A Computational Account of Default Mode Function by Control Theory and Reinforcement Learning

Elvis Dohmatob, Guillaume Dumas, Danilo Bzdok

Abstract

The default mode network (DMN) is the brain manifestation of the human baseline mental activity. Many research streams agree on its likely role in evolutionarily adaptive envisioning of mental scenarios to predict the environment. The DMN would hence be dedicated to continuous autobiographical memory retrieval, generation of hypothetical outcomes, and reward contingency evaluation when letting the mind go. It explains its highest energy consumption in the brain and its intimate coupling with conscious awareness. This concept paper proposes a process model that describes *how* the DMN may actually implement continuous environmental assessment and prediction to guide action choices. DMN function is recast in mathematical terms from the perspective of reinforcement learning. We argue that our formal account of DMN function explains existing cognitive accounts as special cases and parsimoneously explains a variety of experimental findings in animals and humans. Formal models of the computational processes subserved in the DMN could offer access routes to the statistical regularities of human predictive behavior.

keywords: systems biology, mind wandering, cognitive science, artificial intelligence, reinforcement learning

Contents

List of Figures

List of Tables

1 Introduction

When left unperturbed, the human brain is not at rest. Rather, the brain continues to metabolize large quantities of oxygen and glucose energy to maintain potentially purposeful neuronal computation in the absence of a focused behavioral goal (Kenet et al., 2003; Fiser et al., 2004). This baseline energy demand is subject to surprisingly little modulations when mentally processing environmental challenges (Raichle et al., 2001). What has early been described as the "stream of consciousness" in psychology (James, 1890) found its neurobiological manifestation in the so-called "default mode network" (DMN) (Shulman et al., 1997). This networked collection of some of the highest regions in the human association cortex (Mesulam, 1998; Margulies et al., 2016) consistently increased in neural activity during unfocused everyday mind wandering (Raichle et al., 2001). Indeed, the human brain features "a hierarchy of brain systems with the DMN at the top and the salience and dorsal attention systems at intermediate levels, above thalamic and unimodal sensory cortex" (Carhart-Harris and Friston, 2010). In the beginning of the 21st century, brain imaging was hence the first technology to allow for the discovery of a unique brain network that subserves baseline mental activities (Bzdok and Eickhoff, 2015).

1.1 Default mode network: Higher-order control of the organism

The DMN appeared to be exclusive in task-induced deactivation during various psychological experiments. This processing of unknown information categories in the DMN has been argued to mediate an evolutionarily conserved function for the individual. This is all the more likely because the DMN contains the two biggest hotspots of energy consumption in the entire central nervous system (Raichle et al., 2001). DMN activity also persists to a substantial degree during sleep and under anesthesia (Buckner et al., 2008). Today, many authors believe that the DMN implements probabilistic estimation of past, hypothetical, and future events. This brain network might have emerged to continuously predict environmental events 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 areas (Weissman et al., 2006) and increased DMN activity was linked to more task-independent, yet sometimes useful thoughts (Mason et al., 2007; Seli et al., 2016). Understanding of the overarching DMN function is complicated by simultaneously controlling perception-action cycles and maintaining baseline contemplations across time, splace, and content domains at the interface of the external world and the self.

Consider an agent faced with the choice of the next action and being guided by rich reenactment of really-happened, hypothetically conceived, and predicted future events to perfect ongoing behavioral performance. A particularly attractive computational framework to describe, quantify, and predict autonomously acting systems like the brain can be the combination of control theory and reinforcement learning. It is known that, the more the external world is predictable, the more mental activity becomes detached from the actual sensory environment (Antrobus et al., 1966; Pope and Singer, 1978). Conversely, the more the currently executed task is unknown and unpracticed, the lower the tendency for stimulus-independent thoughts (Filler and Giambra, 1973; Teasdale et al., 1995). These "offline" processes may however contribute to optimized control of the organism. Informed by outcomes of performed action, the DMN dynamics are constantly adapted in feedback loops that are calibrated by prediction error.

A DMN framework incorporating RL can naturally embed human behavior into the tension between exploitative action with immediate gains and exploratory action with longer-term reward schedules.

1.2 Sketch of our main contributions

Our theory proposes that the implication of the DMN in a diversity of humans' most advanced cognitive processes can be parsimoniously recast as prediction error minimization based on pervasive probabilistic simulations, thus maximizing reward outcome at various time scales. Such a purposeful optimization objective may be solved by a stochastic approximation based on a brain implementation of Markov Chain Monte Carlo sampling (Tenenbaum et al., 2011). *Control* refers to the influences that an agent exerts when interacting with the environment to encourage preferred states. Even (necessarily imperfect) memories of past experience, random mind-wandering and dreams during sleep may provide meaningful building blocks to iteratively improve the predictive DMN machinery to optimize the behavior of the organism. In short, it has been proposed earlier that the human brain's energy budget is largely dedicated to "the development and maintenance of [a] probabilistic model of anticipated events" (Raichle and Gusnard, 2005). The idea has been invigorated by empirical evidence from neuroscientific experiments (Körding and Wolpert, 2004; Fiser et al., 2004). This paper proposes a formal model that satisfies this contention.

2 Known Neurobiological Properties of the Default Mode Network

We begin by a deconstruction the overall DMN function by the mechanistic relevance of each individual network node based on published findings from neuroscience experiments.

2.1 The posteriormedial cortex: Global monitoring and integration

The posteriormedial cortex (PMC) and the dorsomedial prefrontal cortex (dmPFC) within the DMN are responsible for the highest metabolic turn-over of glucose energy consumption in humans (Raichle et al., 2001). The PMC myelinates relatively late during postnatal development in monkeys (Goldman-Rakic, 1987), generally considered to be a sign of phylogenetic sophistication (Flechsig, 1920). Physiological and disturbed metabolic fluctuations in the human PMC have been repeatedly related to phenomena of changed conscious awareness, including anesthesia, sleep, and forms of coma (Cavanna and Trimble, 2006). The PMC has long been speculated to reflect constant computation of the environmental statistics and its internal representation as an "inner minds eye" (Cavanna and Trimble, 2006). For instance, Bálint's syndrome is a neurological disorder of conscious awareness that results from damage in the bilateral parietal cortex (Bálint et al., 1909). Patients are plagued by an inability to bind various individual features of the visual environment into an integrated whole (i.e., simultanagnosia) as well as inability to direct action towards currently unattended environmental objects (i.e., optic ataxia). This can be viewed as a high-level impairment in the gathering of 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 modalities with the amygdala involved in significance evaluation and the nucleus accumbens (NAc), involved in reward evaluation. Specifically, among all parts of the PMC, the human ventral posterior cingulate cortex was most connected to the laterobasal nuclei group of the amygdala (Bzdok et al., 2015). This amygdalar subregion has been proposed to continuously scan environmental input for biological significance assessment. Indeed, electrophysiological recordings in animals implicated the PMC in strategic selection (Pearson et al., 2009), risk assessment (McCoy and Platt, 2005), and outcome-contingent behavioral modulation (Hayden et al., 2009), while its retrosplenial portion was more specifically implicated in approach-avoidance behavior (Vann et al., 2009). Neuron spiking activity in the PMC allowed distinguishing whether a monkey would persue an exploratory or exploitative behavioral strategy in a complex food foraging task (Pearson et al., 2009). Single-cell recordings in the monkey PMC demonstrated this brain region's sensitivity to subjective target utility (McCoy and Platt, 2005) and integration across individual decision-making instances (Pearson et al., 2009).

