Dark Control: A Unified Account of Default Mode Function by Markov Decision Processes

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

1 Université Paris-Saclay, INRIA, Parietal Team, Saclay, France
2 Université Paris-Saclay, CEA, Neurospin, Gif-sur-Yvette, France
3 Department of Psychiatry, Psychotherapy and Psychosomatics, RWTH Aachen University, Aachen, Germany
4 Institut Pasteur, Human Genetics and Cognitive Functions Unit, Paris, France
5 CNRS UMR 3571 Genes, Synapses and Cognition, Institut Pasteur, Paris, France
6 University Paris Diderot, Sorbonne Paris Cité, Paris, France
7 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 support 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. DMN function is recast in terms of control theory and reinforcement learning by basing it on Markov decision processes. We argue that our formal account of DMN function naturally accommodates as special cases the previously proposed cognitive accounts 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: mind-wandering, artificial intelligence, Markov decision processes

1 Introduction

In the absence of external stimulation, the human brain is not at rest. In the beginning of 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 (**???**). The "default mode network" (DMN) continues to metabolize large quantities of oxygen and glucose energy to maintain neuronal computation during free-ranging thought (**??**). The baseline energy demand is only weakly modulated at the onset of defined psychological tasks (**?**). 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 (**?**).

This *dark matter of brain physiology* (**?**) begs the question of the biological purpose underlying DMN activity. What has early been described as the "stream of consciousness" in psychology (**?**) found a potential neurobiological manifestation in the DMN (**??**). We propose that this set of some of the most advanced regions in the association cortex (**??**) are

responsible for higher-order control of human behavior. Our functional account 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" (**?**).

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

The network nodes that compose the human DMN are responsible for extended parts of the baseline neural activity, which typically decreases when engaged in psychological experiments (**?**). The standard mode of neural information maintenance and manipulation has been argued to mediate evolutionarily conserved functions (**???**). Today, many psychologists and neuroscientists believe that the DMN implements some form of probabilistic estimation of past, hypothetical, and future events (**??????**). This brain network might have emerged to continuously predict the environment using mental imagery as an evolutionary advantage (**?**). 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 (**?**) and increased DMN activity was linked to more task-independent, yet sometimes useful thoughts (**??**). Gaining insight into DMN function is particularly challenging because this network appears to simultaneously modulate perception-action cycles in the present and to support mental travel across time, space, and content domains (**?**).

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. Formally, a particularly attractive framework to describe, quantify, and predict intelligent systems, such as the brain, is proposed to be the combination of control theory and reinforcement learning (RL). 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. Henceforth, *control* refers to the influence that an agent exerts when interacting with the environment to reach preferred states.

Psychologically, the more the ongoing executed task is unknown and unpracticed, the less stimulus-independent thoughts occur (**???**). Conversely, it is known that, the more the world is easy to predict, the more human mental activity becomes detached from the actual sensory environment (**??**). Without requiring explicit awareness, these "offline" processes may contribute to optimizing control of the organism. 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 rewards. 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. Such an RL account of DMN function 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 advanced cognitive processes in humans can be recast as prediction error minimization based on probabilistic mental simulations, thus maximizing action outcome across multiple time scales. Such a purposeful optimization objective may be solved by a stochastic approximation based on a brain implementation of Markov Chain Monte Carlo (MCMC) sampling (**?**). Even necessarily imperfect memory recall, random day-time mind-wandering, and seemingly arbitrary dreams during sleep may provide blocks of pseudo-experience to iteratively optimize the behavior of the organism.

Evidence from computional modeling of human behavior (**?**) and cell recording experiments in ferrets (**?**) suggest that the human brain is largely dedicated to "the development and maintenance of [a] probabilistic model of anticipated events" (**?**). The present paper proposes a process model that satisfies this contention. 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 (Fig. 2). Finally, we detail how our process model relates to previous accounts of DMN function and we derive explicit hypotheses to be tested in future neuroscience experiments (Box 1).

Box 1: Hypotheses for testing the MDP account of DMN function

1. Experiment (Humans): We hypothesize 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 continuous learning paradigms. We expect that the more DMN activity is measured to be increased, supposedly the higher the tendency for stimulus-independent thoughts, and the more the fMRI signals in the reward circuits should be independent of the reward context in the sensory environment.
2. Experiment (Humans): We hypothesize a functional dissociation between computations for behavioral policy versus adapting stimulus-value associations as these are conceivably implemented in different subsystems of the DMN. We first expect that fMRI signals in the right temporo-parietal junction can predict behavioral changes caused by adaptation in the action choice tendencies (policy matrix) related to non-value-related prediction error. Second, fMRI signals in the ventromedial prefrontal cortex should predict behavioral changes caused by adaptation in value estimation (value matrix) due to reward-related stimulus-value association. We finally predict that fMRI signals in the posteromedial cortex, as a potential global information integrator, is able to predict shifts in overt behavior based on previous adaptations in policy or value estimation.
3. Experiment (Animals): We hypothesize that experience replay for browsing problem solutions subserved by the DMN is necessary for choice behavior in mice. Hippocampal single-cell recordings have shown that neural patterns during experimental choice behavior are reiterated during sleep and before make analogous choices in the future. Necessity of the DMN, in addition to the hippocampus, for mind-searching actions during choice behavior can be demonstrated by causal disruption of DMN regions, such as circumscribed brain lesion or optogenetic intervention in the inferior parietal and prefrontal cortices.

2 Known neurobiological properties of the default mode network

We begin by a neurobiological deconstruction of the DMN based on experimental findings in the neuroscience literature. This walkthrough across each region of the DMN will outline the individual functional profiles, paving the way for their algorithmic interpretation in our formal account (section 3).

2.1 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 the highest turn-over of energy consumption (**??**). These metabolic characteristics go hand-in-hand with neuroimaging analyses that suggested the PMC and mPFC to potentially represent the functional backbone of the DMN (**??**).

Normal and disturbed metabolic fluctuations in the human PMC have been closely related to changes of conscious awareness (**?**). Indeed, the PMC matures relatively late (i.e., myelination) during postnatal development in monkeys (**?**), 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" (**??**). For instance, Bálint's syndrome is a neurological disorder of conscious awareness that results from medial damage in the parietal cortex (**?**). Neurological patients are plagued by an inability to combine various individual features of the visual environment into an

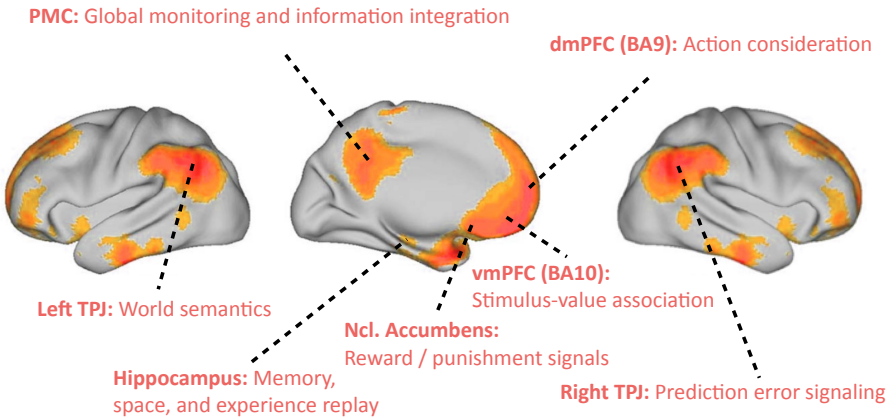

Fig 1. Default mode network: key functions. Neurobiological overview of the DMN with its major constituent nodes and the associated functional roles relevant in our functional interpretation.

integrated whole (i.e., simultanagnosia) as well as an inability to direct action towards currently unattended environmental objects (i.e., optic ataxia). This can be viewed as a high-level impairment in gathering information about alternative objects (i.e., exploration) as well as leveraging these environmental opportunities (i.e., exploitation). Congruently, the human PMC was coupled in two functional connectivity analyses (**?**) 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 (**?**). This amygdalar subregion has been proposed to continuously scan environmental input for biological relevance assessment (**?**).

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 (**?**), risk assessment (**?**), outcome-dependent behavioral modulation (**?**), as well as approach-avoidance behavior (**?**). Neuron spiking activity in the PMC allowed distinguishing whether a monkey would pursue an exploratory or exploitative behavioral strategy during food foraging (**?**). Further, single-cell recordings in the monkey PMC demonstrated this brain region's sensitivity to subjective target utility (**?**) and integration across individual decision-making instances (**?**). This DMN region encoded the preference for or aversion to options with uncertain reward outcomes and its 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" (**?**). In fact, direct stimulation of PMC neurons promoted exploratory actions, which would otherwise be shunned (**?**). 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 (**?**).

