical memory, prospection, navigation, theory of mind, and the default mode: a quantitative meta-analysis. Journal of cognitive neuroscience, 21(3):489–510, 2009.

C. K. Starkweather, B. M. Babayan, N. Uchida, and S. J. Gershman. Dopamine reward prediction errors reflect hidden-state inference across time. Nature Neuroscience, 2017.

K. E. Stephan, G. R. Fink, and J. C. Marshall. Mechanisms of hemispheric specialization: insights from analyses of connectivity. Neuropsychologia, 45(2):209–228, 2007.

D. Stuss and D. Benson. The frontal lobes (raven, new york). StussThe Frontal Lobes1986, 1986.

T. Suddendorf and M. C. Corballis. The evolution of foresight: What is mental time travel, and is it unique to humans? Behavioral and Brain Sciences, 30(03):299–313, 2007.

R. S. Sutton and A. G. Barto. Reinforcement learning: An introduction. MIT press, 1998.

T. Takeuchi, A. J. Duszkiewicz, and R. G. Morris. The synaptic plasticity and memory hypothesis: encoding, storage and persistence. Phil. Trans. R. Soc. B, 369(1633):20130288, 2014.

P. Taylor, J. Hobbs, J. Burroni, and H. Siegelmann. The global landscape of cognition: hierarchical aggregation as an organizational principle of human cortical networks and functions. Scientific reports, 5:18112, 2015.

J. D. Teasdale, B. H. Dritschel, M. J. Taylor, L. Proctor, C. A. Lloyd, I. Nimmo-Smith, and A. D. Baddeley. Stimulus-independent thought depends on central executive resources. Memory & cognition, 23(5):551–559, 1995.

M. Tomasello. The cultural origins of human cognition. Harvard university press, 2009.

L. Q. Uddin, K. Supekar, H. Amin, E. Rykhlevskaia, D. A. Nguyen, M. D. Greicius, and V. Menon. Dissociable connectivity within human angular gyrus and intraparietal sulcus: evidence from functional and structural connectivity. Cereb Cortex, 20(11):2636–46, 2010.

C. Valiquette and T. P. McNamara. Different mental representations for place recognition and goal localization. Psychonomic Bulletin & Review, 14(4):676–680, 2007.

S. D. Vann, J. P. Aggleton, and E. A. Maguire. What does the retrosplenial cortex do? Nature Reviews Neuroscience, 10(11):792–802, 2009.

N. R. Varney and H. Damasio. Locus of lesion in impaired pantomime recognition. Cortex, 23 (4):699–703, 1987.

D. Vatansever, D. K. Menon, and E. A. Stamatakis. Default mode contributions to automated information processing. Proceedings of the National Academy of Sciences, page 201710521, 2017.

P. Vetter, B. Butterworth, and B. Bahrami. A candidate for the attentional bottleneck: set-size specific modulation of the right tpj during attentive enumeration. Journal of Cognitive Neuroscience, 23(3):728–736, 2011.

J. L. Vincent, A. Z. Snyder, M. D. Fox, B. J. Shannon, J. R. Andrews, M. E. Raichle, and R. L. Buckner. Coherent spontaneous activity identifies a hippocampal-parietal memory network. J Neurophysiol, 96(6):3517–31, 2006.

X.-J. Wang. Decision making in recurrent neuronal circuits. Neuron, 60(2):215–234, 2008.

C. J. C. H. Watkins and P. Dayan. Technical note q-learning. Machine Learning, 8:279–292, 1992.

D. H. Weissman, K. C. Roberts, K. M. Visscher, and M. G. Woldorff. The neural bases of momentary lapses in attention. Nat Neurosci, 9(7):971–978, 2006.

A. Whiten and R. W. Byrne. The machiavellian intelligence hypotheses: Editorial. 1988.

D. M. Wolpert, Z. Ghahramani, and M. I. Jordan. An internal model for sensorimotor integration. Science, 269(5232):1880, 1995.