This DMN node encoded the preference 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 factual value, based on an internal currency of value. In fact, direct stimulation of PMC neurons promoted exploratory action towards options with unsafe reward outcomes that were previously shunned. Graded changes in firing rates of PMC neurons indicated changing choices in upcoming trials and their neural patterns were distinct from spike firings that indicated choosing either option. Also in humans, the DMN has been shown to gather and integrate information over different paragraphs in an fMRI study on auditory narratives (Simony et al., 2016). The retrosplenial portion of the PMC can subserve action possibilities and evaluation of

reward contingencies by integrating these with information from memory and altered perspective frames. Regarding memory retrieval, retrosplenial lesions have been consistently associated with anterograde and retrograde memory impairments of various kinds of sensory information in rabbits and humans (Vann et al., 2009). Regarding perspective frames, this PMC subregion has been proposed to mediate between the organism's egocentric (i.e., focused on sensory input) and allocentric (i.e., focused on world knowledge) viewpoints in animals and humans (Epstein, 2008; Burgess, 2008; Valiquette and McNamara, 2007). The PMC may consequently monitor the subjective outcomes of possible decisions and integrates that information with memory, perspective frames, and reward schedules into higher-level strategies. Perceived value that differs across individuals updates statistical assessment of the environment to predict delayed reward opportunities in the future. In doing so, the PMC continuously adapts to changes in both the external environment and its internal representation that modulate strategic behavioral adjustment in volatile environments.

2.2 The prefrontal cortex: Stimulus-value assocation and action consideration

The dmPFC subserves predominantly ambiguous amodal processes across time, space, and content domains in sensory-independent top-down pathways. This DMN node has been described as a mental sketchpad (Goldman-Rakic et al., 1996), potentially implicated in de-novo generation and binding of meaning representations instructed by stored semantics and memories. The dmPFC may thus enable inference, representation, and assessment of one's own and other individuals' action and thoughts. Generally, patients with neurological lesions in the prefrontal cortex are known to adapt to novel situations and stimuli with difficulty (Stuss and Benson, 1986). Specifically, neural activity in human dmPFC reflected expectations about other peoples actions and errors thereof. dmPFC activity indeed explained the proficiency decline of inferring other peoples' thoughts in the elderly (Moran et al., 2012). Some dmPFC neurons in macaque monkeys exhibited a preference for processing others', rather than own, behavior with fine-grained adjustment of contextual circumstances (Yoshida et al., 2010). Also the topographically neighboring dorsal anterior cingulate cortex has been linked to computing values and efforts of persisting a behavioral plan versus switching the environmental context in several lesion studies (Kolling et al., 2016). Such highly abstract neural computations necessarily rely on the generation of probabilistic internal information drawing from episodic memory retrieval, generative construction processes, and explicit knowledge of the external world. According to computational neuroimaging experiments, the dmPFC activity preferentially models stimulus-value associations of possible actions that are not actually executed, whereas the vmPFC activity models value in the environment that leads performed actions (Nicolle et al., 2012).

The ventromedial prefrontal cortex (vmPFC) subserves less ambiguous subjective-value-related evaluative processes and reward-informed risk estimation of self-relevant environmental stimuli. This DMN node is more closely associated with orchestrating adapted behavior by bottom-up-driven processing of what matters now, probably drawing on model-based value representations (ODoherty et al., 2015). Quantative lesion findings across 344 human individuals confirmed a gross impairment in value-based decision making (Gläscher et al., 2012). The vmPFC is preferentially connected with limbic and reward-related areas. The vmPFC has been observed to have monosynaptical connections with the nucleus accumbens in axonal tracing studies in monkeys (Haber et al., 1995). Additionally, the gray-matter volume of the vmPFC and nucleus accumbens correlated with indices of value-guided behavior and social reward attitudes (Lebreton et al., 2009). NAc activity is thought to reflect reward prediction

signals from dopaminergic pathways (Schultz, 1998) that not only modulate motivated behavior towards basic survival needs but also subserve RL in humans more broadly (ODoherty et al., 2015). This is consistent with diffusion MRI (dMRI) tractography in humans and monkeys (Croxson et al., 2005) that quantified this brain regions to be substantially more likely connected to the vmPFC than dmPFC in both species. Two functional connectivity modalities in humans strongly connected the vmPFC with the nucleus accumbens, hippocampus (HC), and posteriomedial cortex. The vmPFC is often proposed to be involved in (external) emotional reactions and own (visceral) arousal. Real or imaginged bodily states could be mapped in the vmPFC as a bioregulatory disposition governing cognition and decision making (Damasio et al., 1996) In neuroeconomic studies of human decision making, the vmPFC consistently reflects an individuals subjective valuation (Behrens et al., 2008). This may be why performance within and across participants was related to state encoding in vmPFC. Such a *cognitive map* of action space was argued to encode the current task state even when states are unobservable from sensory input, which was shown to be critical for behavior. Additionally, independent whole-brain analyses from structural neuroimaging studies related the gray-matter volume GMV of the vmPFC, more consistently than any other brain region, to indices of social competence and social network complexity, among the most complicated decision that humans and monkeys take (Behrens et al., 2009).

2.3 The hippocampus: Memory, space, and experience replay

The HC is well known to be involved in memory and spatial navigation in animals and humans. Its highly recursive anatomical architecture may be specifically designed to allow reconstructing entire episodes of experience from memory fragments. While the HC in the medial temporal lobe system is traditionally believed to allow remembering the past, there is now increasing evidence for a role in constructing mental models in general (Zeidman and Maguire, 2016; Schacter et al., 2007; Gelbard-Sagiv et al., 2008). Indeed, hippocampal damage is not only associated with an impairment in reexperiencing the past (i.e., amnesia), but also thinking about ones own future and imagining new fictitious experiences more broadly (Hassabis et al., 2007). Mental scenes created by hippocampus-lesioned patients exposed a lack of spatial integrity, richness in detail, and overall coherence. Single-cell recordings in the animal hippocampus revealed some constantly active neuronal ensembles whose firing coincided with distinct locations in space while the animal navigated through its surroundings. London taxi drivers, individuals with high performance in spatial navigation, were shown to exhibit increased grey matter volume in the posterior hippocampus (Maguire et al., 2000). Indeed, place-specific neuronal populations in the hippocampus spike one after another when an animal is choosing between alternative paths (Johnson and Redish, 2007); and these neuronal patterns apear to also indicate upcoming behavior (Pfeiffer and Foster, 2013).

But encoding and generative reconstruction in the hippocampus extends beyond mere spatial knowledge of the environment. Based on large-scale recordings of hippocampal neuronal populations, complex spiking patterns can be followed across extended periods including their modification of input-free self-generated patterns after environmental events (Buzsáki, 2004). Specific spiking sequences, that were elicited in experimental task conditions, have been shown to be reenacted spontaneously during quiet wakefulness and sleep (Hartley et al., 2014; ONeill et al., 2010). Moreover, spike sequences measured in hippocampal place cells of rats featured reoccurrence directly after experimental trials as well as directly before upcoming experimental trials (Diba and Buzsáki, 2007). Such hippocampal ensemble burts 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 nodes of the DMN. These mechanisms of the hippocampus probably make important contributions to the recollection of

autobiographical memory episodes and other reexperienced or newly generated mental scenarios (Hassabis et al., 2007). The HC thus orchestrates elements of experienced environmental aspects for consolidations based on reenactment and for integration into rich mental scene construction. In this way, the HC may even influence ongoing perception in the current environment (Zeidman and Maguire, 2016).

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

Finally, the right and left TPJ are known to have hemispheric differences according to their cytoarchitectonic borders and gyrification pattern (Seghier, 2013). Neuroscientific investigations on hemispheric functional specialization have highlighted the right versus left cerebral hemisphere as dominant for attention versus language functions. The anterior RTPJ is a key node of the DMN that is central for action initiation during externally structured tasks and sensorimotor control by integrating supramodal stimulus-guided attention (Corbetta and Shulman, 2002). Involvement of this DMN node was repeatedly reported in multi-step action execution (Hartmann et al., 2005), visuo-proprioceptive conflict (Balslev et al., 2005) and multi-modal detection of sensory changes across visual, auditory, or tactile stimulation in a multi-modal fMRI study (Downar et al., 2000). In humans, direct electrical stimulation of the RTPJ during neurosurgery was associated with altered perception and stimulus awareness (Blanke et al., 2002). Importantly, the RTPJ has been shown to be intimately related to supramodal prediction and error signaling. It was argued that the RTPJ encodes actions and ensuing outcomes without necessarily relating those to outcome value (Liljeholm et al., 2013; Hamilton and Grafton, 2008; Jakobs et al., 2009). Neural activity in the RTPJ has been argued to be responsible for stimulus-driven attentional reallocation to salient and surprising sources of information as a circuit breaker that recalibrates control and maintenance systems (Bzdok et al., 2013; Corbetta et al., 2008). Indeed, patients with right TPJ damage have particular difficulties with multistep actions (Hartmann et al., 2005). In the face of large discrepancies between actual and previously predicted environmental events the RTPJ acts a potential switch between externally-oriented mind sets focussed on the sensory world and internally-oriented mind sets focussed on self-relevant mind wandering through mental scenarios. Transient RTPJ in humans for instance disrupted the impact of predicted intentions of other individuals (Young et al., 2010), a capacity believed to be subserved by the DMN. The RTPJ might hence be an important player that shifts away from the internally directed baseline processes to deal instead with immediate environmental objects and contexts.