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 (**?**). Regarding perspective frames, the retrosplenial subregion of the PMC has been proposed to mediate between the organism's egocentric (i.e., focused on sensory environment) and allocentric (i.e., focused on world knowledge) viewpoints in animals and humans (**???**).

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, that differs across individuals, might enrich statistical assessment of the environment to map and predict delayed reward opportunities in the future. In doing so, the PMC may continuously adapt the organism to changes in both the external environment and its internal representation to enable strategic behavior.

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

Analogous to the PMC, the dorsomedial PFC (dmPFC) 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" (**?**), as it potentially subserves the de-novo construction and manipulation of meaning representations instructed by stored semantics and memories (**?**). The dmPFC may subserve inference, representation, and assessment of one's own and other individuals' action considerations. Generally, neurological patients with tissue damage in the prefrontal cortex are known to struggle with adaptation to new stimuli and events (**?**). 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 (**?**). Certain dmPFC neurons in macaque monkeys exhibited a preference for processing others', rather than own, behavior with fine-grained adjustment of contextual circumstances (**?**).

Comparing to the dmPFC, the vmPFC probably subserves subjective value evaluation and risk estimation of relevant environmental stimuli (Fig. **??**). The ventromedial prefrontal DMN may subserve adaptive behavior by bottom-up-driven processing of what matters now, drawing on sophisticated value representations (**??**). Quantitative lesion findings across 344 human individuals confirmed a substantial impairment in value-based action choice (**?**). 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 (**?**). Congruently, the gray-matter volume of the vmPFC and NAc correlated with indices of value-guided behavior and reward attitudes in humans (**?**). NAc activity is thought to reflect reward prediction signals from dopaminergic neurotransmitter pathways (**?**) that not only channel action towards basic survival needs but also enable more abstract reward processings, and thus perhaps RL, in humans (**?**).

Consistently, diffusion MRI tractography in humans and monkeys (**?**) quantified the NAc to be more connected to the vmPFC than dmPFC in both species. Two different functional connectivity analyses in humans also strongly connected the vmPFC with the NAc, hippocampus (HC), and PMC (**?**). In line with these connectivity findings in animals and humans, the vmPFC is often proposed to represent triggered emotional and motivational states (**?**). 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 estimates (**?**). This may be why performance within and across participants was related to state encoding in the vmPFC (**?**). Such a "cognitive map" of the action space was argued to encode the current task state even when states are unobservable from the sensory environment.

2.3 The hippocampus: memory, space, and experience replay

The DMN midline has close functional links with the HC in the medial temporal lobe (**??**) —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 (**?????**). Its recursive anatomical architecture may be specifically designed to allow reconstructing entire episodes of experience from memory fragments. Indeed, hippocampal damage is not only associated with an impairment in re-experiencing the past (i.e., amnesia), but also imagination dedicated to one's own future and imagination of experiences more broadly (**?**).

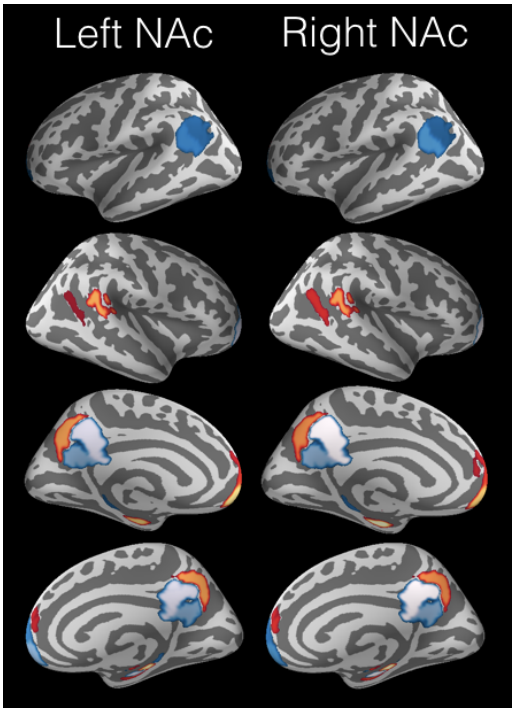

Fig 2. Morphological coupling between reward system and default mode network. Based on 9,932 human subjects from the UK Biobank, inter-individual differences in left NAc volume ($R^2 = 0.11 \pm 0.02$) and right NAc volume ($R^2 = 0.14 \pm 0.02$) could be predicted from volume in the DMN regions. These out-of-sample generalization performances were obtained from support vector regression applied to normalized region volumes in the DMN in a 10-fold cross-validation procedure. Consistent for the left and right reward system, NAc volume in a given subject is positively coupled with the vmPFC and HC. The colors are indicative of the (red = positive, blue = negative) and size (the lighter the higher) of the regression coefficients. Code and data for reproduction and visualization: `www.github.com/banilo/to_be_added_later`.

Mental scenes created by neurological patients with HC lesion exposed a lack of spatial integrity, richness in detail, and overall coherence. 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 (**?**). Such neuronal patterns in the HC appear to directly indicate upcoming behavior, such as in planning navigational trajectories (**?**) and memory consolidation of choice relevance (**?**). Congruently, London taxi drivers, humans with high performance in spatial navigation, were shown to exhibit increased gray-matter volume in the HC (**?**).

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 (**?**). Specific spiking sequences, which were elicited by experimental task design, have been shown to be re-enacted spontaneously during quiet wakefulness and sleep (**??**). Moreover, neuronal spike sequences measured in hippocampal place cells of rats featured re-occurrence directly after experimental trials as well as directly before upcoming experimental trials (**?**). 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 also in the regions of the DMN. Moreover, in mice, invasively triggering spatial experience recall in the HC during sleep has been demonstrated to subsequently alter action choice during wakefulness (**?**). These HC-subserved mechanisms conceivably contribute to advanced cognitive processes that require re-experiencing or newly constructed mental scenarios, such as in recalling autobiographical memory episodes (**?**). Thus, the HC would orchestrate re-experience of environmental aspects for consolidations based on re-enactment and for integration into rich mental scene construction (**??**). As such, the HC may impact ongoing perception of and action on the environment (**??**).

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

The DMN emerges with its midline structures early in human development (**?**), while the right and left TPJs may become fully integrated into this major brain network only after birth. The TPJs are known to exhibit hemispheric differences based on microanatomical properties and gyrification patterns (**?**). Globally, neuroscientific investigations on hemispheric functional specialization have highlighted the right versus left cerebral hemisphere as dominant for attentional versus semantic functions (**????**).

The TPJ in the right-hemispheric DMN (RTPJ) has been shown to be closely related to multi-sensory prediction and prediction error signaling. It is central for action initiation during goal-directed psychological tasks and for sensorimotor behavior by integrating multi-sensory attention (**?**). Involvement of this DMN region was repeatedly reported in multi-step action execution (**?**), visuo-proprioceptive conflict (**?**), and detection of environmental changes across visual, auditory, or tactile stimulation (**?**). Direct electrical stimulation of the human RTPJ during neurosurgery was associated with altered perception and stimulus awareness (**?**). It was argued that the RTPJ encodes actions and ensuing outcomes, without necessarily relating those to value estimation (**???**). Additionally, 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 (**??**). Indeed, neurological patients with RTPJ damage have particular difficulties with multi-step actions (**?**). 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 (**?**), a capacity believed to be enabled by the DMN. The RTPJ might hence be an important relay that shifts away from the internally directed baseline processes to, instead, deal with unexpected environmental stimuli and events.

The TPJ in the left-hemispheric DMN (LTPJ), in turn, has a close relationship to Wernicke's area involved in semantic processes, such as in spoken and written language. Neurological patients with damage in Wernicke's area 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 mediates access to and integration of world knowledge, such as required during action considerations (**??**). Consistent with this view, LTPJ damage in humans also entails problems in recognizing others' pantomimed action towards objects without obvious relation to processing explicit language content (**?**). 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 (**?**). Further, episodic memory recall and mental imagery 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 future events. The LTPJ may hence facilitate the automated environmental predictions by incorporating experience-derived building blocks of world regularities into ongoing action, planning, and problem solving.

