CAN WORKING MEMORY BE EXPLAINED BY PREDICTIVE CODING?

A PREPRINT

Mengli Feng*
Informatics Institute
University of Amsterdam
Amsterdam, The Netherlands

Michael D. Nunez
Psychological Methods
University of Amsterdam
Amsterdam, The Netherlands

June 6, 2025

ABSTRACT

Predictive coding (PC) is a theory in cognitive/computational neuroscience which explains cortical functions with a hierarchical process of minimizing prediction errors. PC provides a neuronal scheme (through neurons or neural populations) for implementing Bayesian inference in the brain. PC recovers the hidden state of the world from sensory input (passive inference) and selects actions to reach the goals the agent has (active inference). Since its discovery, PC has been found to be a unifying theory explaining more and more cognitive functions, including perception, attention, and action planning. In this paper, we review and discuss how PC can be used also as a powerful tool to understand working memory (WM), an essential function for executive control. Specifically, we sought answers in the literature to the following questions: 1. How is WM maintained and updated? 2. What is the relationship between attention and WM and how do they interact? 3. Why does WM have limited capacity? and 4. Why is WM hierarchical? Modelling WM in PC frameworks provides alternative explanations to some long-standing questions about WM and may help with resolving the conflicts between WM theories. We expect such alternative explanations can help future researchers, such as in developing computational tools to improve treatments for brain disorders and more robust artificial working memory in artificial intelligence.

Keywords Predictive coding · Working memory · Active inference · Markov decision process

1 Introduction

Working memory (WM) has been at the centre stage of psychology and neuroscience for decades. While numerous models of WM have been developed over the years, there are still many unresolved debates and mysteries about how WM is maintained and updated, why it has limited capacity, and why it has hierarchical representations, i.e. where different levels of abstraction emerge Luu and Stocker [2021], Hasson et al. [2015], Honey et al. [2012], Brady and Alvarez [2011], Lerner et al. [2011], Hasson et al. [2008]. Meanwhile, predictive coding (PC) has recently been identified as a promising unifying theory for explaining different functions of the brain [Millidge et al., 2021a, Smith et al., 2022], such as perception, learning and action planning. The functional role of WM links to PC through many related concepts (e.g. executive control, perception, attention and action planning), upon which PC has already established relatively mature theories (see Figure 1). We expect therefore for PC to be able to help with delineating some similar concepts from WM, answering some long-standing questions about WM and promoting a more formal understanding of WM. In this work, we draw connections between the literature to seek support for the hypothesis that WM can be explained by PC. Details of the literature and what we consider support for this hypothesis are described in this paper.