The left TPJ in turn exhibits a close topographical relationship to Wernicke's area involved in the comprehension or understanding of written and spoken language. Neurological patients with damage caused to Wernicke's area have a major impairment of language comprehension when listening to others or reading a book. Their speech preserves natural rhythm and about normal syntax, yet the voiced sentences are devoid of meaning (i.e., aphasia). Abstracting from the typical semantic interpretations in linguistics and neuropsychology, the LTPJ probably mediates access to and integrating of world knowlege to model action concepts (Binder and Desai, 2011; Seghier, 2013). For instance, LTPJ lesions also entail problems in recognizing others' pantomimed action towards objects without obvious relation to processing any language content (Varney and Damasio, 1987) Also inner speech hinges on knowledge retrieval of statistical structure about the physical and inter-personal world. Indeed, the internal production of formulated thought ("language of the mind") was closely related to the left TPJ in a multivariate analysis of brain volume (Geva et al., 2011). Further, episodic memory recall and imagination strongly draw on complex world knowledge. Isolated

building blocks of world statistics probably get reassembled in internally generated visual scenarios that navigate present action, weigh hypothetical possibilities, and forcast the future. The LTPJ may hence facilitate the automated prediction of events by incorporating experience-derived models of the world into ongoing action, planning, and problem solving.

3 Reinforcement learning: A process model for DMN function

The reviewed neurobiological properties and likely functions of the DMN are now argued to be sufficient for implementing all the major components of a full-fledged *reinforcement learning (RL)* algorithm. Memory, reward signaling, information gathering, random sampling, and other components of intelligent RL agents appear to cross-talk in the DMN in a plausible way.

Reinforcement learning is learning to achieve a goal by interacting with an external *environment*. **XXX Don't define something with the same thing** In a general RL framework, an *agent* interacts with an environment in a trial-by-error fashion by taking *actions* a, which trigger changes in the *state* of the environment $s \to s'$, accompanied by a real-valued *reward* $r = r(s, a)$ (or regret!, since it can be negative) collected by the agent. In this view, the environment is partially controlled by the action of the agent and the reward should be thought of as an instantaneous **XXX instant is problematic here, because it may change behavior only later in RL** satisfaction **XXX or punishment?** accompanying the execution of an action, and may be delayed, only becoming substantial when accumulated over an extended period of time. The environment is generally taken as stochastic, **XXX One sentence, one bullet. Please explain stochast and PO in one sentence each.** is *partially observed*, in the sense that only part of the current state is observed by the agent (Starkweather et al., 2017). This unpredictability **XXX Explain this sentence better because it is important.** is realistic in a model which sets out to explain DMN function. Indeed, it is known that, the more the external world is predictable, the more mental activity becomes detached from the actual sensory inputs (Antrobus et al., 1966; Pope and Singer, 1978), the more stimulus-independent thoughts occur, and the higher the neural activity in the DMN. Conversely, the more the currently executed task is unknown and unpracticed, the lower the DMN activity (Filler and Giambra, 1973; Teasdale et al., 1995). These "offline" processes may however contribute to optimizing control of the organism. A DMN framework incorporating reinforcement learning can naturally embed human behavior into the tension between exploitative action with immediate gains and explorative action with longer-term reward schedules (Dayan and Daw, 2008). The implication of the DMN in a diversity of human cognitive processes can be parsimoneously explained as prediction error minimization coupled with probabilistic simulations in form of mental scenes, thus calibrating action portfolios and maximizing reward outcome at various 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 (Tenenbaum et al., 2011).

3.1 Partially Observable Markov Decision Processes

It is our view that at a sufficiently coarse scale, the brain is a physical system governed by the laws of **XXX the paper is about the DMN more specifically, perhaps try to sharpen** physics and can be locally described by Markov processes. Indeed, as noted in Tegmark (2016), 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,

defining $s(t)$, the state at time t, as the position in phase space. If the process has *memory* such that the next state depends not only on the current state but also on some 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 elements of the past.

In artificial intelligence (AI), a popular abstract model for multi-step decision processes in such an environment are so-called *Partially Observable Markov Decision Processes (POMDPs)*. A POMDP models a 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, it tries to optimize a *subjective* reward signal. This is performed alongside each transition in the environments state-space, and maintains a probability distribution over the set of possible states, based on a set of observations and observation probabilities. This is a minimal amount of assumptions that can be made about an environment, and is characteristic of so-called *model-free RL*, the approach adapted by our work. Such model-free RL is plausible in the human brain. Indeed, as was noted in Dayan and Daw (2008), it has long been suggested in the literature that there is a rather direct mapping of model-free RL learning algorithms onto the brain, **XXX New sentence** with the neuromodulator dopamine serving as a teaching signal to train values or policies by controlling synaptic plasticity at targets such as the ventral and dorsolateral striatum. **XXX Loose all boxes by integrating the text please since it is not a usual layouting**

In contrast, *model-based RL* would start off with some mechanistic assumptions about the dynamics of world. These can be assumptions about the physical laws governing the agent's world or constraints on the state space and transitions between states. For instance, if a rat in a maze knows that standing still will produce no change in the environment, and in particular will not eventually lead to finding the food. **XXX transition back to model-free RL is not clear. please make explicit** Our rat might represent such knowledge about the world as:

- $r(s, \text{“stand still”}) = 0$ if s does not correspond to a cell / chamber containing food.
- $p(s'|s, \text{“stand still”}) = 1$ if $s' = s$ and 0 otherwise.
- etc.

Mathematically, a POMDP is simply a quintuplet $(\mathcal{S}, \mathcal{A}, r, p, \mu)$ where

- $\mathcal{S}$ is the set of states, such as $\mathcal{S} = \{\text{happy}, \text{sad}\}$.
- $p : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1],\ (s, a, s') \mapsto p(s'|s, a)$, 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.
- $\mathcal{A}$ is the set of actions, such as $\mathcal{A} = \{\text{read}, \text{run}, \text{laugh}, \text{sympathize}, \text{empathize}\}$.
- $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the *reward function*, so that $r(s, a)$ is the instant reward for taking action a in state s.
- $\mu : \mathcal{S} \to [0, 1]$ is the prior probability on the states so that $\mu(s)$ is the probability that the environment starts off in state s.

3.2 Long-term rewards and action policies

The behavior of the agent is governed by some kind of *policy*, which maps states of the world to a set of candidate actions to perform given this state. Starting a time $t = 0$, a

policy π generates a trajectory of action cascades as follows

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

Since an action taken in the present moment t might have repercussions long into the 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 the future. A standard for modeling this accumulation is the so-called time-discounted *cumulative reward* function

$$G^\pi = \sum_{t=0}^{\infty} \gamma^t R_t. \tag{1}$$

This random variable[1] measures the cumulative reward of following a policy π.

Where is "value" buffered in the brain ? You mean where in the DMN! then the answer is in the vmPFC which is connected to the NAc The amygdala is known to be involved in significance evaluation) and the nucleus accumbens (NAc) is involved in reward evaluation.