3 Reinforcement learning control: a process model for DMN function

We now 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 expected 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 learning optimal behavioral through interactions with an *environment*, the goal being to maximize some notion of *cumulative reward*.

The optimal behavior typical takes the future into account, as some rewards could be *delayed*. Through repeated interactions with the environment, the agent learns to reach goals and optimize reward signals in an iterative trial-and-error fashion (Fig. **??**). At a given moment, each taken *action* a triggers a change in the *state* of the environment $s \rightarrow s'$, accompanied by environmental feedback signals as *reward* $r = r(s, a, s')$ collected by the agent. If the collected reward outcome yields a negative value it can be more naturally interpreted as *punishment*. In this view, the environment is partially controlled by the action of the agent and the reward can be thought of as satisfaction —or aversion —accompanying the execution of a particular action.

The environment is generally taken as *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. (**?**). 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 a functional account of the DMN 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 (**?**). In short, DMN implication in a diversity of particularly advanced cognitive processes can be parsimoniously explained as probabilistic mental scene of experience simulations coupled with prediction error minimization to calibrate action trajectories for reward outcome maximization at different time scales. Such a purposeful optimization objective may be solved by a stochastic approximation based on a brain implementation of MCMC sampling (**?**).

3.1 Markov decision processes

Model-free RL has had great success in many real-world problems, including robotics (**??**), super-human performance in complex video games (**?**), and strategic board games like the breakthrough results upon recently on the game of Go (**?**), considered to be a golden benchmark problem in artificial intelligence. We emphasize that the brain in general, and the DMN in particular, is a physical system governed by the laws of physics and can be formally described by Markov processes at a sufficiently coarse scale. It has indeed been previously proposed (**?**) that any system obeying the laws of classical physics can be accurately modeled as a Markov process as long as the time step is sufficiently short. The process has *memory* if the next state depends not only on the current state but also on a finite number of past states. Rational probabilistic planning can be reformulated as a standard memoryless Markov process by simply expanding the definition of the state s to include experience episodes of the past.

In artificial intelligence and machine learning, a popular computational model for multi-step decision processes in such an environment are *Markov decision processes* (MDPs) (**?**). 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., is likely to be different for another agent in the same state), by maintaining probability distributions over actions 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 active learning.

Model-free RL can be plausibly realized in the human brain (**?**). Indeed, it has been proposed (**?**) that a core property of human intelligence underlie improvement of expected utility outcomes as a strategy for action choice in uncertain environments, a situation perfectly captured by the formalism of MDPs. It has also long been proposed (**?**) that there is a rather direct mapping of model-free RL learning algorithms onto aspects of the brain. The neurotransmitter dopamine could serve as a "teaching signal" to better estimate value associations and action policies by controlling 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 can be assumptions about the physical laws governing the agent's environment or constraints on the state space and transitions between states.

In our adopted model-free RL framework, 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.

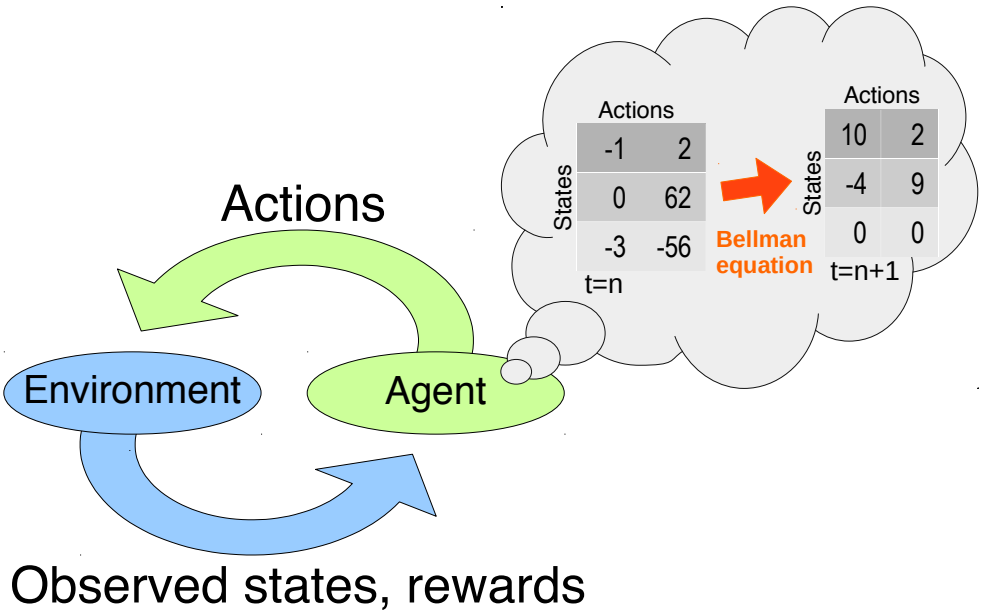

Fig 3. Reinforcement learning in a nutshell. Given the current state of the environment, the agent takes an action by following the policy matrix as updated by the Bellman equation. The agent receives a consequential reward and observes the next state. The process goes on until interrupted or a goal state is reached.

- etc.

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

- $\mathcal{S}$ *is the set of states, such as* $\mathcal{S} = \{happy, sad\}$.
- $\mathcal{A}$ *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* $s \to s'$.
- $p : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$, $(s, a, s') \mapsto p(s'|s, a)$, *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.*

Why MDPs ? At this point, it is righteous to question the applicability of MPDs as a model for something as complex as human behavior. Indeed a critique could argue that Markov chains are state-machines without memory. These may describe insects, or other behaviorally primitive beings like ATM machines, insects, etc., but won't be sufficient to describe human beings, as for example, human behavior is heavily influenced not only by the current state but by past states as well (memory).

Answer:

- First, let us emphasize that the brain in general, and the DMN in particular, is a physical system governed by the laws of physics and can be formally described by Markov processes at a sufficiently coarse scale. It has indeed been previously proposed (**?**) that any system obeying the laws of classical physics can be accurately modeled as a Markov process as long as the time step is sufficiently short.
- Rational probabilistic planning can be reformulated as a standard memory Markov process by simply expanding the definition of the state s to include experience episodes

of the past. This adds the capacity for memory, since then, the next state depends not only on the current state but also on a finite number of past states. This is precisely what partially observable MDPs do (**??**).

- Also, it should be noted that MDPs have been successfully used as neural model for animal behavior. One can mention (**?**) (random dots), (**?**) (mal-adaptive decision-making in a rat version of a gambling task) just to name a few.
- Successes in AI/ML:
  - Robotics: (**??**)
  - Games and other: (**???**)
- Algorithmic trading. In financial trading, the main actions which the agent (human or computer acting on behalf of a human operator) must execute strategically are: sell (sell a given quantity of a commodity), buy (by a given quantity of a commodity), do nothing. According to how the market responds, a definite gain / loss is recorded as a consequence of the executed action, perhaps with some delay. Over the years, research in algorithmic trading has identified MDPs as a powerful framework for modelling this behavioral process (**??????**), for which dynamic-programming based algorithms can be employed to produce numerical solutions. (**?**) is an excellent reference with the mathematical details on how MDPs work in finance.

3.1.1 Accumulated rewards and policies

The behavior of the agent is governed by a *policy*, which maps states of the world to probability distributions over actions. Starting at time $t = 0$, following a policy π generates a trajectory of action choices as follows:

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

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 accumulation 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[1] measures the cumulative reward of following an action policy π. Note that value buffering may be realized in the vmPFC. This DMN region has direct connections to to the NAc, known to be involved in reward evaluation.

The goal of the RL agent is then to update this action policy in order to maximize G^π on average (cf. below). In (**??**), the definition of cumulative reward G^π, the constant γ ($0 \leq \gamma < 1$) is the *reward discount factor*, viewed to be characteristic for a certain agent. On

[1] Random as it depends both on the environment's dynamic and the policy π being played (which can be stochastic).

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 long-term goal 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 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 longer-term future. More specifically, rewards that are not expected to come within $\tau := 1/(1-\gamma)$ time steps from the present point are disregarded. This reduces 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 farsighted future, it is important to appreciate that a stochastic policy estimation is more advantageous in many RL settings.

3.2 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 number

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

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 (**??**) provides the subjective utility of executing a specific action. It thus answers the question "What is the expected utility of taking action a in this situation ?". $Q^{\pi}(s,a)$ offers a formalization of optimal behavior that may well capture processing aspects of the DMN in human agents.