* *Corresponding author* Email: mengli.feng@student.uva.nl

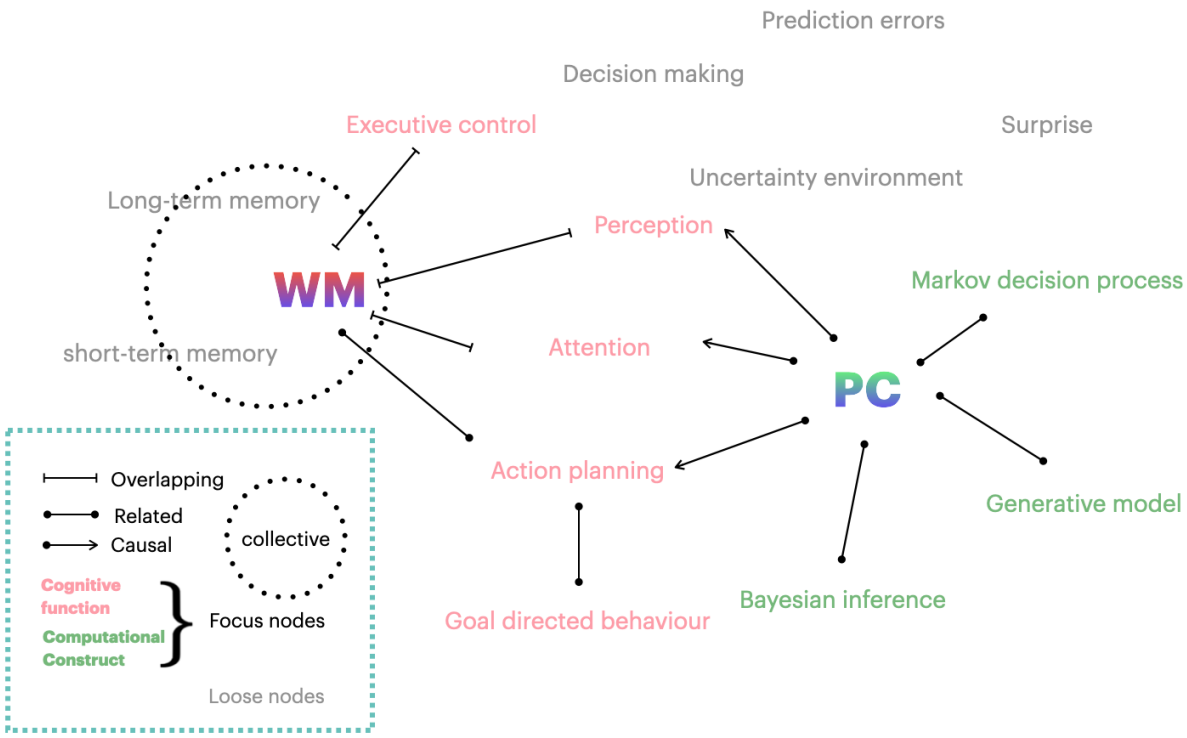

Figure 1: Links between working memory and predictive coding.
This graph illustrates how WM and PC are connected through many key concepts across multiple different disciplines. On the left we have the core concept working memory 'WM' and on the right we have the core concept predictive coding 'PC'. Nodes that related to the two terms are placed around them according to the principle of proximity. The nodes highlighted in colours are the ones that are *focused*, with the pink ones being cognitive functions and green ones being computational constructs. The grey nodes are *loose* nodes, which are less focused. Three different types of edges are used to show the relationship between nodes. The ones with two bars at the ends denotes nodes that have conceptual overlapping; the ones with two circles at the ends denotes nodes that are strongly related; the ones with one circle at an end and an arrow at the other end denotes causal relations. The dotted circle is used to show a collective of concepts that belong to the same category. In this case, it shows that working memory, short-term memory and long-term memory belongs to the category 'memory'. This network does not exhaust all possible connections but provides a literature-searching strategies. 1. One can use the related or overlapping nodes to substitute the core concepts. For example, one can search for "modelling WM with Bayesian inference" instead of "modelling WM with PC". 2. One can look into specific aspects of the link between the core concepts. For example, one can search for the links between WM and PC in terms of attention.

1.1 Four objectives

In this literature review, we investigate whether *predictive coding* (PC) can help with answering four research questions:

1. How is working memory maintained and updated?
2. Why does working memory have limited capacity?
3. What is the relationship between attention and working memory and how do they interact?
4. Why is working memory hierarchical?

We attempt to answer these questions with the literature using a theoretical perspective. We put more effort into analyzing literature that have the most established models available on these topics, while supplementing them with other relevant studies and evidence. An auxiliary objective of this literature review was to find the ingredients to better formalize WM in PC framework for modelling work in the future. This was achieved by summarizing current modelling

attempts and trying to be specific about modelling details when answering the research questions. Note that we sought to introduce the mathematical concepts and derivations in this paper as intuitively as possible. For those readers who have difficulties understanding these concepts, we refer to readers to the original papers cited around the math and derivations. We expect that mathematical formalization will help experimental work of working memory and predictive coding in the future.

2 Working Memory (WM)

Although WM has drawn attention since the 1890s [Yu and Friston, 2014], many psychological and neuroscientific questions about WM remain unanswered or under debate. We do not intend to exhaust all the details in each topic, but instead give an overview of those topics and provide perspectives for future work, as guided by the four questions above.

Working memory (WM) can be understood as *short-term memory* (STM) with an emphasis on the functional role of maintaining useful information for *executive control*, which is the ability to carry out goal-directed behaviour [Cowan, 2009, Oberauer and Lin, 2017, Miller et al., 2018, Oberauer, 2019, Trapp et al., 2021, Friedman and Robbins, 2022]. WM provides an interface for perception, long-term memory and action [Baddeley, 2003] and plays an important role in comprehension, learning, planning and reasoning [Cowan, 2014]. The importance of WM in human cognition is apparent from human behavior in patients with deficits in WM, such as in behavioral and brain disorders like ADHD Ortega et al. [2020], schizophrenia [Eryilmaz et al., 2016], and dementia Jahn [2013]. WM is also very important in the field of artificial intelligence to ensure robust performance Zheng et al. [2016].