The goal of RL is then to adapt this policy so as to maximize the cumulative rewards. In the definition of G^π above, the constant γ ($0 \leq \gamma < 1$) is the reward *discount factor*. Taking $\gamma = 0$ corresponds to a perfectly myopic agent who is solely concerned about their immediate rewards and has no horizon for the future, which is not compatible with long-term goal planning as potentially implemented in the DMN. To allow a learning process to arise, it is thus necessary that $0 < \gamma < 1$. Thus, the agent is positively biased towards considering rewards with a short time delay. The reward importance hence decays exponentially with distance in the future, **XXX correct grammar in previous sentences** given that there is more uncertainty in the farsighted future. The reward discount factor γ can thus be seen as calibrating the delay discounting behavior of the intelligent agent, that is the decision related to reward-delay tradeoff schedules. It is also important to appreciate that a stochastic policy estimation is more advantageous in many RL settings. For instance, a deterministic strategy in playing rock-paper-scissors can be easily exploited and a uniformly random choice is optimal. **XXX I love this example :-)**

3.3 The Q-value: A goodness measure for actions

The purpose of the DMN can then be formulated as finding a policy π for the agent which maximizes the expected cumulative value of an state-action pair (a, s), also known as the *Q-value*, namely

$$Q^\pi(s, a) = \mathbb{E}[G^\pi | s_0 = s, a_0 = a, a_1 \sim \pi, a_2 \sim \pi, \dots]. \tag{2}$$

In other words, the Q-value $Q^\pi(s, a)$ corresponds to the expected reward over all possible action trajectories in which the agent starts off starts at with an environment

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

in state s, chooses action a, and then follows the policy π thereafter to select future actions. For the brain, $Q^\pi(s,a)$ defined in (2) provides the subjective utility of executing a specific action; it answers the question "What is the expected utiltiy of taking action a in this situation ?". When a choice of actions is available, the $Q^\pi(s,a)$ provides a formalization of optimal behavior, potentially driven by the DMN.

3.4 Computing optimal behavior

By construction, maximizing the expectation in (2) is computationally intractable in general due to the enormous number of possible action trajectories the agent might need to consider simultaneously. One popular solution that remedies this issue is to only consider the sub-family of *deterministic policies*, which map each state to the best single best action which should be taken at thin state, and therefore define functions from states to actions. **XXX sentence too long, please cut into 2**

3.4.1 Q-learning

Q-learning (Watkins & Dayan, 1992), a commonly used off-policy algorithm **XXX WHATTTEEE?** , uses the *greedy policy*

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

The *Bellman equation* (Sutton and Barto, 1998) then takes the simple form

$$\begin{aligned} Q^\pi(s,a) &= \mathbb{E}_{s'\sim p(s'|s,a)}[r(s,a) + \gamma \max_{a'\in\mathcal{A}} Q^\pi(s',a')] \\ &= \mathbb{E}_{s'\sim p(s'|s,a)}[r(s,a) + \gamma Q^\pi(s',\pi(s'))], \end{aligned} \tag{4}$$

a *fixed-point* equation which provides a recursive decomposition of optimal behavior. A complicated dynamic programming optimization can be broken into simpler sub-problems at different time points, which has often been closely related to the medial prefrontal part of the DMN (Koechlin et al., 1999; Braver and Bongiolatti, 2002). Using the Bellman equation, each state can be associated with a certain value, to guide action towards a better state, thus improving on the agent's current policy. Notice that in (4), the sampling is done over only things which **XXX Ahhmm, please make a bigger effort here** depend on the environment, and so can be learned off-policy by observing state-transitions triggered by another behavioral policy, which can be stochastic.

3.4.2 Efficient learning via value approximation and experience replay

A successful RL theory needs ways to approximate the state-action value function which scale up to large state and action spaces. In state $s\in\mathcal{S}$, the agent takes an action $a\in\mathcal{A}$ by sampling from its current policy matrix, and collects a reward r, and the environment transitions to a new state $s'\in\mathcal{S}$. At the end of this step, a new *experience* $e=(s,a,r,s')$ is produced; this represents an exemplar behavior of the agent and is recorded in replay memory buffer: $D\leftarrow D\cup\{e\}$ (possibly discarding the oldest entries to make space). Now, at iteration $k+1$, replay consists in sampling (uniform or importance-weighted [2] ?) mini-batches of experiences $(s,a,r,s')\sim\mathcal{U}(D)$ from the replay memory $\mathcal{D}$ and trying to approximate the would-be Q-value for the state-action pair (s,a) as predicted by Bellman's equation (4), namely $r+\gamma\max_{a'} Q(s',a'|\theta^Q_k)$, with the output of a parametrized regression model $(s,a,r,s')\mapsto Q(s,a|\theta^Q_k)$, where $Q(.,.|\theta^Q_k)$ is an approximator for the Q-value operator, parameterized by θ^Q_{k+1}. **XXX please cut previous sentence into smaller pieces**

[2] e.g weighted by TD error of the state transition $s\xrightarrow{a}s'$.

In the human brain, such an experience replay mechanism is likely to rely on the hyppocampus (HC). Indeed, hippocampal replay is a phenomenon observed in rats, mice, cats, rabbits, songbirds, and monkeys **XXX citation !!!**. Indeed, during sleep or awake rest, replay refers to the re-occurrence of a sequence of cell activations that also occurred during activity, but the replay has a much faster time scale.

Computing an optimal policy then corresponds to finding the parameters θ^Q_{k+1} which minimize the following mean-squared loss function

$$\mathcal{L}(\theta^Q_{k+1}) = \mathbb{E}_{(s,a,r,s')\sim\mathcal{U}(D)}\left[\frac{1}{2}(Q(s,a|\theta^Q_{k+1}) - y)^2\right], \tag{5}$$

where $y = r + \gamma \max_{a'\in\mathcal{A}} Q(s',a'|\theta^Q_k) = r + \gamma Q(s',\pi(s')|\theta^Q_k)$ is the prediction target **XXX unpack what the target is !!!**. For instance, a general linear model with a kernel ϕ would be of the form

$$Q(s,a|\theta^Q) = \phi(s,a)^T\theta^Q.$$

$\phi(s,a)$ would represent a high-level hand-crafted representation of the state-action (s,a) into an appropriate feature space. Such are the techniques that have been proposed by Song et al. (2016). A recently proposed breakthrough alternative (Mnih et al., 2015; Silver et al., 2016), is to learn this representation using a deep neural net, leading to the so-called *Deep Q-learning* family of methods which are by far the current state-of-the-art in RL. In such a case, θ^Q would correspond to the strengths of low-level synaptic or high-level functional connections between different neurons / nodes / regions in the network.

Learning of the the entire model parameters can effictively be achieved via *backprop*, wherein prediction errors (aka *regrets*) are propagated from lower to higher regions to modulate the choice of future actions

$$\delta\theta^Q_{k+1} \propto -\nabla_{\theta^Q_{k+1}}\mathcal{L}(\theta^Q_{k+1}) = -\mathbb{E}_{(s,a,r,s')\sim\mathcal{U}(D)}\left[\underbrace{(Q(s,a|\theta^Q_{k+1}) - y)}_{\text{regret}}\underbrace{\nabla_{\theta^Q_k}Q(s,a|\theta^Q_{k+1})}_{???}\right]. \tag{6}$$

A note on classical RL algorithms. Note that most RL algorithms, including *Temporal Difference (TD)* learning (Sutton and Barto, 1998), REINFORCE (Williams, 1992), and SARSA can be cast in this general variational framework. For example TD corresponds to the above framework using a linear value approximator with feature encoding $\phi(s,a) = \delta_{(s,a)} =$ point mass at (s,a) on the grid $\mathcal{S}\times\mathcal{A}$, and so

$$\nabla_{\theta^Q_{k+1}}Q(s,a|\theta^Q_{k+1}) = \phi(s,a) = \delta_{(s,a)}.$$

3.4.3 Interlude: The role of the hippocampus

Does the hippocampus do Monte-Carlo sampling ? In RL, Monte-Carlo simulation can be used to update the agent's belief state (Silver and Veness, 2010). Such methods have a sample complexity that is determined only by the underlying difficulty of the POMDP, rather than the size of the state space or observation space, which can be prohibitively large. Monte-Carlo simulation provides a simple method for evaluating the value / fitness of a state. They provide an effective mechanism both for tree search and for belief state updates, breaking the curse of dimensionality and allowing much greater scalability than has previously been possible.