3.2.1 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,

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

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

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

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

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

The Bellman equation (**??**) is a temporal consistency equation which provides a dynamic decomposition of optimal behavior by dividing the Q-value function into the immediate reward and the discounted rewards of the upcoming states. The optimal Q-value operator Q^* is a fixed point for this equation. As a consequence of this decomposition, the complicated dynamic programming problem (**??**) 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 (**??**). 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 (**??**) 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.

Box 2: Neural correlates of the Bellman equation in the DMN

Relating decomposition of consecutive action choices by the Bellman equation to neuroscience, specific neural activity in the dorsal prefrontal cortex (BA9) was linked to processing "goal-tree sequences" in human neuroimaging experiments (**??**). 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 neuroimaging experiments (**?**) corroborated the prefrontal DMN to subserve "processes related to the management and monitoring of sub-goals while maintaining information in working memory". Moreover, neurological patients with lesions in this DMN region were reported to be impaired in aspects of realizing "multiple sub-goal scheduling" (**?**). Hence, the various advanced human mental 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.

3.2.2 Value approximation and the policy matrix

As already mentioned in the previous section, Q-learning optimizes over the class of deterministic policies of the form (**??**). State spaces may be extremely large, and tracking all possible states and actions may require prohibitively excessive computation and memory resources. 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 (**?**), or artificial neural-network models as demonstrated in recent seminal investigations (**??**) for playing complex games (atari, Go, etc.) at super-human levels. In the DMN, the dmPFC would 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 (**??**). The approximating Q-value function $\tilde{Q}(s, a|\theta)$ would tell the DMN the (expected) usefulness of taking an action a in state s. The DMN, and in particular its dmPFC node, could then contribute to the choice, at a given state s, of an action a which maximizes these approximate Q-values. This mapping from states to actions that is conventionally called *policy matrix* (**??**). Learning consists in starting from a given table and updating it during action choices, which take the agent to different table entries.

3.2.3 Self-training and the loss function

Successful learning in brains and computer algorithms is not possible without a defined learning goal —the *loss function*. The action a chosen in state s according to the policy matrix defined in (**??**) 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 be subserved by 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. This warrants a learning process and definition of a loss function. The Bellman equation (**??**) provides a way to obtain such a loss function (**??**) 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 (**??**), namely

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

with the prediction 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 occurred during specific prior actions and environmental states. The HC is a strong candidate to contribute to such a mechanism because neuroscience experiments have repeatedly indicated in rats, mice, cats, rabbits, songbirds, and monkeys (**????**).

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 loss function

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

where y_k is defined in (**??**). A recently proposed, practically successful alternative approach (**??**) is to learn this representation using an artificial deep neural-network model, leading to the so-called *deep Q-learning* family of methods which This 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 artifical 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 (**??**).

3.2.4 Optimal control via stochastic gradient descent in the DMN

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 true gradient is estimated from batches of training samples, which, in our case, 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 next 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 a standard result 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 (**??**).

3.2.5 Does the hippocampus subserve MCMC sampling?

In RL, MCMC simulation is a common means to update the agent's belief state (**?**). MCMC simulation provides a simple method for evaluating the value of a state. They provide 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

an RL agent without stochastic resampling procedures. Such methods have scaling 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 (**???**). These simulations of experience batches would be used to update the value function, without ever looking inside the black box describing the model's dynamics (**?**). This would be a simple control algorithm by evaluating all legal actions and selecting the action with highest expected cumulative rewards. In MDPs, MCMC 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 (**?**).

3.3 Putting everything together

The DMN is today known to consistently increase in neural activity when humans engage in cognitive processes that are detached from the current sensory environment. Additionally, this network 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 unimodal sensory cortices at the bottom (**??**). Its putative involvement in thinking about the past, hypothetical experiences, and the future appears to tie in with the implicit computation of action and state cascades as a function of what happened in the past. A policy matrix encapsulates the repertoire of possible actions on the world given a current state. The policy matrix encodes the probabilites of choosing actions to be executed in a certain situation. The DMN may subserve constant exploration of possible action trajectories and nested estimaton of their cumulative reward outcomes. Implicit computation of future choices provides an explanation for the evolutionary emergence and practical usefulness of mind-wandering at day-time and dreams during sleep in humans.

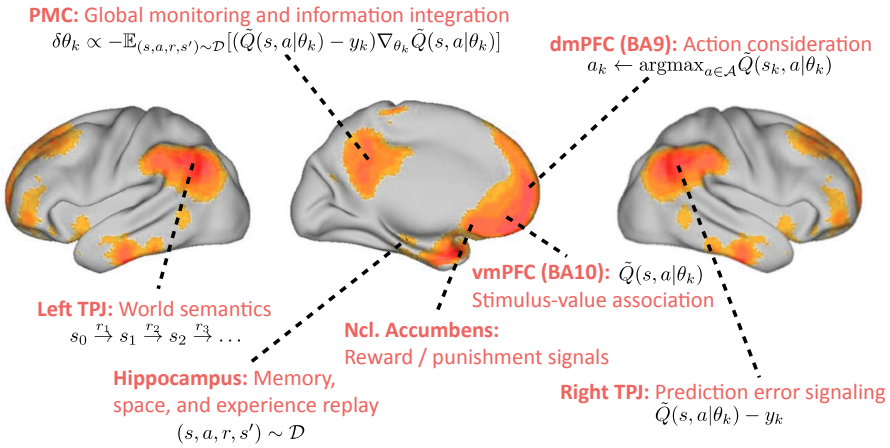

Fig 4. Default mode network: neurobiological implementation of reinforcement learning. Overview of how the constituent nodes of the DMN (refer to section **??**) may map onto computational components necessary for an RL agent.

The HC may contribute to generation of perturbed action-transition-state-reward samples as batches of pseudo-experience (i.e., imagined, hypothesized, and recalled mental scenarios). The small variations in these experience samplings allow searching a larger space of model parameters and possible experiences. Taken to its extreme, stochastic recombination of experience building blocks can further optimize the behavior of the RL agent by model learning from scenarios in the environment that the agent might only very rarely or never encounter. 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 situation. In the absence of environmental input and feedback (e.g., mind-wandering or sleep), mental scene construction allows for pseudo-experiencing possible future scenarios and action outcomes. Our formal account of DMN function thus acknowledges the unavoidable stochasticity of computation in neural systems (**?**).

From the perspective of a model-free RL agent, *inference* in the human brain 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.

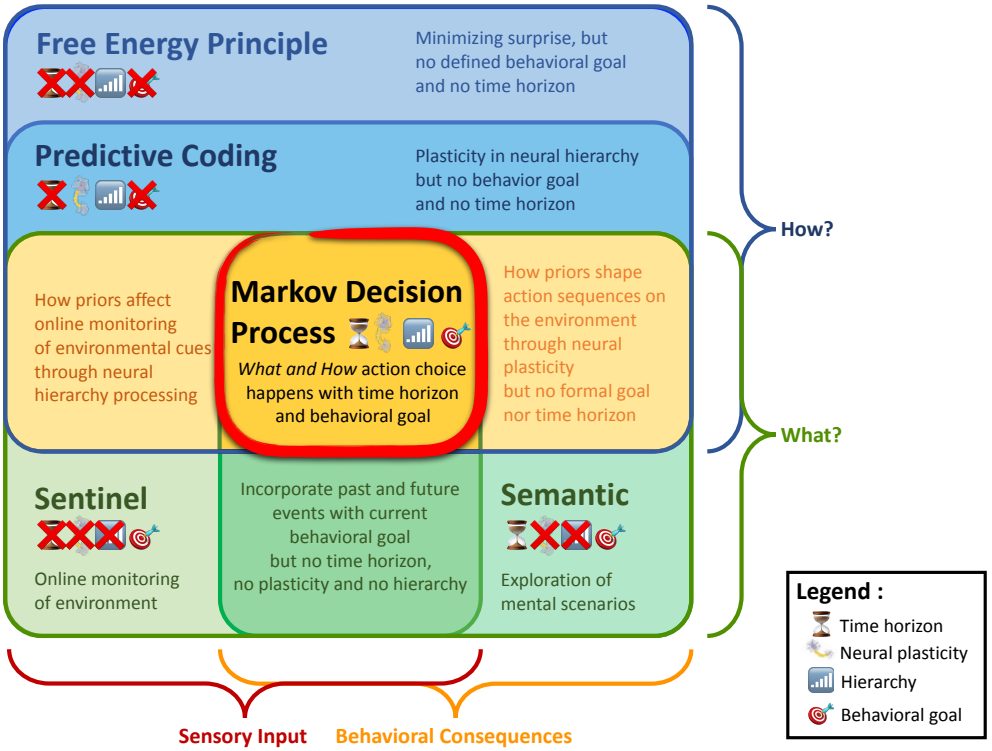

Fig 5. Situating Markov Decision Processes among other accounts of default mode function This Venn diagram summarizes the relationship between four existing explanations of the functional role of the DMN and our present account. Viewing empirical findings in the DMN from the MDP viewpoint incorporates 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.