In a typical WM task, a participant needs to keep a piece of information in mind in order to respond to a stimulus after a delay period. Figure 2 shows some components that are commonly seen in a WM task. Often, a working memory experimental trial starts with an initial cue to remember a stimulus, and ends with a target cue, where the participants need to recall what was shown at the beginning [Berlot and de Lange, 2022] or decide whether it is the same as the initial cue [Parr and Friston, 2017, Yu and Friston, 2014]. Sometimes, there is an anticipatory cue shown before the initial cue, to set up an expectation, e.g. whether there will be an update. In the middle of an experimental trial, sometimes a retro-cue showing partial or all the information about the initial stimuli [Parr and Friston, 2017] or a new cue [Yu and Friston, 2014] is presented. In addition, noise can be presented at any point in time to disturb WM [Feng et al., 2023]. During a task, WM is anticipated, initiated, disturbed by noise, updated and recalled. The temporal complexity of WM tasks can test how WM functions in a world with uncertainties.

One of the essential functions of WM is to retain information accurately and update information flexibly. Robust WM maintenance is very important to cope with noisy environments. While it keeps information on hold, it also needs to adapt to new incoming evidence. So one classical question about WM is how it balances the maintenance of old information and the assimilation of new information [Yu and Friston, 2014].

Attention and working memory share many similarities, both psychologically [Oberauer, 2019] and neuroscientifically [Mayer et al., 2007]. Functionally, they are both part of executive control that supports many other cognitive functions and conscious awareness. Those functional overlaps are accompanied by shared neural substrates [Knudsen, 2007]. Although they are seen as closely related [Oberauer, 2019], how to formally delineate them is not often discussed.

There are many studies that use PC to explain attention [Yu and Friston, 2014, Feldman and Friston, 2010, Pauszek, 2019, Hohwy, 2012, Ransom et al., 2016]. They provides an opportunity for me to inspect attention and working memory in the same theoretical and modelling framework and ask whether WM can be delineated from attention and how they interact. This can potentially help us to reach a clearer understanding of the functional and computational role of WM.

Only a limited number of items can be held in WM. It was estimated experimentally that humans can keep four items at once in their STM [Cowan, 2000]. The limit of WM capacity was mechanistically accounted for by information decay or interference [Barrouillet and Camos, 2009, Oberauer, 2019], 'subcycles' in the brain oscillation [Lisman and Idiart, 1995] or the refresh rate of synaptic weight change [Miller et al., 2018].

It has been pointed out that the capacity limit of WM might be related to the capacity limit of attention Palmer [1990]. Another study showed that dopamine-related heterogeneity reflects individual differences in WM capacity [Cools and D'Esposito, 2011]. Considering dopamine is a neural signature of predictive processing, PC might bring new insights into why WM has a capacity limit.

Both psychological and neuroscientific research show evidence of the hierarchical structure of WM [Luu and Stocker, 2021, Berlot and de Lange, 2022, Hasson et al., 2015], meaning WM can maintain information at different levels of abstraction. Why and how WM is hierarchically constructed may be answered by hierarchical PC framework.

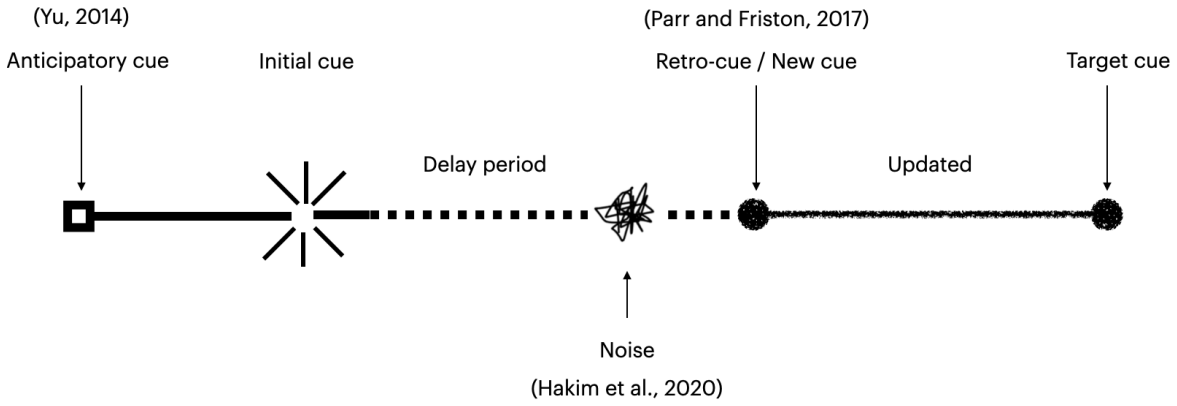

Figure 2: an illustration of WM tasks.
The graph shows possible elements in a trial of a working memory task during behavioral experiments. A trial can start with an *anticipatory cue*, which is a cue indicating what might happen next. Then an *initial cue* might be presented, which is the cue that the participants need to remember and to be tested at the end. Afterwards, there is normally a *delay period*, which is a period the participants are presented with no cues. During the delay period, *noise* stimuli are often presented to create distractions that are irrelevant to the task. After the delay period, a *retro-cue* might be presented, which is a cue that provides additional information about the initial cue. A *new cue* may also be presented, which is a cue to be tested that is different from the initial cue. After the retro-cue or the new cue being presented, there is a period of *update*, for WM to be updated after the retro-cue/new cue. Finally, a *target cue* will be presented, with a typical testing question being 'is the cue the same as the initial cue?'