In the human brain, the HC could contribute to synthesizing imagined sequences of world states, observations and rewards. These simulations 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 evaluation. In these challenging POMDPs, Monte-Carlo simulation provides an effective mechanism both for tree search and for belief state updates, breaking the curse of dimensionality and allowing much greater scalability than has previously been possible. Much recent work points to Monte Carlo or stochastic sampling based approximations as a unifying framework for understanding how Bayesian inference may work practically across all these levels, in minds, brains, and machines. For sampling-based methods it is still almost non-existent beyond the description of receptive fields.

Hippocampal replay = inverse reinforcement-learning ? Given the trace $s_0, a_0, s_1, \ldots$, of an optimal agent's strategy π^* in a POMDP (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 example, given traces of motor actions from a 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 Abbeel and Ng (2004). They are also been rigourously 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.) ...

3.5 Putting everything together

The DMN is today known to consistently increase in neural activity when humans engage in goal-directed behavior that are detached from current sensory environment (Kenet et al., 2003; Fiser et al., 2004) and it was proposed to be situated at the top of the functional network hierarchies (Carhart-Harris and Friston, 2010; Margulies et al., 2016). Its involvement in thinking about the past, the future, and hypothetical possibilities ties in with the implicit computation of action and state cascades as a function of what happened in the past. The DMN may subserve constant exploration of possible future actions and their cumulative outcomes. Implicit computation of future choices provides an explanation for purposeful mind-wandering.

The HC provides perturbed action-transition-state-reward samples as a block of "imagined", "hypothesized", "recalled" experience; the small variations in these experience blocks allow searching a larger space of parameters and possible experiences. In the absence of environmental input and feedback (e.g., mind-wandering or sleep) pseudo-experiencing (i.e., emulating) possible future scenarios and action outcomes. Our approach acknowledges the unavoidable stochasticity of computation in neuronal processes (Faisal et al., 2008).

In our proposed model-free RL-based view, *inference* in the human brain reduces to generalization of policy and value adaptations 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 given environment moves to novel, yet unexperienced environments, reward prediction errors will massively increase. This will lead to adaptation of the policy until the system converges to a new steady-state of

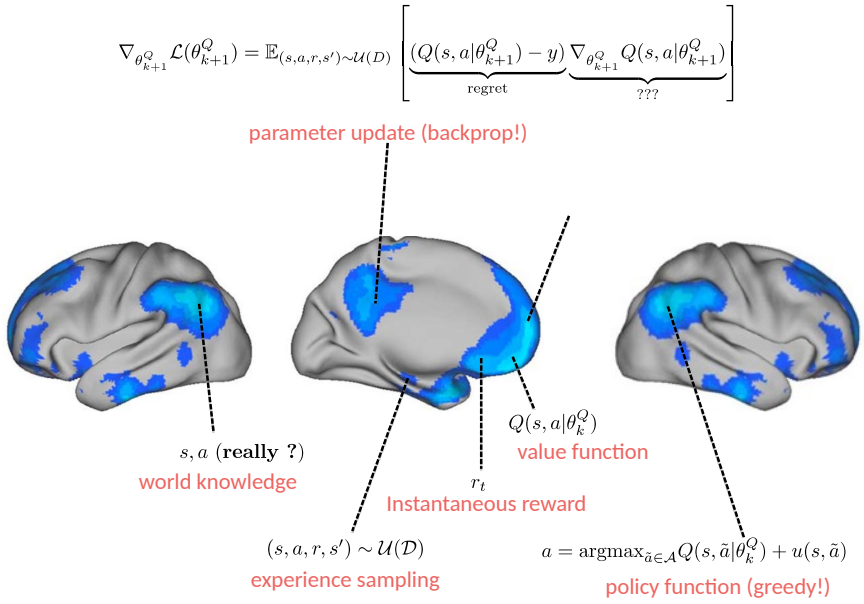

Fig 1. The DMN as an RL agent...

optimal action decisions in the volatile environment.

4 Relation to other models decision-making and optimal control for the brain

4.1 The free-energy principle and active inference

In Friston's free-energy principle (FEP) (Friston, 2010; Friston et al., 2009), the brain is portrayed as biomechanical inference engine which must minimize the long-term average of surprise. Precursors of this theory can be traced back to (Dayan et al., 1995) in which they introduced the so-called *Helmholtz machine*, a hierarchical factorial *directional deep belief-net (DBN)*. According to FEP's account, the goal of the brain is then to optimize over the generative model G: to iteratively modify its internal representation $p_G(\mathbf{z}|\mathbf{x})$ about objects in the world, their interactions and dynamics, etc., so as to minimize surprise when these representations are confronted with sensory input $\mathbf{x}$ during perception cycles. These is called the *generative* model. FEP also postulates a dual model called the *recognition* model, which works in tandem with the generative model $p_R(\mathbf{z}|\mathbf{x})$, to accomplish approximate inference. The recognition model dreams / fantasizes imaginary worlds $\mathbf{z}$ whilst the latter tries to generate sensations $\mathbf{x}$ which match these dreams!

Because surprisal is intractably difficult to optimize (since we would need to sum over all hidden causes of the sensations), FEP sets out to instead minimize an upper-bound thereupon, namely the free-energy $F^R_G(\mathbf{x})$ given by

$$\begin{aligned}\text{generative surprise } &:= -\log(p_G(\mathbf{x})) = F_G(\mathbf{x}) \\ &= \underbrace{F_G^R(\mathbf{x})}_{\text{accuracy}} - \underbrace{D_{KL}(P_R(.|\mathbf{x})||P_G(.|\mathbf{x}))}_{\text{complexity}} \\ &\leq F_G^R(\mathbf{x}), \text{ with equality if } p_R(\mathbf{z}|\mathbf{x}) = p_G(\mathbf{z}|\mathbf{x}) \text{ for all } \mathbf{z}\end{aligned} \tag{7}$$

The main algorithm for minimizing free-energy $F_G^R(\mathbf{x})$ is the *wake-sleep algorithm.* (Dayan et al., 1995). As noted already in (Dayan et al., 1995), a crucial drawback of the wake-sleep algorithm is that it requires a concurrent models (generative and recogntiion), which together do not correspond to optimization of (a bound of) the marginal likelihood (because of the incorrect KL used therein, etc.). Thus the brain could not possible be running such a algorithm, not even in principle! To the rescue, we note that the recent theory of *variational auto-encoders (VAEs)* (Kingma and Welling, 2013) might provide an efficient alternative to the wake-sleep algorithm, as it overcomes the technical limits of the former, by using a reparametrization trick. For example, unlike the wake-sleep algorithm for minimizing free-energy, VAEs can be efficiently trained via back-propagation of learning errors.

4.1.1 Comparison to our proposed theory

On the surface, a common point between the FEP and our proposed RL-based framework is placing the minimization of a surprise signal at the core of brain function. Indeed in RL, surprise minimization is subsumed by accurate prediction of rewarding outcomes in the future. A "free-energy" agent is barely a biomechanical machine which has the tendency to resist undesired / harmful phase-transitions. (Friston, 2010; Friston et al., 2009; Ortega and Braun, 2013). Such a theory cannot by itself, explain the emergence of strategic behavior inherent in humans (e.g., cite dark-room experiment).

XXX: add tons of more examples and connnections XXX: integrate aspects from section "ThE MARkOv DECISION PROBLEM" in Dayan and Daw XXX: integrate aspects from WakeNsleep by Hinton XXX: integrate aspects from Friston2014(Dropbox) on active inference XXX: integrate aspects from Sutton/Barto book chapter

4.2 Predictive Coding Hypothesis

The predictive coding framework (Clark, 2013; Friston, 2008) is a frequently evoked idea related to the default mode function (Bar et al., 2007). According to this cognitive account, cortical responses emerge from continuous functional interaction between higher and lower levels of the neural processing hierarchy. This dynamic intercourse enables inference on the world by reconciling gaps between fresh sensory input and stored prior information. Feed-forward sensory experience is constantly calibrated by top-down modulation at various hierarchical processing levels. In fact, axonal back projections probably outnumber by far the input projection existing in the brain (Salin and Bullier, 1995). This can explain the constructive, generative nature of sensory perception (Friston, 2010) and why motor action is intimately related to sensory expectations (Wolpert et al., 1995; Körding and Wolpert, 2004). At each stage of neural processing, an internally generated prediction of the sensory input is directly compared against the actual input from the environment. A prediction error at one of the processing levels, i.e. the difference between sensory input and internally predicted sensation, incurs plasticity changes of neuronal back projections (i.e., model parameters) for gradually improved future prediction of the environment. Contextual integration is hence maintained by top-down modulation of sensorimotor processing by a-priori

information about the environment. The generative model of how perceived sensory events arise in the world would be incorporated into the current neuronal wiring. Indeed, This process permits updates of the internal model of the environment to best fit the constant influx of sensory examples.