P. Yakovlev. The myelogenetic cycles of regional maturation of the brain. Regional development of the brain in early life, pages 3–70, 1967.

S. Yang, M. Paddrik, R. Hayes, A. Todd, A. Kirilenko, P. Beling, and W. Scherer. Behavior based learning in identifying high frequency trading strategies. In Computational Intelligence for Financial Engineering & Economics (CIFEr), 2012 IEEE Conference on, pages 1–8. IEEE, 2012.

S. Y. Yang, Q. Qiao, P. A. Beling, and W. T. Scherer. Algorithmic trading behavior identification using reward learning method. In 2014 International Joint Conference on Neural Networks, IJCNN 2014, Beijing, China, July 6-11, 2014, pages 3807–3414, 2014.

S. Y. Yang, Q. Qiao, P. A. Beling, W. T. Scherer, and A. A. Kirilenko. Gaussian process-based algorithmic trading strategy identification. Quantitative Finance, 15(10):1683–1703, 2015.

W. Yoshida, B. Seymour, K. J. Friston, and R. J. Dolan. Neural mechanisms of belief inference during cooperative games. The Journal of Neuroscience, 30(32):10744–10751, 2010.

L. Young, J. A. Camprodon, M. Hauser, A. Pascual-Leone, and R. Saxe. Disruption of the right temporoparietal junction with transcranial magnetic stimulation reduces the role of beliefs in moral judgments. Proceedings of the National Academy of Sciences, 107(15): 6753–6758, 2010.

P. Zeidman and E. A. Maguire. Anterior hippocampus: the anatomy of perception, imagination and episodic memory. Nat Rev Neurosci, 17(3):173–182, 2016.