.

3 Predictive Coding (PC)

The classical *Predictive Coding* (PC) algorithm proposes that cognitive functions are realised through the minimization of prediction errors in the brain, the difference between the actual sensory input (e.g. from a stimulus) and the predicted sensory input[Aitchison and Lengyel, 2017],

$$\text{prediction error} = \text{input} - \text{prediction} \tag{1}$$

The term 'predictive coding' was first introduced in the paper by Rao and Ballard [1999]. It introduced a model of visual processing where *feedback* connections from higher to lower areas carry predictions of the signals in the lower area, while the *feedforward* connections carry the difference between the prediction and actual signals in the lower areas. Over the years, this simple idea has formed a unifying theory that explains many cortical functions in computational and cognitive neuroscience [Millidge et al., 2021a, Clark, 2013, Friston, 2017]. It explains *perception* as a process of inferring hidden states of the world, and explains *learning* as a process of updating the internal model of the world. It also explains *adaptive behaviours* as the process of inferring action plans, and *attention* as the precision of inferred states [Feldman and Friston, 2010]. Both neuroimaging and neuro-computational studies have been providing supporting evidence to this theory [Cullen, 2020, Isomura et al., 2022, Caucheteux et al., 2023, Mikulasch et al., 2023].

3.1 Predictive coding as an implementation of Bayesian inference

Bayesian inference uses Bayes theorem (Equation 2) to infer the hidden state of the world or latent cause (denoted as s) given an observation/sensory input (denoted as o).

$$p(s \mid o) = \frac{p(o \mid s)p(s)}{p(o)} \tag{2}$$

In this framework, the latent state can be inferred by first calculating the posterior probability distribution $p(s \mid o)$ of the latent states, then finding the state that results in the maximum value of the posterior. This would result in the so-called

Maximum A Posteriori (MAP) estimate, the mode of the posterior distribution that results in the maximum posterior value.

Aitchison and Lengyel [2017] describe the relationship between Bayesian inference and predictive coding as "complementary", in the sense that the latent cause s provided by Bayesian inference can be used to calculate the *prediction* of the current sensory input as the *expected value* of sensory input o given the latent cause s. The *prediction* is thus mathematically defined by the following integral to calculate an expected value E with the subscript denoting what we are integrating with respect to:[2]

$$\text{prediction} = \mathrm{E}_{p(o|s)}\left[o\right] = \int op(o \mid s)do \tag{3}$$

This prediction can then be used to calculate the *prediction error* to correct the weights for the neural activity encoding possible states. This is done in order to perform Bayesian inference to infer the state given an observation [Spratling, 2017]. See Rao and Ballard [1999] for a detailed algorithm for updating beliefs through prediction errors. The reciprocal process combining Bayesian inference and predictive coding is introduced as *Bayesian predictive coding* in [Aitchison and Lengyel, 2017].

We do not make the claim that Bayesian inference *is* predictive coding. Aitchison and Lengyel [2017] make the point that predictive coding is not the only neural arithmetic to implement Bayesian inference, and Bayesian inference is not the only computational goal that predictive coding serves.

Formally, predictive coding could instead be seen as a type of variational Bayesian inference [Zhang et al., 2019], where the posterior distribution is approximated instead of directly solved calculation in the brain. This approximation can be achieved by minimizing the difference between the estimated posterior $q(s)$ and the true joint distribution $p(o, s)$. We can therefore approximate the true posterior through an iterative algorithm, without worrying about the intractable marginal distribution of sensory inputs $p(o)$ due to the high dimensionality of the representation of an observation (e.g. many pixels in an image). [3] A thorough mathematical derivation can is given by Millidge et al. [2021a].

3.2 Predictive coding for active inference

Passive inference passively infers the state of the world based on observations. *Active inference* (AC) is an extension of passive inference in that AC also infers the actions which lead to the observations, in order to prevent surprises (e.g. being hit by a car while riding a bike through an intersection). Because of the close link between goal-directed action planning and WM, we expected active inference could be very useful in explaining WM.