- Both the predictive coding account and POMDPs are process theories backed up by plausible neurobiological evidence that view the brain as a "statistical organ" generalizing from the past to new experiences.
- Both provide a parsimonious explanation why the human brain decreases processing load devoted to incoming information when the environment is predictable and allocate increased attentional resources when novel stimuli are encountered. Similarly, both propose explanations why predicting the future is inherently linked to information from the past.
- Sensory experience is a generative process from both views. In predictive coding, sensory perception of the external world is a generative experience due to modulatory top down experience at various hierarchical levels of sensory input processing. In our RL view, the (partially observed) environmental model is incorporated in the POMDP, *which can be fully recovered based on the last state alone*.
- Both naturally expose a mechanisms of brain plasticity in that neuronal wiring gets increasingly adapted when faced by unprecedented environmental challenges.
- The hierarchical aspects from predictive coding is re-expressed in POMDPs in form of nested prediction of probable upcoming actions and rewards.
- Both model the consequences of action. In RL, the horizon of that look into the future is explicitly manifested in the γ parameter in the Bellman equation, but not explicitly modeled in the PC account that tends to emphasize prediction about the immediate future.
- Mismatch negativity of PC is immanent in reward prediction error by the difference between actually predicted reward of an action given a state and the reward predicted by non-linear regression with graded discounting of future rewards.
- Adapting the neuronal connections for improved top-down modulation takes the concrete form of gradient computation *and back-propagation* in MDPs and RL, although this may not be neurobiologically plausible.
- In sum, the MDP account may serve as a concrete conceptualization of the PC account in cognitive neuroscience. MPDs have the advantage of exposing an explicit mechanism of the horizon of future considerations or how the internal model of the world gets updated, and collapses sensory input processing and action output preparation. As such, our account can be viewed as a RL implemention of the prediction coding hypothesis.

4.3 Semantic Hypothesis

Another frequently proposed cognitive account of DMN function revolves around forming logical associations and analogies between the current experience and the conceptual knowledge derived from past experiences (Bar, 2007; Binder et al., 1999). Analogies might naturally tie incoming novel sensory stimuli to explicit world knowledge (i.e., semantics) extractable from the environment (Bar, 2009). In this way, the encoding

of complex environments could be facilitated by relative similarity associations between states. Going much beyond human language ifself, semantic building blocks provide the basis to mentally envision non-existent scenarios that would improve optimal behavior in the environment by simulating future events. Such cognitive processes can afford the internal generation of necessary information that is not presented in the surrounding environment by recombining building blocks of concept knowledge and episodic memories (Hassabis and Maguire, 2009). Indeed, in aging humans, remembering the past and imaging the future equally decreased in the level of detail and are associated with concurrent deficits in forming and integrating relationships between items (Addis et al., 2008; Spreng and Levine, 2006). Further, a constructive account explains the reciprocal relationship between an egocentric first person perspective and an allocentric birds eye perspective immersed in self-reflection, social knowledge, and autobiographical memories. Cognitive aspects of egocentric-allocentric switching are also closely related to episodic memory, language, problem solving, planning, estimating other people's thoughts, and spatial navigation as these necessitate abstract world knowledge and abstract assocations for binding the constituent elements in mental scene construction (Schacter et al., 2007). These scene generation processes could contribute to interpreting the present and foretelling the future. This view is for instance supported by evidence in animals that could learn a *cognitive map of the environment*, even without reward incentives, and exploit it later for other means (Tolman, 1948).

- Both the semantic hypothesis and MDPs expose mechanisms of how alternative decision trees could be mentally explored.
- In both semantic hypothesis and MDPs, there is no evidence to indicate that predictions of various levels of complexity, abstraction, timescale and purpose use mechanisms that are qualitatively different. This concurs with DMN activity increases across time, space, and content domains demonstrated in many neuroimaging studies. The semantic hypothesis and RL account provide explanations why hippocampus lesion does not only impair retrieving memories, but also hypothetical and future thinking (Hassabis et al., 2007).
- The notion of semantic or knowledge association is incorporated into the MDP as the Markov property, that is, the current state directly results from the agent's history of states and actions. The learned value matrix and action transition schedules drive stimulus processing and action choice in the present.
- While mental scene construction implies a distinction between environmental perception and internally generated mind-wandering[3], the MDPs naturally integrate the former egocentric (more related to current action, state and reward) and the later allocentric (more related to past and future actions, states, and rewards) angles on the world in a same optimization problem.
- 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.
- 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.

[3] "A computational model of how such a process might be structured is far from being defined, but it 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."(Buckner and Carroll, 2007)

- The semantic hypothesis does not explain why memory recall for scene construction in humans is typically fragmentary and noisy instead of accurate and reliable. Yet, the MDP framework provides an algorithmic explanation in that stochasticity of the parameter space search implemented by Monte Carlo solvers provably yields better models of the world. 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.
- While both semantic and MDP account propose memory-based internally generated information for probabilistic mental models of action outcomes, only MDPs render explicit the grounds on which the final action is eventually chosen (i.e., the estimated cumulative reward).
- In sum, episodic scene construction according to the semantic hypothesis is lacking an explicit time and incentive model; neither does it explain how randomness of human mental experience can be beneficial.

4.4 Sentinel Hypothesis

Processing self-relevant information was perhaps the first cognitive account that was proposed for the DMN (Gusnard et al., 2001). Since then, different investigators have speculated that neural activity in the DMN may reflect the brains relentless tracking of relevance in the environment as an advantageous evolutionary adaptation. According to this cognitive account, the brain's baseline realizes a "radar" function to detect subjectively salient and unexpected events that are the most important for unfolding behavior, including resources, the securing of mates, and protection. In fact, environmental stimuli important for humans are frequently of social nature. This is probably unsurprising given that a key property of the human species is the complexity of their social systems (Tomasello, 2009). More specifically, according to the "social brain hypothesis", the human brain did not evolve to solve problems of the physical environment, but to form and maintain increasingly complex social systems to solve ecological problems by means of social relationships (Whiten and Byrne, 1988). Highly efficient neurobiological processing of social cues is exemplified by facial judgments being routinely processed in less than 100 ms (Bar et al., 2006). Our human ancestors were probably among the few organisms that were more likely to be killed by members of their own species rather than predators from another species. Being able to detect the intention and upcoming actions of conspecifics turned into a decisive evolutionary advantage (Frith and Frith, 2010). This may explain why social topics amount to roughly two thirds of human everyday communication (Dunbar et al., 1997) and why mind-wandering and dreams are so rich in stories about people and the complex relationships between them (Schilbach et al., 2008).

XXX Your focus in this subsection ought to be on the SC itself not POMPDs, etc.

- Processing social information has been proposed to underlie the physiological baseline of human brain function (Schilbach et al., 2008). This was later challenged by observing analogues of the DMN in monkeys (Mantini et al., 2011) and rats (Lu et al., 2012), with supposedly less advanced social capacities. The MDPs parsimoneously explain the dominance of social content in human mental scenes by their extremely high relevance to humans reflected by high values for social information in the value matrix (Baker et al., 2009).
- How can a same neurobiological circuit be equally important for baseline house-keeping functions and specific task performance? Our computational

account can explain why the DMN is implicated in both a goal-directed task and an idlying rest cognitive set, if environmental relevance is processed, manipulated, and retained also as a baseline function of human mental activity. In tasks, the policy and value matrices are updated to optimize short-term action, whereas, at rest, these parameter updates may improve especially mid- and long-term action.

- The sentinel account does not provide a formal mechanisms of how attention and memory resources are exactly reallocated when encountering a salient environmental stimulus. **XXX remove the next sentence. It's probably false.** In contrast, the Bellman recursion is weighted at each step by $\sum_{a \in A} \pi(a|s)$. This factor weighs which action cascades are recursively explored by the agent and which decision trees are neglected. This directly implies allocation of attention and computational load.