4 Relation to existing accounts

4.1 Predictive coding hypothesis

Predictive coding mechanisms (**??**) are a frequently evoked idea in the context of default mode function (**?**). 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 expectations computed based on stored prior information. At each stage of neural processing, an internally generated prediction 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 (i.e., adapting model parameters) 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 (**??**) and the intimate relation of motor action to sensory expectations (**??**). Contextual integration of sensorimotor perception-action cycles may be maintained by top-down modulation using a-priori 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 stimulation and performing action on the environment. 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" (**?**) with the biological purpose to generalize from the past to new experiences. Neuroanatomically, axonal back projections indeed outnumber by far the axonal input projections existing in the monkey and probably also human brain (**?**). These many and diverse modulatory influences 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 (**?**).

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 predictive coding ideas from

cognitive neuroscience. MDPs have the advantage of exposing an explicit mechanism for 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.

4.2 Semantic hypothesis

Another 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 (**???**). Analogies might naturally tie incoming new sensory stimuli to explicit world knowledge (i.e., semantics) (**?**). 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 of non-existing scenarios. As such, semantic knowledge would be a prerequisite for optimizing behavior by constantly simulating possible future scenarios (**??**). Such cognitive processes can afford the internal construction and elaboration of necessary information that is not presented in the surrounding sensory environment by recombining building blocks of concept knowledge and episodic memories (**?**). 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 (**??**). 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 (**?**). Such scene construction processes could contribute to interpreting the present and foretelling the future. Further, mental scene construction has been proposed to imply a distinction between engagement in the sensory environment and internally generated mind-wandering (**?**). 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 mentally explored. In both accounts, there is also no reason to assume that predictions of various levels of complexity, abstraction, timescale, and purpose rely on mechanisms that are qualitatively different. This concurs with DMN activity increases across time, space, and content domains demonstrated in many neuroimaging studies (**????**). Further, the semantic hypothesis and MDP account provide explanations why HC damage does not only impair recalling memories, but also hypothetical and future thinking (**?**). While both semantic hypothesis and our formal account propose memory-based internally generated information for probabilistic mental models 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 offer a mechanistic explanation *how* explicit world knowledge and semantic analogies thereof lead to prediction of future actions and states. The semantic hypothesis does also not 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 MCMC approximation achieves better fine-tuning of the action policies and estimation of expected reward outcomes. That is, the purposeful stochasticity of policy and value estimation in MDPs provides a candidate explanation for why humans 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 model, both of which are integral parts of the MDP interpretation of DMN function.

4.3 Sentinel hypothesis

The DMN regions have been associated with processing the experienced or expected relevance of environment cues (**?**). Processing self-relevant information was perhaps the first cognitive account that was proposed for DMN function (**??**). 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 (**??**). 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. 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. However, 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 become mentally 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 (**??**). Both environment-detached and environment-engaged cognitive processes may entail DMN recruitment if real or imagined experience is processed, manipulated, and used for predictions. During tasks, the policy and value estimates may be updated to optimize especially short-term action. At 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 stimuli especially important for humans are frequently of social nature. This is not surprising given that the complexity of the social systems is likely to be a human-defining property (**?**). 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 (**?**). Indeed, social topics amounted to roughly two thirds of human everyday communication (**?**), while mind-wandering at daytime and dreams during sleep are rich in stories about people and the complex relationships between them. In line with this, the DMN was argued to be specialized in continuous processing of social information as a physiological baseline of human brain function (**?**). This view was later challenged by observing analogues of the DMN in monkeys (**?**), cats (**?**), and rats (**?**), three species with social-cognitive capacities that are supposedly less advanced than in humans.

Further, 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 (**?**). Neurocomputational 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 (**?**). The DMN is moreover largely overlapping with neural networks associated with higher level social cognition (**?**). For instance, the vmPFC, PMC, and RTPJ together play a key role in bridging the gap between self and other by integrating low-level embodied processes within higher level inference-based mentalizing (**?**).

Rather than functional specificity for processing social information, 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 (**???**). 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.

4.4 The free-energy principle and active inference

According to theories of the *free-energy principle* (FEP) and *active inference* (**??**) (see also (**?**)), the brain corresponds to a biomechanical reasoning engine. It 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 sum 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* functional defined by $\mathcal{H}(p) := -\sum_{\mathbf{z}} p(\mathbf{z}) \log(p(\mathbf{z})$, while $\mathrm{KL}(.\|.)$ 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 how different two probability distributions are. In this framework, the goal of the agent is then to iteratively refine the generative model p_G and the recognition model p_R (i.e two Bayesian belief nets) so as to minimize the free energy $F_G^R(\mathbf{x})$ over sensory input $\mathbf{x}$.

Importantly, one notes that $F_G^R(x)$ is low just in case:

- $p_R(\mathbf{z}|\mathbf{x})$ puts a lot of mass on configurations $(\mathbf{z}, \mathbf{x})$ which are p_G-likely, **and**
- $p_R(\mathbf{z}|\mathbf{x})$ 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 $\mathbf{x}$.

Main criticisms. Despite its popularity, criticism against the FEP has arisen over the years, some of which is outlined in the following. The main algorithm for minimizing free energy $F_G^R(\mathbf{x})$ is the *wake-sleep algorithm* (**?**). As these P. Dayan and G. Hinton noted, a crucial drawback of the wake-sleep algorithm (and therefore of theories like the FEP (**?**) based on it) is that it involves a pair of forward (generation) and backward (recognition) models p_G and p_R respectively that together do not correspond to optimization of (a bound of) the marginal likelihood, because KL divergence is not symmetric in its arguments. The brain may therefore be unlikely to implement a variant of the wake-sleep algorithm. The recent theory of *variational auto-encoders* (VAEs) (**?**) might provide an efficient alternative to the wake-sleep algorithm. VAEs 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, VAEs can be efficiently trained via back-propagation of prediction errors.

On another front, since theories based on the FEP (**??**) conceptualize ongoing behavior in an organism to be geared towards the surprise-minimizing goal, an organism entering a dark room (Fig. **??**) would strive to remain in this location because its sensory inputs are perfectly predictable given the environmental state (**?**). However, such a behavioral tendency is seldom observed in animals or humans in the real world: In a dark room, intelligent agents would search for light sources to *explore* its surroundings or leave it. Defenders of the FEP have retorted by advancing the "full package" (**?**): FEP is proposed to be multi-scale and there would be a meta-scale at which the organism would be *surprised* by such a *lack of surprise*. According to this argument, a dark room would paradoxically correspond to a state of particularly high surprise.

Driven by surprise-minimization objective, the FEP agent would eventually 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 **??**) would serve to sample memories or fantasies of alternative situations with reward structure. Such artificially generated *internal* sensory input subserved by the DMN can entice the organism to explore the room, for instance by looking for and using the light switch or simply finding the room exit.

Fig 6. The dark room experiment. An intelligent agent situated in a light-deprived closed space can be used as a thought experiment for the complete absence of external sensory input.

Links between FEP and RL. We note that FEP and active inference can be reframed in terms of our model-free RL framework. This becomes 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 corresponds to using free-energy as a Q-value approximator for the MDP, i.e

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

Such a surprise-guided reinforcement learning scheme has previously been advocated under the equivalent framework of energy-based reinforcement-learning[2] (**??**) and information compression (**??**). Nevertheless, minimization of surprise quantities alone may be insufficient to explain the diversity of behaviors observed in humans and other intelligent animals.

5 Conclusion

Which brain function could be important enough for the existence and survival of the human species to justify constantly high energy costs? MDPs motivate an attractive formal account of how the human association cortex might implement multi-sensory representation and control of the environment to optimize the organism's interaction with the world. This idealized process model explains a number of previous experimental observations in the DMN by simple but non-trivial mechanisms. From the view of a Markovian sequential decision process, human behavior unfolds by integrating action outcomes from policy matrix (mapping from states to actions which maximize expected cumulative reward; see subsection **??**) and extrapolation to upcoming events for guiding action choice in the present context. MDPs also provide a mathematical formalism how opportunity in the environment can be recursively exploited when confronted with challenging decisions. This functional 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-related prospection. Improvement of the internal world representation by injecting stochasticity into the recall of past actions and the estimation of action outcomes may explain why highly accurate memories have been disfavored in human evolution and why human creativity might be adaptive.