For readers to better understand the mathematical formulation of active inference, we would like to first introduce the concepts of entropy and relative entropy. Entropy is a measurement of expected surprise for a random variable X, that is $\mathrm{E}[-\log P(x)]$, where surprise is defined as the logarithm of the inverse of a probability of an event, such that $\log(\frac{1}{p(X=x)}) = -\log P(x)$. Thus, the less probable the event, the more surprising it is. *Relative entropy* measures the similarity between two statistical distributions ($\mathrm{E}[P\left(\log \frac{P(x)}{Q(x)}\right)]$), which can be understand as the surprise caused by the difference between the two distribution. In equations, it is often written as $D_{KL}(q\|p)$ for the similarity of distributions q and p. Another name for relative entropy is *KL divergence*.

The core of active inference is the concept of *variational free energy* F (VFE, see Equation 4). Active inference (AC) minimizes variational free energy F. The idea of 'free energy' was borrowed from thermodynamics, where free energy refers to the usable energy that can do work. In thermodynamics, minimising free energy drives the system towards equilibrium. Similarly in active inference, minimising free energy also drives the system towards equilibrium, where the approximated distribution is equal to the true distribution [Friston, 2008]. In mathematical terms, *Variational free energy* F is the *relative entropy* between the true joint distribution of sensory inputs o and states s given a *policy* (π, also known as *selected actions*) $p(o, s \mid \pi)$ and the *approximate* posterior distribution of states given selected actions $q(s \mid \pi)$ (Equation 5). Note that sensory inputs o are not included in the *approximate* posterior distribution q of the true posterior distribution $p(s, o \mid \pi)$. This is because the estimate q is obtained by an iterative optimisation procedure and not actually dependent on observations [Smith et al., 2022]. Minimizing F therefore enables us to acquire an estimated posterior.

[2] Note that an *expected value* is analogous to a weighted average if sensory input o were discrete.

[3] intractability issue: an observation can be multidimensional such as an image having many pixels. for each dimension/pixel, there can be N different states, then for D dimensions, we have N^D states, which explore exponentially with the number of dimensions, making the computation impossible

$$F_\pi = \mathrm{E}_{q(s|\pi)}\left[\ln \frac{q(s \mid \pi)}{p(o, s \mid \pi)}\right] \tag{4}$$

$$F = \mathrm{E}_{q(s)}\left[\ln \frac{q(s \mid \pi)}{p(s \mid o, \pi)}\right] - \ln p(o) \tag{5}$$

After a rearrangement of the equation, F can be understood as complexity minus accuracy (see Equation 6). The first term measures the difference between the prior and posterior belief. A larger change of beliefs to account for an observation entails a greater complexity. The second term measures the expected likelihood of observations given beliefs about the states. The accuracy of the inference increases with increasing expected likelihood. This can be further reformulated into entropy minus energy, which is in the end a term of linear combinations of squared prediction errors [Millidge et al., 2021a]. Because F can is also the upper bound of negative log evidence (see Inequality 7), minimizing F also leads to the minimization of expected surprise (i.e. entropy). A more extensive mathematical derivation can be seen in Smith et al. [2022].

$$F = \underbrace{D_{KL}[q(s \mid o, \pi) \| p(s \mid \pi)]}_{\text{Complexity}} - \underbrace{\mathrm{E}_{q(s|\pi)}[\ln p(o \mid s)]}_{\text{Accuracy}} \tag{6}$$

$$-\ln p(o) = -\ln \mathrm{E}_{q(s|\pi)}\left[\frac{p(o, s \mid \pi)}{q(s, \pi)}\right] \leq -\mathrm{E}_{q(s,\pi)}\left[\ln \frac{p(o, s \mid \pi)}{q(s \mid \pi)}\right] = F \tag{7}$$

3.3 Expected free energy

While free energy is a measurement for the present, *expected free energy* (EFE) G is a measure for the future. It is the free energy calculated with respect to the expected observations (Equation 8). This means calculating future VFE based on beliefs about future observations, which enters the expectation operator E_q as a random variable [Millidge et al., 2021b, Smith et al., 2022], leading to the main difference from Equation 4. To accommodate preferred future observations, the equation can be rearranged into 9, where C denotes preference and $p(o|C)$ represents the probability of a preferred observation. This is often referred as *pragmatic value* [Smith et al., 2022], that is to guide action selection to achieve the desired outcome/observation. To minimize G, one needs to select actions that can produce targeted/preferred future observations. The expected value reaches its maximum when $q(o|\pi)$ approaches $p(o|C)$. Intuitively, this means assigning larger probabilities to larger values of $p(o|C)$ and vice versa.