- In sum, MDPs imply a "radar" function of monitoring the environment for salient information in general and explain social thinking as a physiological brain baseline as an important special case, without excluding other topics contemplated in continuous mental activity.

5 Conclusion

What brain function could be important enough for the existence and survival of the human species to warrant constant, high energy costs? MDPs provide an attractive process model how the human association cortex might implement supra-modal representation and control of the environment to optimize the organism's fate. This idealized process model explains a number of previous experimental observations in the DMN by a simple but non-trivial mechanism. From the computational viewed of a Markov decision process, behavior unfolds by integrating happened past events and possible future events to guide action choice in the present context. This functional account is more compatible with the DMN's poorly understood involvement across autobiographical memory retrieval, problem solving, abstract reasoning based on internally generated scenes, social cognition, as well as delay discounting and self-related prospection. MDPs provide a mathematical formalism how optimal substructure in the environment can be recursively exploited when confronted with complicated decisions. Algorithmic gains by injecting stochasticity in the recall of past actions and outcomes to improve the internal world model may explain why very accurate memories have been disfavored in human evolution and why human creativity might be adaptive. In principle, neuroscientific experiments can be designed that operationalize the set of action, value, and state variables that determine the behavior of intelligent RL systems. The proposed machine-learning perspective on DMN function is hence not only practically computable but also neuroscientifically falsifiable. At the least, we propose an alternative vocabulary to describe and interpret experimental findings in neuroscience studies on the default mode network. Ultimately, the DMN can be viewed as a smoothing kernel extending from the relatively recent events to anticipated upcoming events to constantly improve sensory processing and action performance in the world.

Acknowledgment

References

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

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

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

Chris L Baker, Rebecca Saxe, and Joshua B Tenenbaum. Action understanding as inverse planning. Cognition, 113(3):329–349, 2009.

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

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

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

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

Moshe Bar, Maital Neta, and Heather Linz. Very first impressions. Emotion, 6(2):269, 2006.

Timothy EJ Behrens, Laurence T Hunt, Mark W Woolrich, and Matthew FS Rushworth. Associative learning of social value. Nature, 456(7219):245–249, 2008.

Timothy EJ Behrens, Laurence T Hunt, and Matthew FS Rushworth. The computation of social behavior. science, 324(5931):1160–1164, 2009.

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

Jeffrey R Binder, Julia A Frost, Thomas A Hammeke, PSF Bellgowan, Stephen M Rao, and Robert W Cox. Conceptual processing during the conscious resting state: a functional mri study. Journal of cognitive neuroscience, 11(1):80–93, 1999.

Olaf Blanke, Stphanie Ortigue, Theodor Landis, and Margitta Seeck. Neuropsychology: Stimulating illusory own-body perceptions. Nature, 419(6904):269–270, 2002.

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

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

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

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

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

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

Danilo Bzdok and Simon Eickhoff. The resting-state physiology of the human cerebral cortex. Technical report, Strukturelle und funktionelle Organisation des Gehirns, 2015.

Danilo Bzdok, Adrian Heeger, Robert Langner, Angela R Laird, Peter T Fox, Nicola Palomero-Gallagher, Brent A Vogt, Karl Zilles, and Simon B Eickhoff. Subspecialization in the human posterior medial cortex. Neuroimage, 106:55–71, 2015.

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

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

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

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

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

Paula L Croxson, Heidi Johansen-Berg, Timothy EJ Behrens, Matthew D Robson, Mark A Pinsk, Charles G Gross, Wolfgang Richter, Marlene C Richter, Sabine Kastner, and Matthew FS Rushworth. Quantitative investigation of connections of the prefrontal cortex in the human and macaque using probabilistic diffusion tractography. The Journal of neuroscience, 25(39):8854–8866, 2005.

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

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

Peter Dayan, Geoffrey E Hinton, Radford M Neal, and Richard S Zemel. The helmholtz machine. Neural computation, 7(5):889–904, 1995.

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

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

Robin IM Dunbar, Anna Marriott, and Neil DC Duncan. Human conversational behavior. Human Nature, 8(3):231–246, 1997.

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

A Aldo Faisal, Luc PJ Selen, and Daniel M Wolpert. Noise in the nervous system. Nature reviews neuroscience, 9(4):292–303, 2008.

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

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

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

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

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

Uta Frith and Chris Frith. The social brain: allowing humans to boldly go where no other species has been. Philosophical Transactions of the Royal Society B: Biological Sciences, 365(1537):165–176, 2010.

Hagar Gelbard-Sagiv, Roy Mukamel, Michal Harel, Rafael Malach, and Itzhak Fried. Internally generated reactivation of single neurons in human hippocampus during free recall. Science, 322(5898):96–101, 2008.

Sharon Geva, P Simon Jones, Jenny T Crinion, Cathy J Price, Jean-Claude Baron, and Elizabeth A Warburton. The neural correlates of inner speech defined by voxel-based lesion–symptom mapping. Brain, 134(10):3071–3082, 2011.

Jan Gläscher, Ralph Adolphs, Hanna Damasio, Antoine Bechara, David Rudrauf, Matthew Calamia, Lynn K Paul, and Daniel Tranel. Lesion mapping of cognitive control and value-based decision making in the prefrontal cortex. Proceedings of the National Academy of Sciences, 109(36):14681–14686, 2012.

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

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

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

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

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

Tom Hartley, Colin Lever, Neil Burgess, and John O'Keefe. Space in the brain: how the hippocampal formation supports spatial cognition. Phil. Trans. R. Soc. B, 369(1635): 20120510, 2014.

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

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

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

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

Oliver Jakobs, Ling E Wang, Manuel Dafotakis, Christian Grefkes, Karl Zilles, and Simon B Eickhoff. Effects of timing and movement uncertainty implicate the temporo-parietal junction in the prediction of forthcoming motor actions. Neuroimage, 47(2):667–677, 2009.

William James. The principles of. Psychology, 2, 1890.

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

Tal Kenet, Dmitri Bibitchkov, Misha Tsodyks, Amiram Grinvald, and Amos Arieli. Spontaneously emerging cortical representations of visual attributes. Nature, 425(6961): 954–956, 2003.

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

Etienne Koechlin, Gianpaolo Basso, Pietro Pietrini, Seth Panzer, and Jordan Grafman. The role of the anterior prefrontal cortex in human cognition. Nature, 399(6732):148–151, 1999.

Nils Kolling, Marco K Wittmann, Tim EJ Behrens, Erie D Boorman, Rogier B Mars, and Matthew FS Rushworth. Value, search, persistence and model updating in anterior cingulate cortex. Nature Neuroscience, 19(10):1280–1285, 2016.

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

Maël Lebreton, Soledad Jorge, Vincent Michel, Bertrand Thirion, and Mathias Pessiglione. An automatic valuation system in the human brain: evidence from functional neuroimaging. Neuron, 64(3):431–439, 2009.

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

Hanbing Lu, Qihong Zou, Hong Gu, Marcus E Raichle, Elliot A Stein, and Yihong Yang. Rat brains also have a default mode network. Proceedings of the National Academy of Sciences, 109(10):3979–3984, 2012.

Eleanor A Maguire, David G Gadian, Ingrid S Johnsrude, Catriona D Good, John Ashburner, Richard SJ Frackowiak, and Christopher D Frith. Navigation-related structural change in the hippocampi of taxi drivers. Proceedings of the National Academy of Sciences, 97(8): 4398–4403, 2000.

Dante Mantini, Annelis Gerits, Koen Nelissen, Jean-Baptiste Durand, Olivier Joly, Luciano Simone, Hiromasa Sawamura, Claire Wardak, Guy A Orban, Randy L Buckner, et al. Default mode of brain function in monkeys. The Journal of Neuroscience, 31(36): 12954–12962, 2011.

Daniel S Margulies, Satrajit S Ghosh, Alexandros Goulas, Marcel Falkiewicz, Julia M Huntenburg, Georg Langs, Gleb Bezgin, Simon B Eickhoff, F Xavier Castellanos, Michael Petrides, et al. Situating the default-mode network along a principal gradient of macroscale cortical organization. Proceedings of the National Academy of Sciences, page 201608282, 2016.