It is an important feature of the proposed artificial intelligence perspective on DMN biology that it is practically computable and yields falsifiable neuroscientific hypotheses. Neuroscience

[2]These methods assume that the posterior distribution of the hidden states $\mathbf{z}$ given the visible states $\mathbf{x}$ factorizes, leading to so-called RBMs –*Restricted Boltzmann Machines*. One notes, that even with these simplifications, such a model can be very hard to train.

experiments could be designed that 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 the DMN. Ultimately, DMN activity may instantiate a holistic integration ranging from real experience over purposeful dreams to anticipated futures for continued refinement of the organism's fate. Indeed, a major hurdle in explaining DMN activity from brain-imaging studies has been its very similar engaged across time scales: thinking about the past (e.g., autobiographical memory), thinking about hypothetical presents (e.g., dreams during sleep or daytime mind-wandering), and thinking about future scenarios (e.g., prospection or delay discounting). The MDP account of DMN function naturally integrates this a-priori diverging mental processes into a common framework.

Remark 1. *Given a system $(\mathcal{S}, \mathcal{A}, r, p, \mu)$ satisfying all the axioms for an MDP except the Markov property, we can always transform it into an MDP by considering a compounded version of the states*

$$S_t \leftarrow (S_0, A_0, R_0, S_1, A_1, S_2 \ldots, S_t, A_{t-1}, R_{t-1}, S_t)$$

made of the system's history up to an including time t.

A More on RL in the brain

A.1 Does hippocampal replay equal inverse reinforcement-learning?

Given the trace $s_0, a_0, s_1, \ldots,$ of an optimal agent's strategy π^* in an MDP (called a *teacher's demonstration*), can we figure out what is the (instantaneous) reward function $r : \mathcal{S} \times A \to \mathbb{R}$ that the agent is optimizing over a prescribed class of reward functions (e.g., linear rewards $r(s, a) \equiv \theta^T \phi(s, a)$)? For instance, given traces of motor actions from an adult to grab a cup on a table, can an observing child figure out what "langragian" functional is being minimized by the former? How? Can they reproduce this optimal behavior? Such questions are of course pertinent to our decision-making theory for the DMN. In the general artificial case, the problem has been extensively studied and partially solved by (**?**). They are also been rigorously studied in general optimal control literature under the name "inverse optimal control", but in model-based certain (where the physical dynamics are known, etc.)...

IRL is is suited for problems in which it's hard to define what the reward function could be (e.g., car-driving, drone maneuvering, etc.) ...

B Free-energy principles!

The so-called free-energy principle in its present form (including notions like "generative density", "recognition density", etc.) can be traced back to works of Dayan & Hinton (**?**) in which they introduced the so-called *Helmholtz machine*, a hierarchical factorial directional deep belief-net (DBN). In this subsection, we will develop from first-principles, the bare-bones minimalistic ideas needed to build a free-energy principle for general decision-making. This ideas were first developed by Hinton et al. in the early 90s in building their Helmholtz machine. Theories like Friston's free-energy principle and active-inference will then emerge as particular instances of this general framework, with particular design choises. For instance, the Friston theory axiomatizes that the brain uses a (problematic, as it implicitly assumes that posterior of each hidden unit is factorial) wake-sleep algorithm to train the underlying Helmholtz machine, etc.

B.1 Helmholtz free-energy and the generative model

symbol	description
$\langle X\rangle_p$	Expectation (a.k.a average, a.k.a mean value) of the random quantity X w.r.t to the probability density p, formally defined by $\langle E\rangle_p := \sum_z p(z)X(z)$.
$\mathcal{H}(p)$	Information-theoretic entropy of a probability density p, formally defined by $\mathcal{H}(p) := -\sum_z p(z)\log(p(z))$, with the usual convention $0\log(0) := 0$.
$D_{KL}(q\|\|p)$	The Kullback-Leibler divergence between the probability densities q and p respectively, formally defined by $D_{KL}(q\|\|p) := \sum_z q(z)\log(q(z)/p(z))$.
$\mathbf{x}$	Observations. In Friston's free-energy principle this has a decomposition in to two terms: the brain's internal state b and sensory inputs s, i.e., $\mathbf{x} = (s, b)$.
$\mathbf{z}$	Hidden variables. This should be understood as the unobservable states of the external environment (to which the brain is trying to adapt by learning).
$p_G(.\|\mathbf{x})$	Generative density for ...
$p_R(.\|\mathbf{x})$	Recognition density for ... Does some kind of predictive coding (?).
$F_G(\mathbf{x})$	Helmholtz free-energy for a model p_G of generating the observation $\mathbf{x}$. This measures the surprise incured upon observing $\mathbf{x}$ generated by the model G.
$F_G^R(\mathbf{x})$	Variational Helmholtz free-energy from model G to R. Note that $F_G^G = F_G$.

Table 1. Table of notations.

Our starting point will be to build an approximation p_G for the true density p of the observations, so that this approximate density corresponds to the partition function of thermodynamic system. So,

$$\begin{aligned}
\text{generative surprise} &= -\log(p_G(\mathbf{x})) = -\log(p_G(\mathbf{x}))\times 1 = -\log(p_G(\mathbf{x}))\sum_{\mathbf{z}} p_G(\mathbf{z}|\mathbf{x}) \\
&= -\sum_{\mathbf{z}} p_G(\mathbf{z},\mathbf{x})\log(p_G(\mathbf{x})) = -\sum_{\mathbf{z}} p_G(\mathbf{z}|\mathbf{x})\log(p_G(\mathbf{z},\mathbf{x})/p_G(\mathbf{z}|\mathbf{x})) \\
&= \sum_{\mathbf{z}} p_G(\mathbf{z}|\mathbf{x})\log(p_G(\mathbf{z}|\mathbf{x})) - \sum_{\mathbf{z}} p_G(\mathbf{z}|\mathbf{x})\log(p_G(\mathbf{z},\mathbf{x})) \\
&= -\langle\log(p_G(.,\mathbf{x}))\rangle_{p_G(.|\mathbf{x})} - \mathcal{H}(p_G(.|\mathbf{x})) \\
&= \langle E_G(.,\mathbf{x})\rangle_{p_G(.|\mathbf{x})} - \mathcal{H}(p_G(.|\mathbf{x}))
\end{aligned} \tag{12}$$

where $E_G(\mathbf{z},\mathbf{x})$ is the energy at *macrostate* $\mathbf{z}$ of a fictive thermodynamic system defined by setting

$$E_G(\mathbf{z},\mathbf{x}) := -\log(p_G(\mathbf{z},\mathbf{x})), \tag{13}$$

The last quantity in (**??**) is nothing but *Helmholtz free-energy* (at unit temperature!), defined formally by

$$F_G(\mathbf{x}) := \langle E_G(.,\mathbf{x})\rangle_{p_G(.|\mathbf{x})} - \mathcal{H}(p_G(.|\mathbf{x})). \tag{14}$$

Thus,

Fact B.1. *Generative surprise and generative Helmholtz free-energy are different views on exactly the same object.*

The goal of the brain is then to optimize over the generative model G: to iteratively or analytically modify the generative density $p_G(.|\mathbf{x})$, so as to minimize surprise. It turns out that a direct attempt to attack this optimization problem by gradient descent on the free-energy $F_G(\mathbf{x})$ is furtile: the parameter update steps are not "very clean", and require rather

cumbersome and heavy computations. A workaround is then to introduce a second density $p_R(.|\mathbf{x})$ called a *recognition* density to work in tandem with the generative density $p_G(.|\mathbf{x})$, as a trick for doing approximate inference. The former dreams / fantacizes whilst the latter tries to generate sensations which match these dreams! This primal-dual idea, first proposed in Hinton et al. 1995, is at the heart of the general free-energy principle that we will introduce shortly.

B.2 Variational Helmholtz free-energy and the bottom-up recognition sub-model

In this subsection, we will present an insightful upper bound for the generative surprise (i.e., generate Helmholtz free-energy), called the *variational* (Helmholtz) free-energy. As an avant-gout of what is to come shorty, let's just note that the well-known *free-energy principle* is simply a workaround whereby the minimization surprise (intractable) is replaced with the minimization a carefully chosen upper bound thereof.