$$G_\pi = \mathrm{E}_{q(o,s|\pi)}\left[\ln \frac{q(s \mid \pi)}{p(o, s \mid \pi)}\right] \tag{8}$$

$$G \approx -\underbrace{\mathrm{E}_{q(o,s|\pi)}[\ln q(s \mid o, \pi) - \ln q(\,\mathrm{s} \mid \pi)]}_{\text{information gain (epistemic value)}} - \underbrace{\mathrm{E}_{q(o|\pi)}[\ln p(o \mid C)]}_{\text{preference (pragmatic value)}} \tag{9}$$

The idea is to get the posterior belief on each policy based on how much the expected observations under a policy will match the preferred observation. The first term in equation 9, on the other hand, is the difference between the posterior and prior beliefs, which can be understood as the information gain, often referred as *epistemic value* [Smith et al., 2022]. To minimize G, the *agent*, that is the animal/person who initiates the action, must select a policy to maximize the information gain. This arrangement shows that active inference maximises both the epistemic and pragmatic values of the predictions.

EFE can again be reformulated in two parts in Equation 10. The left term is exploitative, used to maximise reward by matching the expected observations with the preferred observations. It is also referred to as risk or expected complexity. The right term is exploratory, used to reduce the uncertainty/ambiguity of predictions of the outcome given a state [4], which has a similar function to the first term in Equation 9 entailing epistemic affordance. (10). The decomposition in Equation 9 and Equation 10 is very useful later in the discussion in Section 5 for understanding the relationship

[4] $H[p()]$ is a notation for entropy. Higher entropy leads to a higher dispersion of the probability distribution. The dispersion can be seen as a measure of uncertainty. To reduce the dispersion, one can try making observations that will lead to most evidence given a state

between WM and goal-directed behaviour (Sect. 4.1), the relationship between WM and attention (Sect. 7) and how WM capacity is constrained by the trade-off between complexity and accuracy (Sect. 5.3).

$$G_\pi = \underbrace{D_{KL}[q(o \mid \pi) \| p(o \mid C)]}_{\text{exploitative (Reward)}} + \underbrace{\mathrm{E}_{q(s|\pi)}[\mathrm{H}[p(o \mid s)]]}_{\text{exploratory (Uncertainty)}} \tag{10}$$

While classical PC defines prediction errors as the difference between predicted sensory input and actual sensory input, there are two types of prediction errors in active inference, one for *state prediction* and one for *outcome prediction* [Smith et al., 2022]. The state prediction error refers to the discrepancy between the predicted and actual internal states while the outcome prediction error referees to the discrepancy between expected and received outcomes/observations. The minimization of those two prediction errors is realised through the minimization of EFE and VFE. This formation allows for active inference to form realistic neuronal messaging schemes. In active inference, Dopamine is modelled as a hyper-parameter (γ) regulating the prior confidence about EFE estimates. It is also referred to as the precision of EFE, which is changed when the observation is inconsistent with the EFE.

3.4 Markov decision processes

Active inference is often applied to a type of generative model called Markov decision process (MDP) (see Figure 3) to model a wide range of psychological phenomena to do with temporal processing in the brain [Paletta et al., 2004, Smith et al., 2022, Kuperberg, 2021]. There are often four main ingredients in a typical MDP: 1. states (s), 2. actions (a), 3. rewards (r) and 4. transition probabilities (p). In an MDP (see Figure 3), the agent makes a sequence of actions. An action at each step maps to a state of the process, which leads to a reward. The state from the current step transits to a state in the next step with a transition probability p. Note that MDP has a property of Markov processes that the actions and rewards are only directly related to the current states, not the future or the past ones.

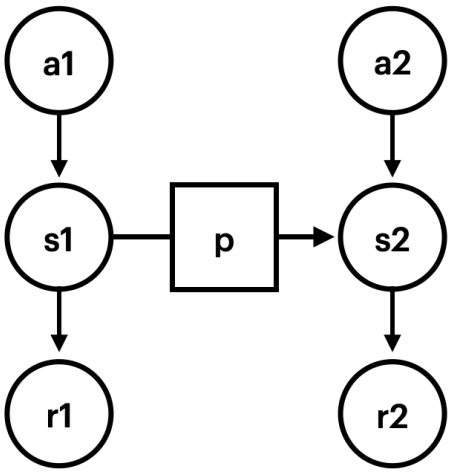

Figure 3: Markov Decision Process. a: action; s: state; r: reward; p: transitional probability. The graph shows a unit of a markov decision process, where the agent makes a sequence of actions a_1 and a_2. An action at each step maps to a state of the process s, which leads to a reward r. The state from the current step transits to a state in the next step with a transition probability p.

.