Figure1

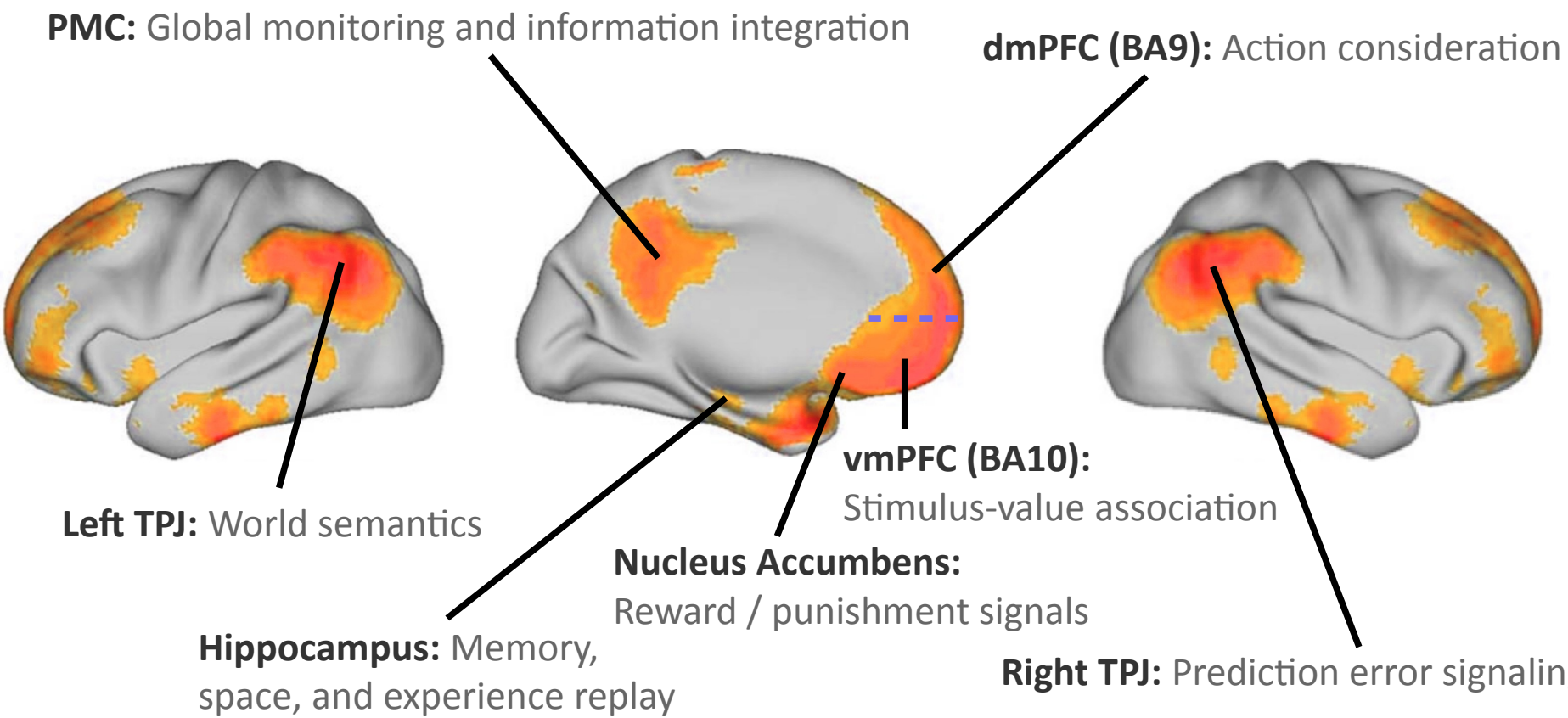

PMC: Global monitoring and information integration
dmPFC (BA9): Action consideration
Left TPJ: World semantics
vmPFC (BA10):
Stimulus-value association
Nucleus Accumbens:
Reward / punishment signals
Hippocampus: Memory,
space, and experience replay
Right TPJ: Prediction error signaling