Invoking (**??**) and applying Bayes rule, we get the Gibbs distribution

$$p_G(\mathbf{z}|\mathbf{x}) = \frac{p_G(\mathbf{z}|\mathbf{x})}{p_G(\mathbf{x})} = \frac{\exp(-E_G(\mathbf{z},\mathbf{x}))}{Z_G(\mathbf{x})} = \frac{\exp(-E_G(\mathbf{z},\mathbf{x}))}{Z_G(\mathbf{x})}, \tag{15}$$

where $Z_G(\mathbf{x}) := \log(p_G(\mathbf{x})) = \sum_{\mathbf{z}'} \exp(-E_G(\mathbf{z}',\mathbf{x}))$, the normalizing *partition function* for the model **??**. Whence, for any macrostate $\mathbf{z}$, we have $p_G(\mathbf{x}) = Z_G(\mathbf{x}) = \exp(-E_G(\mathbf{z},\mathbf{x}))/p_G(\mathbf{z}|\mathbf{x})$, and so it holds that

$$F_G(\mathbf{x}) \overset{(??)}{=} -\log(p_G(\mathbf{x})) = -\log(Z_G(\mathbf{x})) = E_G(\mathbf{z},\mathbf{x}) + \log(p_G(\mathbf{z}|\mathbf{x})). \tag{16}$$

Now, in the above equation, the LHS only depends on the generative model G and the data point $\mathbf{x}$: it doesn't depend on the hidden variable $\mathbf{z}$, etc. So, taking expectations w.r.t an arbitrary density[3] $p_R(.|\mathbf{x})$ yields

$$\begin{aligned} F_G(\mathbf{x}) &= -\log(Z_G(\mathbf{x})) = \langle E_G(.,\mathbf{x})\rangle_{p_R(.|\mathbf{x})} + \sum_{\mathbf{z}} p_R(\mathbf{z}|\mathbf{x}) \log(p_G(\mathbf{z}|\mathbf{x})) \\ &= \langle E_G(.,\mathbf{x})\rangle_{p_R(.|\mathbf{x})} - \mathcal{H}(p_R(.|\mathbf{x})) - \sum_{\mathbf{z}} p_R(\mathbf{z}|\mathbf{x}) \log(p_R(\mathbf{z}|\mathbf{x})/p_G(\mathbf{z}|\mathbf{x})) \\ &= F_G^R(\mathbf{x}) - D_{KL}(p_R(.|\mathbf{x})||p_G(.|\mathbf{x})), \end{aligned} \tag{17}$$

where $F_G^R(\mathbf{x})$ is the *variational* Helmholtz free-energy from R to G defined by

$$F_G^R(\mathbf{x}) := \langle E_G(.,\mathbf{x})\rangle_{p_R(.|\mathbf{x})} - \mathcal{H}(p_R(.|\mathbf{x})) \tag{18}$$

and $D_{KL}(p_R(.|\mathbf{x})||p_G(.|\mathbf{x}))$ is the Kullback-Leibler divergence between the $p_R(.|\mathbf{x})$ and the generative density $p_G(.|\mathbf{x})$. Note that $F_G^G = F_G$.

B.3 A general free-energy principle

We can resume the situation as follows[4]:

$$\begin{aligned} \text{generative surprise } &:= -\log(p_G(\mathbf{x})) = F_G(\mathbf{x}) \\ &= \underbrace{F_G^R(\mathbf{x})}_{\text{accuracy}} - \underbrace{D_{KL}(p_R(.|\mathbf{x})||p_G(.|\mathbf{x}))}_{\text{complexity}} \\ &\leq F_G^R(\mathbf{x}), \text{ with equality if } p_R(\mathbf{z}|\mathbf{x}) = p_G(\mathbf{z}|\mathbf{x}) \text{ for all } \mathbf{z} \end{aligned} \tag{19}$$

[3] conditioning in $p_R(.|\mathbf{x})$ is because this density is selected from a world in which the sensory inputs and internal brain state vector $\mathbf{x}$ is assumed already observed.

[4] Where we have used the fact that KL divergence is always nonnegative.

B.4 Helmholtz machines and the wake-sleep algorithm

Assumption: In both generative and recognition components of the network, there is conditional independence of neurons in the same layer, given the data (i.e., input from lower more primitive layers). Precisely

$$p_G(\mathbf{z}^{(l)}|\mathbf{x}) = \Pi_{k=1}^{h_l} p_G(\mathbf{z}_k^{(l)}|\mathbf{x}), \qquad p_R(\mathbf{z}^{(l)}|\mathbf{x}) = \Pi_{k=1}^{h_l} p_R(\mathbf{z}_k^{(l)}|\mathbf{x})$$

B.5 Friston's active-inference and agency

This is nothing but an application of the Dayan's wake-sleep algorithm for training a Helmholtz machine model of the brain...

The following critics can be made:

- As noted by Dayan et al. (*Variants of Helmholtz machines*), the inter-neuronal intra-layer independence assumption which is at the center of the HM becomes severely problematic as it is agnostic to the known organization of cortical layers...
- A drawback of the wake-sleep algorithm is that it requires a concurrent models (generative and recognitiion), which together do not correspond to optimization of (a bound of) the marginal likelihood (because of the incorrect KL used therein, etc.).
- Also, note that the wake-sleep algorithm doesn't do backprop! This is due to technical difficulty in getting derivatives of loss function w.r.t recognition weights $\mathbf{W}^R$).
- This difficulty was removed in the 2010s by (**?**), an other groups, via a "reparametrization trick".

B.6 Minimizing free-energy via backprop: variational auto-encoders

Here, we present a way to alleviate some conceptual and computational issues with the free-energy framework presented thus far, by using the recent *variational auto-encoder* (VAE) theory (**?**). Define the data-dependent auxiliary random function

$$f_{G,R}(.,\mathbf{x}) : \mathbf{z} \mapsto \log(p_G(\mathbf{z},\mathbf{x})) - \log(p_R(\mathbf{z}|\mathbf{x})). \tag{20}$$

Then we can rewrite the variational free-energy as

$$\begin{aligned} F_G^R(\mathbf{x}) &:= \langle E_G(.,\mathbf{x})\rangle_{p_R(.|\mathbf{x})} - \mathcal{H}(p_R(.|\mathbf{x})) = \langle E_G(.,\mathbf{x}) + \log(p_R(.|\mathbf{x}))\rangle_{p_R(.|\mathbf{x})} \\ &= \langle -\log(p_G(.,\mathbf{x})) + \log(p_R(.|\mathbf{x}))\rangle_{p_R(.|\mathbf{x})} \\ &= -\langle f_{G,R}\rangle_{p_R(.|\mathbf{x})} \approx -\frac{1}{M}\sum_{m=1}^{M} f_{G,R}(\mathbf{z}^{(m)}), \text{ with } \mathbf{z}^{(1)},\ldots,\mathbf{z}^{(M)} \sim p_R(.|\mathbf{x}), \text{ and } M \to \infty. \end{aligned}$$

Problem: How do we sample from the recognition density $p_R(.|\mathbf{x})$ in such a way that the sampling process is differentiable w.r.t the weights of the recognition network $\mathbf{W}^R$?

Solution: the reparametrization trick.

- Choose $\epsilon \sim p_{\text{noise}}$ (noise distribution, independent of $\mathbf{W}^R$!)
- Set $\mathbf{z} = g(\mathbf{W}^R, \mathbf{x}, \epsilon)$, where g is an appropriate class $\mathcal{C}^1$ function
  - results in a sample $\mathbf{z} \sim p_R(.|\mathbf{x})$, from the correct posterior

The mapping g should be taught of as a "blurring" function which produces noisy versions $\mathbf{z}$, called *sensations*, of the true world state $\mathbf{x}$. The result is a scheme for training DBNs via good-old backprop! Refer to Fig. vae.pdf. Some examples of the reparametrization trick for a number of choices of the posterior distribution are given in Tab. **??**.