An MDP for active inference (see Figure 4) is a partially observed MDP, where the agent does not have complete access to the true state of the environment and needs to rely on observations to infer the current state. This corresponds to real-life situations where we are uncertain about the state of the world and need to infer how likely it is to be in a state based on observations [Smith et al., 2022]. Because of this setting, the process became non-deterministic and the state variables are probabilistic. The outcome of the states are observations, instead of rewards, which the algorithm tries to optimise to match the preferred observations. On top of that, actions, which have deterministic relationships with the states, are substituted by policies, that choose an action for each time step. They affect the transition probabilities

between states. To summarize, in a typical active inference MDP, observations/action outcomes depend on the hidden states, which depend on policy and the previous state (Figure 4).

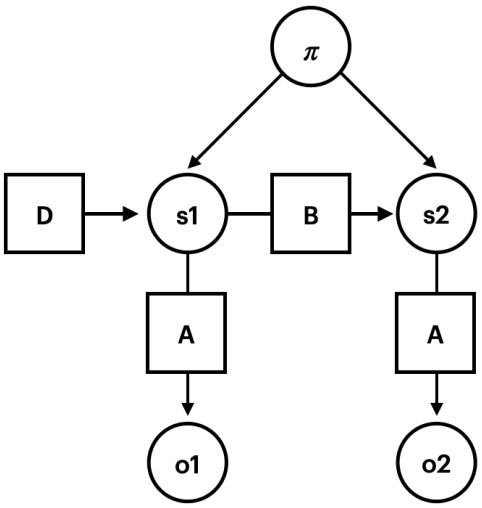

Figure 4: The generative model for Markov Decision Process with active inference.
π: policy, a sequence of actions to achieve a desired outcome; D: prior; s: state posterior; B: transition probability; A: likelihood; o: observation. Different from a vanilla MDP, the MDP in this graph is partially observed, meaning the agent's knowledge about the states is accumulated through observations. Here we can see that the states lead to observations with likelihood A. The initial state s_1 is co-determined by the prior D and the policy π and is updated by observation o_1. The second state s_2 is co-determined by the initial state with transitional probability B and the policy, and it is updated by observation o_2

.

To update the state variable in this framework, the following equation applies:

$$S_{\pi,\tau} = \sigma\left(\ln \mathbf{B}_{\pi,\tau-1} S_{\pi,\tau-1} + \ln \mathbf{B}^{\mathrm{T}}_{\pi,\tau} S_{\pi,\tau+1} + \ln \mathbf{A}^{\mathrm{T}} o_{\tau}\right) \tag{11}$$

Here S is the probability distribution of states given an action π at a time step τ. B is transitional probabilities from either the past or the future. A is the likelihood probabilities of an observation given a state. $\sigma()$ is the softmax function to transform/normalise the term inside into a probability distribution. The input to the softmax function is a sum such that: the first term $\ln \mathbf{B}_{\pi,\tau-1} S_{\pi,\tau-1}$ is the prior from the previous time point $\tau - 1$; the second term $\ln \mathbf{B}^{\mathrm{T}}_{\pi,\tau} S_{\pi,\tau+1}$ is the prior from the future time point $\tau + 1$; the third term $\ln \mathbf{A}^{\mathrm{T}} o_{\tau}$ is the likelihood of the observation at the current time point τ.

Active inference in MDP can be summarised in Figure 5. During state inference, we can infer the current state given the observation through minimizing VFE, the difference between the estimated posterior and the generative model. We can then predict the current observation given the inferred state and compare it to the actual observation, with which we can update the generative model. During policy inference, we first infer the future states based on the current state, from which we calculate the EFE for each future state. Then we select the policy that minimizes the sum of the EFE for all time points that we aim to predict. Note that the policy we select will, in turn, affect the state inference for the next round.

3.5 Generalised predictive coding

Generalised predictive coding put the active inference model in the context of dynamical systems, where the hidden states are presented by their instantaneous first and higher-order derivatives (generalised motions) [Yu and Friston, 2014, Hijne, 2020]. This framework is relevant here because of its emphasis on temporal trajectories, therefore mentioned in some studies to situate WM under PC. Readers who would like to know more can refer to Friston et al. [2011]

To perform *hierarchical* inference, more layers are added to the simplest model mentioned above, where the state variable at the lower layer would serve as the input to the higher layer (Figure 6). This enables the model to build up