Figure2

Click here to download high resolution image

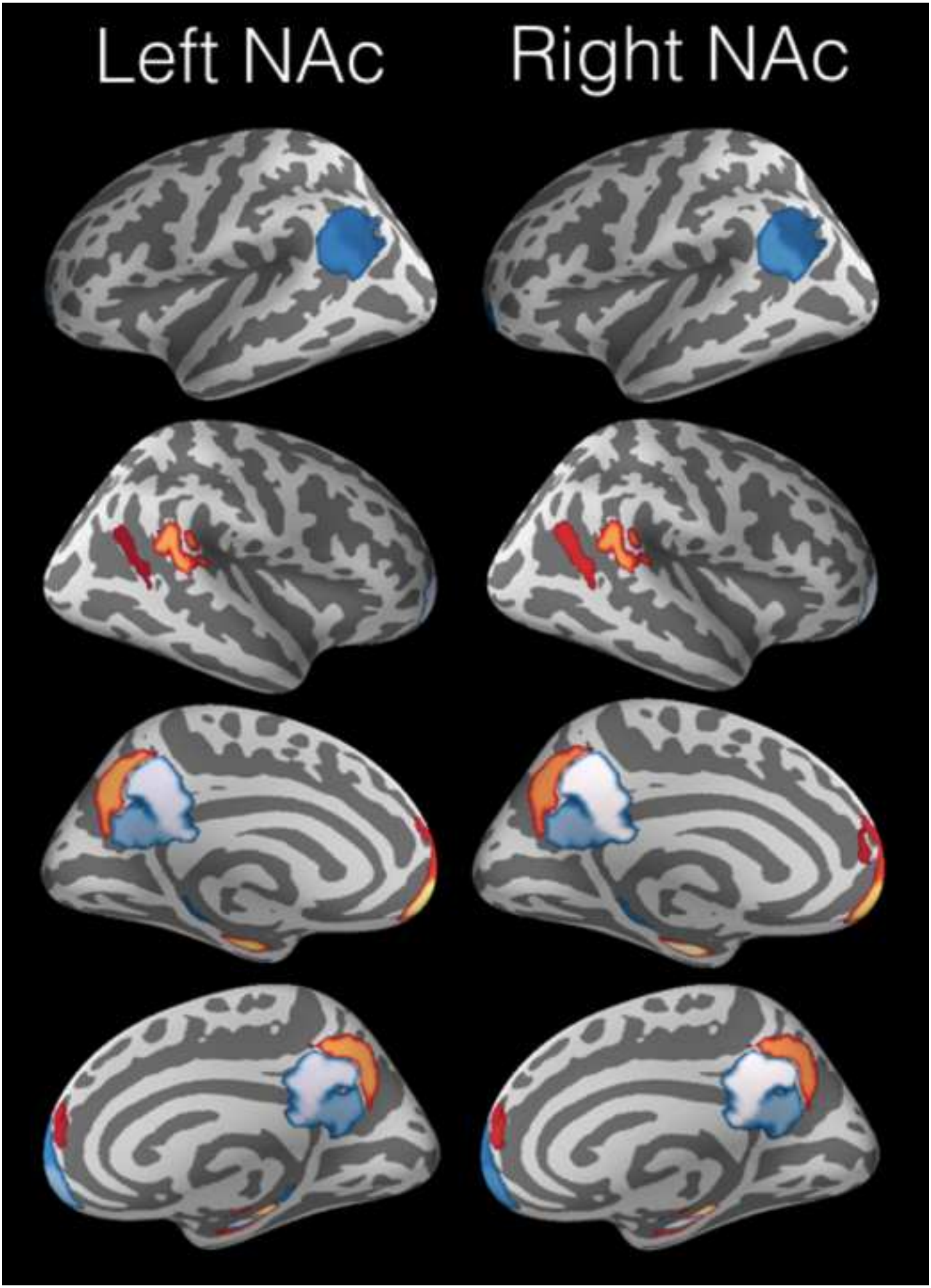


Figure3

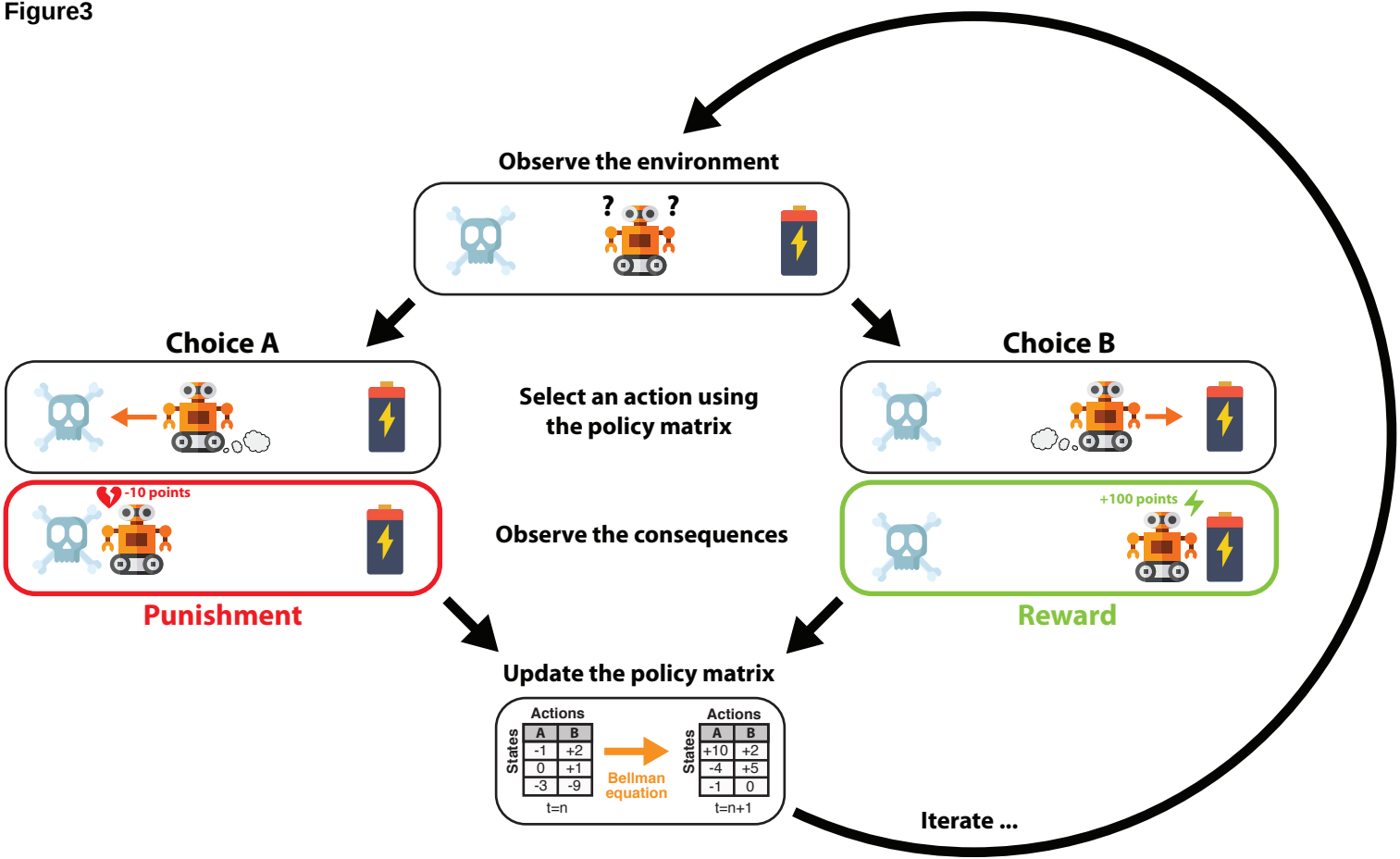

Observe the environment
Choice A
Choice B
Select an action using the policy matrix
-10 points
Observe the consequences
+100 points
Punishment
Reward
Update the policy matrix
Actions
States
A
B
-1
+2
0
+1
-3
-9
t=n
Bellman equation
Actions
States
A
B
+10
+2
-4
+5
-1
0
t=n+1
Iterate ...