Posterior	$p_R(.\|\mathbf{x})$	noise	$g(\mathbf{W}^R, \mathbf{x}, \epsilon)$	Also
Normal	$\mathcal{N}(\mu, \sigma)$	$\epsilon \sim \mathcal{N}(0,1)$	$\mu + \sigma \odot \epsilon$	Location-scale family: Laplace, Elliptical, Students t, Logistic, Uniform, Triangular, ...
Exponential	$\exp(\lambda)$	$\epsilon \sim \mathcal{U}([0,1))$	$-\log(1-\epsilon)/\lambda$	Invertible CDF: Cauchy, Logistic, Rayleigh, Pareta, Weibull, Reciprocal, Gompert, Gumbel, Erlan, ...
Other	$\log \mathcal{N}(\mu, \sigma)$	$\epsilon \sim \mathcal{N}(0,1)$	$\exp(\mu + \sigma \odot \epsilon)$	Gamma, Dirichlet, Beta, Chi-squared, F, ...

Table 2. Reparametrization trick (**?**) for a variety of models.

For fixed $p_G(.|\mathbf{x})$, the optimal choice for $p_R(.|\mathbf{x})$ is given analytically by

$$p_R(\mathbf{z}|\mathbf{x}) \propto p_G(\mathbf{z}|\mathbf{x}) \exp(-\Delta E_{G \to R}(\mathbf{z}, \mathbf{x})). \tag{21}$$

B.7 A thermodynamic model for bounded rationality, aka robust optimality

- Recall that an agent is said to have *bounded rationality* if they must take into account the cost of finding solutions to problems, and not just the utility of the final state.
- For example, consider an agent that must operate under a limited lifetime and/or computation cost.
- This is in contrast to agents with *unbounded rationality* considered in classical game theory.

The material presented is a revisit of (**?**). See also (**?**)

Utility functions and conjugate pairs. Let $(\Omega, \mathcal{F}, P)$ be a probablity space. A function $U : \mathcal{F} \to \mathbb{R}$ is said to be a *utility function* for this space if the conditional utility $U(A|B) := U(A \cap B) - U(B)$ has the following propertites:

- additivity: $U(A_1 \cap A_2|B) = U(A_1|B) + U(A_2|B)$, for all events $A_1, A_2, B \in \mathcal{F}$.
- statistic: there exists a function $f_U : \mathbb{R}_+ \to \mathbb{R}$ such that $U(A|B) = f(P(A|B))$, for all events $A, B \in \mathcal{F}$.
- monotonicity: f_U is strictly increasing.

Theorem 1. *The only functions $f : \mathbb{R}_+ \to \mathbb{R}$ which is such that $U(A|B) \equiv f(P(A|B))$ any probability space $(\Omega, \mathcal{F}, P)$ and utility function U thereupon are of the form*

$$f = \alpha \log(.), \tag{22}$$

where $\alpha > 0$.

XXX: equation (??) above looks like Boltzmann's formula (on his gravestone...)!

Such U and P are said to form a *conjugate pair* at temperature α.

Example. Given a utility function U on a probability space $(\Omega, \mathcal{F}, *)$, the *Gibbs measure* at temperature $\alpha > 0$ and energy levels $(-U(\omega))_{\omega \in \Omega}$ is defined to be the probability measure (on thesame measurable space)

$$P(\omega) = \frac{1}{Z_U(\alpha)} \exp\left(\frac{1}{\alpha} U(\omega)\right), \ \forall \omega \in \Omega, \tag{23}$$

where

$$Z_U(\alpha) := \sum_{\omega\in\Omega} \exp\left(\frac{1}{\alpha}U(\omega)\right) \tag{24}$$

is a normalization constant called the *partition function* of U. It's not hard to see that U and the P above form a conjugate pair.

Free-utility functional. Let (U, P) be a conjugate pair at temperature $\alpha > 0$ on a measurable space $(\Omega, \mathcal{F})$. Given another probability measure P' on the same space, define it's *free utility* as

$$J(P'|U, P) = \langle U\rangle_{P'} + \alpha\mathcal{H}(P'), \tag{25}$$

where

$$\mathcal{H}(P') := \langle \log(P')\rangle_{P'} := -\sum_{\omega\in\Omega} P'(\omega)\log(P'(\omega)) \tag{26}$$

is the *entropy* of P' (measured in the Naperian base $e \approx 2.73$). It's not difficult to establish the upper bound

$$J(P'|U) \leq J(P|U) = \sum_{\omega\in\Omega} U(\omega) =: U(\Omega). \tag{27}$$

In particular, if P is the Gibbs measure at temperature α corresponding to U, then the upper bound above reduces to the *log-partition function*

$$J(P'|U) \leq U(\Omega) = -\alpha\log(Z_U(\alpha)). \tag{28}$$

The free-energy / utility principle (of Friston ?). We are now in shape to introduce the notion of free-energy for model transitions, and a variational principle for optimizing it. Consider thus an initial system described by a conjugate pair $(U_{\text{ini}}, P_{\text{ini}})$ at temperature $\alpha > 0$. We want to transform this to a new model by adding constraints represented by the utility function ΔU. The resulting system has final utility $U_{\text{fin}} = U_{\text{ini}} + \Delta U$. The difference in free-utility is then

$$\Delta J_{(U_{\text{ini}},P_{\text{ini}})\to P_{\text{fin}}} := J_{\text{fin}} - J_{\text{ini}} = \underbrace{\langle \Delta U\rangle_{P_{\text{fin}}}}_{\textbf{accuracy}} - \underbrace{\alpha D_{\text{KL}}(P_{\text{ini}}\|P_{\text{fin}})}_{\textbf{complexity}}, \tag{29}$$

where

$$D_{\text{KL}}(P_{\text{fin}}\|P_{\text{ini}}) := \langle \log(\mathbf{P}_{\text{fin}}/\mathbf{P}_{\text{ini}})\rangle_{\mathbf{P}_{\text{fin}}} := \sum_{\omega\in\Omega} \mathbf{P}_{\text{fin}}(\omega)\log(\mathbf{P}_{\text{fin}}(\omega)/\mathbf{P}_{\text{ini}}(\omega)) \tag{30}$$

is the Kullback-Leibler divergence, and represents the information cost (measured in energy units) of changing the initial system. In the above formula, we've extensively used the fact that $(U_{\text{ini}}, P_{\text{ini}})$ is a congugate pair and so $U_{\text{ini}}(\omega) \equiv \alpha\log(P_{\text{ini}}(\omega))$ by virtue of (**??**). The two terms in (**??**) (accuracy or expected gain in utility, and the complexity of the transition) can be viewed as dertiminants of bounded rational decision-making. They formalize a trade-off between an expected utility ΔU (first term) and the information cost of transforming Pi into P_{fin} (second term). In this interpretation P_{ini} represents an initial choice probability or policy, which includes the special case of the uniform distribution where the decision-maker has initially no preferences between the different choices. The probability measure P_{fin} is the final choice probability that we are looking for since it considers the utility constraint U that we want to optimize. We can then formulate a variational principle for bounded rationality in the probabilities $P_{\text{fin}}(\omega)$

$$P^*_{\text{fin}} := \underset{P_{\text{fin}}}{\text{argmax}}\ \Delta J_{(U_{\text{ini}},P_{\text{ini}})\to P_{\text{fin}}} \tag{31}$$

By differentiating the RHS of (**??**) w.r.t P_{fin} and setting to zero, we obtain the closed-form solution

$$P^*_{\text{fin}}(\omega) \propto P_{\text{ini}}(\omega)\exp\left(\frac{1}{\alpha}\Delta U(\omega)\right). \tag{32}$$

Two limit cases are worth considering.

- **Low-temperature regime $\alpha \approx 0$:** Here $\Delta J_{(U_{\text{ini}},P_{\text{ini}})\to P_{\text{fin}}} \approx \langle \Delta U \rangle_{P'_{\text{fin}}}$, and so it's optimal to take
$$P^*_{\text{fin}}(\omega) \equiv \text{dirac}(\omega - \omega^*) = \begin{cases} 1, & \text{if } \omega = \omega^*, \\ 0, & \text{otherwise,} \end{cases}$$
where $\omega^* := \text{argmax}_{\omega\in\Omega} U(\omega)$. This corresponds to unbounded rational decision-making, in which the cost of transition / problem-solving is completely disregarded.
- **High-temperature regime $\alpha \to +\infty$:** In this limiting case, it's optimal to take $P^*_{\text{fin}}(\omega) \equiv P_{\text{ini}}(\omega)$, i.e the change is so costly that it's optimal to maintain the current choice probabilities.

To conclude this section, let's note that (**?**) show how their free-energy framework (on paths) links with the well-known Hamilton-Jacobi-Bellman optimal control framework. For example, one can re-derive the Linear Quadratic Gaussian (LQG) controller, which is a generalization of the LQR in (**??**)...

B.8 GANs and other likelihood-free methods

...