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

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

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

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. Nature, 518(7540):529–533, Feb 2015. Letter.

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

Antoinette Nicolle, Miriam C Klein-Flügge, Laurence T Hunt, Ivo Vlaev, Raymond J Dolan, and Timothy EJ Behrens. An agent independent axis for executed and modeled choice in medial prefrontal cortex. Neuron, 75(6):1114–1121, 2012.

Pedro A Ortega and Daniel A Braun. Thermodynamics as a theory of decision-making with information-processing costs. In Proc. R. Soc. A, volume 469, page 20120683. The Royal Society, 2013.

John P ODoherty, Sang Wan Lee, and Daniel McNamee. The structure of reinforcement-learning mechanisms in the human brain. Current Opinion in Behavioral Sciences, 1:94–100, 2015.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. Nature, 529(7587):484–489, 2016.

Erez Simony, Christopher J Honey, Janice Chen, Olga Lositsky, Yaara Yeshurun, Ami Wiesel, and Uri Hasson. Dynamic reconfiguration of the default mode network during narrative comprehension. Nature Communications, 7, 2016.

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

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

Clara Kwon Starkweather, Benedicte M Babayan, Naoshige Uchida, and Samuel J Gershman. Dopamine reward prediction errors reflect hidden-state inference across time. Nature Neuroscience, 2017.

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

Richard S Sutton and Andrew G Barto. Reinforcement learning: An introduction. MIT press, 1998.

John D Teasdale, Barbara H Dritschel, Melanie J Taylor, Linda Proctor, Charlotte A Lloyd, Ian Nimmo-Smith, and Alan D Baddeley. Stimulus-independent thought depends on central executive resources. Memory & cognition, 23(5):551–559, 1995.

Max Tegmark. Improved measures of integrated information. PLOS Computational Biology, 12(11):e1005123, 2016.

Joshua B Tenenbaum, Charles Kemp, Thomas L Griffiths, and Noah D Goodman. How to grow a mind: Statistics, structure, and abstraction. science, 331(6022):1279–1285, 2011.

Edward C Tolman. Cognitive maps in rats and men. Psychological review, 55(4):189, 1948.

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

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

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

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

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

Andrew Whiten and Richard W Byrne. The machiavellian intelligence hypotheses: Editorial. 1988.

Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. Machine Learning, 8(3):229–256, 1992.

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

Wako Yoshida, Ben Seymour, Karl J Friston, and Raymond J Dolan. Neural mechanisms of belief inference during cooperative games. The Journal of Neuroscience, 30(32):10744–10751, 2010.

Liane Young, Joan Albert Camprodon, Marc Hauser, Alvaro Pascual-Leone, and Rebecca Saxe. Disruption of the right temporoparietal junction with transcranial magnetic stimulation reduces the role of beliefs in moral judgments. Proceedings of the National Academy of Sciences, 107(15):6753–6758, 2010.

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

A 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 (Dayan et al., 1995) 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 example 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.

A.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,

$$\begin{aligned}
\text{generative surprise } &= -\log(p_G(\mathbf{x})) = -\log(p_G(\mathbf{x}))\times 1 = -\log(P_G(\mathbf{x}))\sum_{\mathbf{z}} p_G(\mathbf{z}|\mathbf{x}) \\
&= -\sum_{\mathbf{z}} p_G(\mathbf{z},\mathbf{x})\log(p_G(\mathbf{x})) = -\sum_{\mathbf{z}} p_G(\mathbf{z}|\mathbf{x})\log(p_G(\mathbf{z},\mathbf{x})/p_G(\mathbf{z}|\mathbf{x})) \\
&= \sum_{\mathbf{z}} p_G(\mathbf{z}|\mathbf{x})\log(p_G(\mathbf{z}|\mathbf{x})) - \sum_{\mathbf{z}} p_G(\mathbf{z}|\mathbf{x})\log(p_G(\mathbf{z},\mathbf{x})) \\
&= -\langle\log(p_G(.,\mathbf{x}))\rangle_{p_G(.|\mathbf{x})} - \mathcal{H}(p_G(.|\mathbf{x})) \\
&= \langle E_G(.,\mathbf{x})\rangle_{p_G(.|\mathbf{x})} - \mathcal{H}(p_G(.|\mathbf{x}))
\end{aligned} \tag{8}$$

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

$$E_G(\mathbf{z},\mathbf{x}) := -\log(p_G(\mathbf{z},\mathbf{x})), \tag{9}$$

The last quantity in (8) is nothing but *Helmholtz free-energy* (at unit temperature!), defined formally by

$$F_G(\mathbf{x}) := \langle E_G(.,\mathbf{x})\rangle_{p_G(.|\mathbf{x})} - \mathcal{H}(p_G(.|\mathbf{x})). \tag{10}$$

Thus,

Fact A.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.

A.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 (9) and applying Bayes rule, we get the Gibbs distribution

$$p_G(\mathbf{z}|\mathbf{x}) = \frac{p_G(\mathbf{z}|\mathbf{x})}{p_G(\mathbf{x})} = \frac{\exp(-E_G(\mathbf{z},\mathbf{x}))}{Z_G(\mathbf{x})} = \frac{\exp(-E_G(\mathbf{z},\mathbf{x}))}{Z_G(\mathbf{x})}, \tag{11}$$

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 9. 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

$$F_G(\mathbf{x}) \overset{(8)}{=} -\log(p_G(\mathbf{x})) = -\log(Z_G(\mathbf{x})) = E_G(\mathbf{z},\mathbf{x}) + \log(p_G(\mathbf{z}|\mathbf{x})). \tag{12}$$

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[4] $p_R(.|\mathbf{x})$ yields

$$\begin{aligned} F_G(\mathbf{x}) &= -\log(Z_G(\mathbf{x})) = \langle E_G(.,\mathbf{x})\rangle_{P_R(.|\mathbf{x})} + \sum_{\mathbf{z}} p_R(\mathbf{z}|\mathbf{x}) \log(p_G(\mathbf{z}|\mathbf{x})) \\ &= \langle E_G(.,\mathbf{x})\rangle_{P_R(.|\mathbf{x})} - \mathcal{H}(p_R(.|\mathbf{x})) - \sum_{\mathbf{z}} p_R(\mathbf{z}|\mathbf{x}) \log(p_R(\mathbf{z}|\mathbf{x})/p_G(\mathbf{z}|\mathbf{x})) \\ &= F_G^R(\mathbf{x}) - D_{KL}(P_R(.|\mathbf{x})||P_G(.|\mathbf{x})), \end{aligned} \tag{13}$$

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

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

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 $F_G^G = F_G$.

A.3 A general free-energy principle

We can resume the situation as follows[5]:

$$\begin{aligned} \text{generative surprise } &:= -\log(p_G(\mathbf{x})) = F_G(\mathbf{x}) \\ &= \underbrace{F_G^R(\mathbf{x})}_{\text{accuracy}} - \underbrace{D_{KL}(P_R(.|\mathbf{x})||P_G(.|\mathbf{x}))}_{\text{complexity}} \\ &\leq F_G^R(\mathbf{x}), \text{ with equality if } p_R(\mathbf{z}|\mathbf{x}) = p_G(\mathbf{z}|\mathbf{x}) \text{ for all } \mathbf{z} \end{aligned} \tag{15}$$

[4] 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.

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

A.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

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

A.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 (Kingma and Welling, 2013), an other groups, via a "reparametrization trick".

A.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 (Kingma and Welling, 2013). Define the data-dependent auxiliary random function

$$f_{G,R}(.,\mathbf{x}) : \mathbf{z} \mapsto \log(p_G(\mathbf{z},\mathbf{x})) - \log(p_R(\mathbf{z}|\mathbf{x})). \tag{16}$$

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. 2. Some examples of the reparametrization trick for a number of choices of the posterior distribution are given in Tab. 2.

Fig 2. Variational autoencoders...

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 (Kingma and Welling, 2013) for a variety of models.

A.7 GANs and other likelihood-free methods

...