Figure4

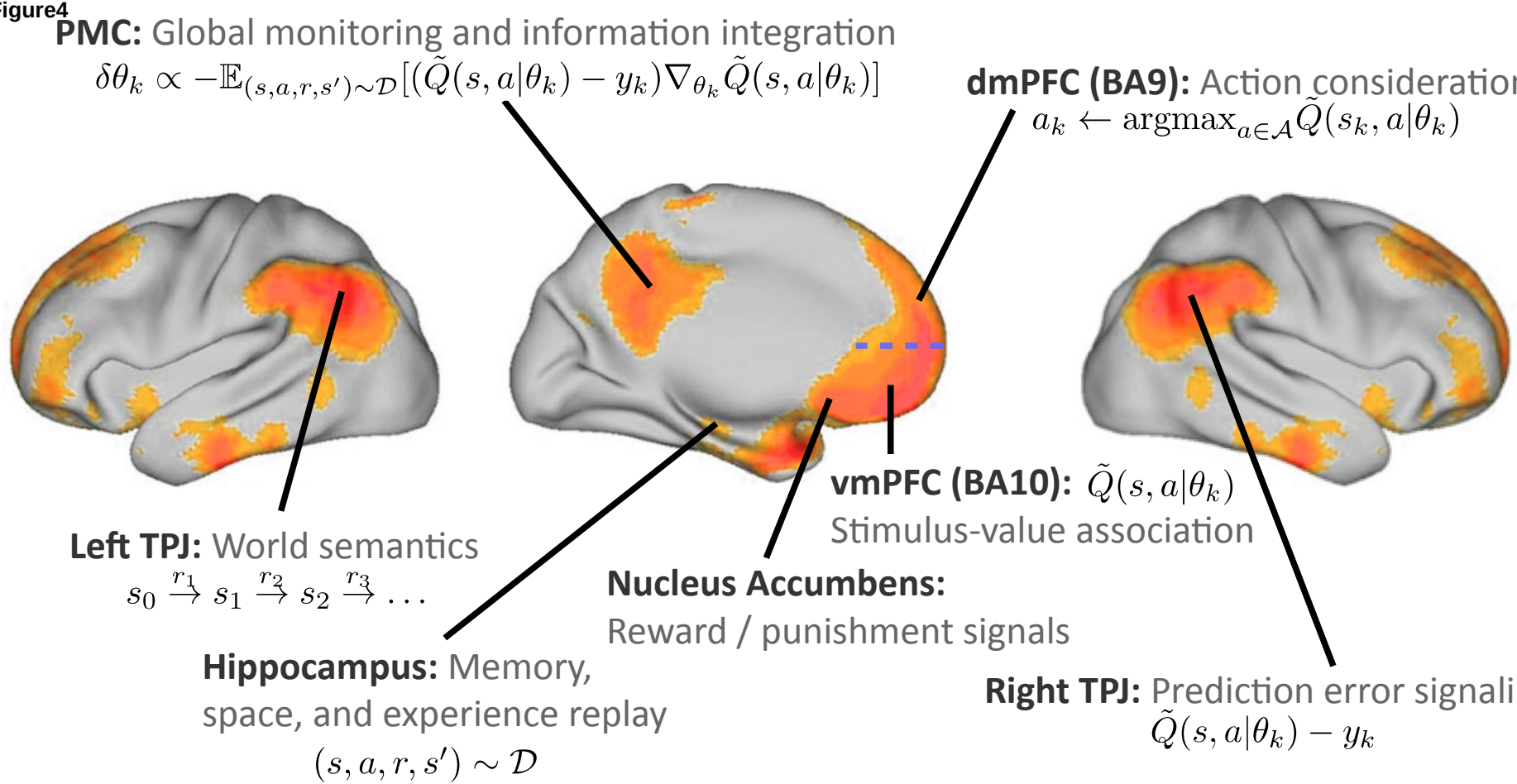
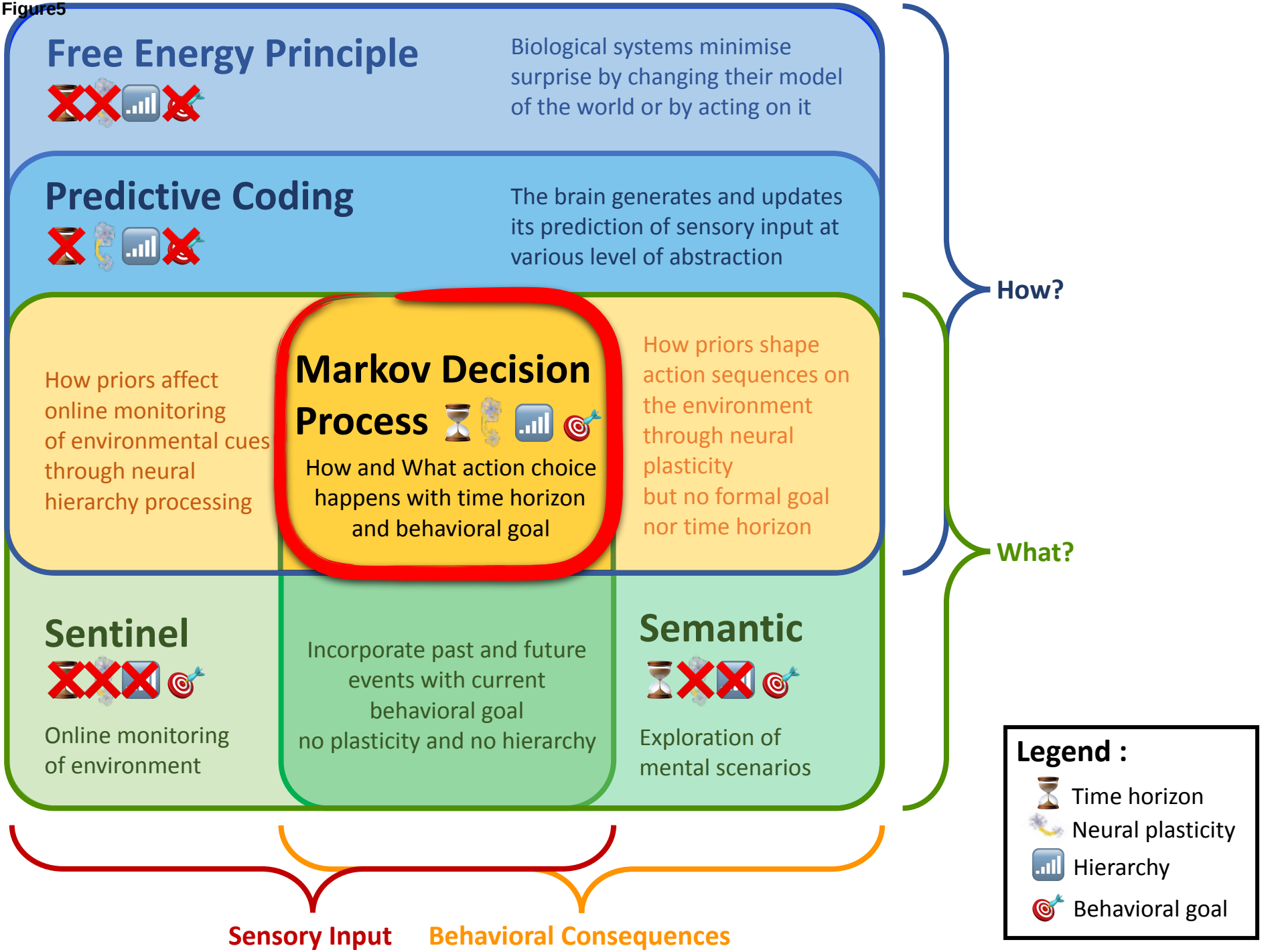

PMC: Global monitoring and information integration
$\delta\theta_k \propto -\mathbb{E}_{(s,a,r,s')\sim\mathcal{D}}[(\tilde{Q}(s,a|\theta_k) - y_k)\nabla_{\theta_k}\tilde{Q}(s,a|\theta_k)]$
dmPFC (BA9): Action consideration
$a_k \leftarrow \text{argmax}_{a\in\mathcal{A}}\tilde{Q}(s_k,a|\theta_k)$
vmPFC (BA10): $\tilde{Q}(s,a|\theta_k)$
Stimulus-value association
Left TPJ: World semantics
$s_0 \xrightarrow{r_1} s_1 \xrightarrow{r_2} s_2 \xrightarrow{r_3} \dots$
Nucleus Accumbens:
Reward / punishment signals
Hippocampus: Memory,
space, and experience replay
$(s,a,r,s') \sim \mathcal{D}$
Right TPJ: Prediction error signaling
$\tilde{Q}(s,a|\theta_k) - y_k$

Figure5


Free Energy Principle
Biological systems minimise surprise by changing their model of the world or by acting on it
Predictive Coding
The brain generates and updates its prediction of sensory input at various level of abstraction
How priors affect online monitoring of environmental cues through neural hierarchy processing
Markov Decision Process
How and What action choice happens with time horizon and behavioral goal
How priors shape action sequences on the environment through neural plasticity but no formal goal nor time horizon
Sentinel
Online monitoring of environment
Incorporate past and future events with current behavioral goal no plasticity and no hierarchy
Semantic
Exploration of mental scenarios
How?
What?
Sensory Input
Behavioral Consequences
Legend :
Time horizon
Neural plasticity
Hierarchy
Behavioral goal