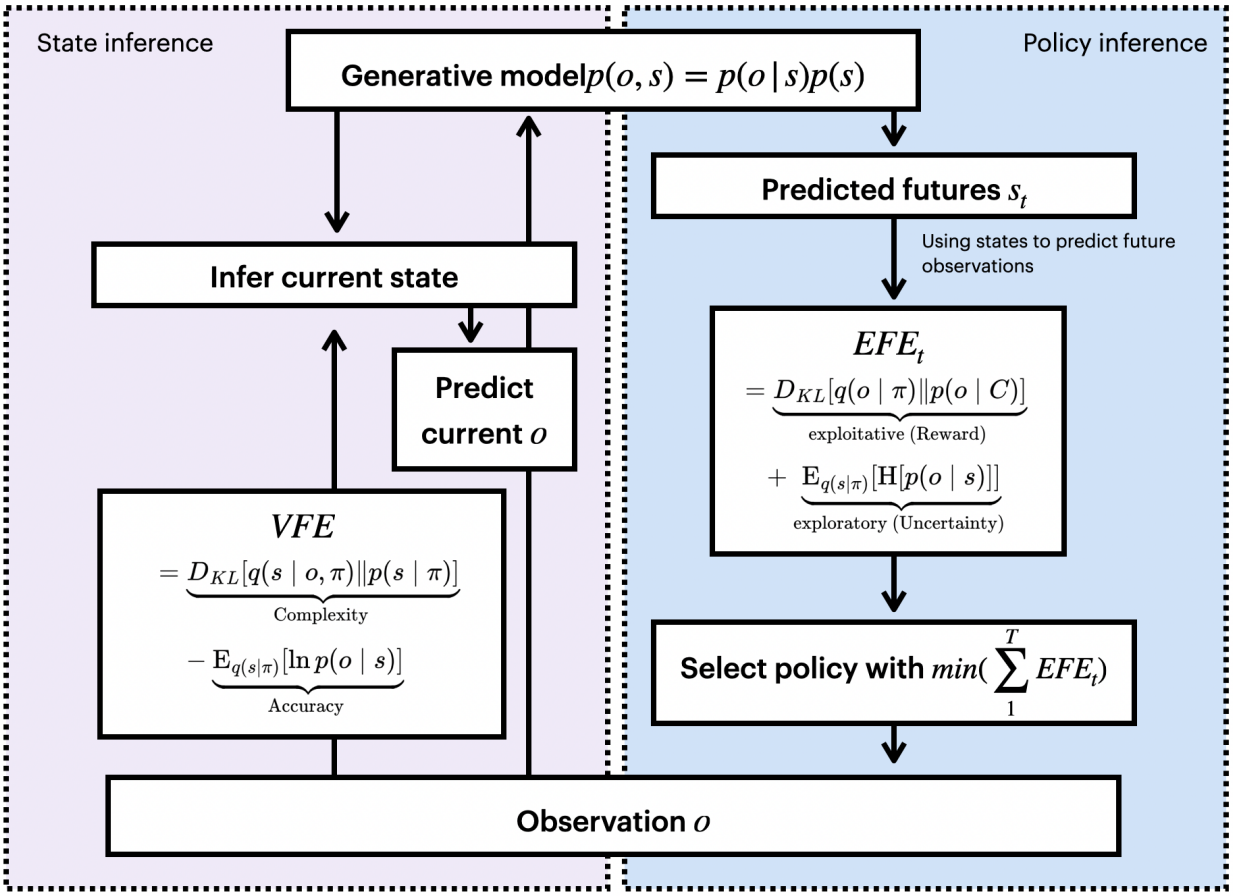

Figure 5: Active inference in Markov Decision Process (MDP).
o: observation; s: state; VFE: variational free energy (see Equation 6); EFE: expected free energy (see Equation 10). This diagram illustrates how states and policies are inferred during active inference using a generative model $p(o, s) = p(o \mid s)p(s)$. On the **state inference** side (left panel), the agent receives an observation o and minimizes VFE to infer the current state s. This process involves comparing the posterior estimate with the generative model. The inferred state is then used to predict the current observation $\hat{o}$, which enables updates to the generative model based on the mismatch with the actual o.On the **policy inference** side (right panel), the inferred state s is used to generate **predicted future states** s_t for future time steps $t = 1, \ldots, T$. These states are used to predict future observations, from which the **expected free energy** EFE_t is computed for each t. The agent then selects the policy that minimizes the cumulative expected free energy over the time horizon, i.e., $\min \sum_{t=1}^{T} \mathrm{EFE}_t$. This leads to a new observation o, and restarts the cycle of inference.w

.

inferences on top of each other and allows for different levels of abstraction [Spratling, 2017, Dora et al., 2021] and timescales [Smith et al., 2022]. Hierarchical PC has been shown to explain neural activities along cortical hierarchy in the human brain during image recognition and speech processing [Dora et al., 2021, Caucheteux et al., 2023].

4 Bridging WM and PC

To remind the readers, we plan to explore whether PC can help with answering four research questions about WM. They are:

1. How is working memory maintained and updated?
2. What is the relationship between attention and working memory and how do they interact?
3. Why does working memory have limited capacity?

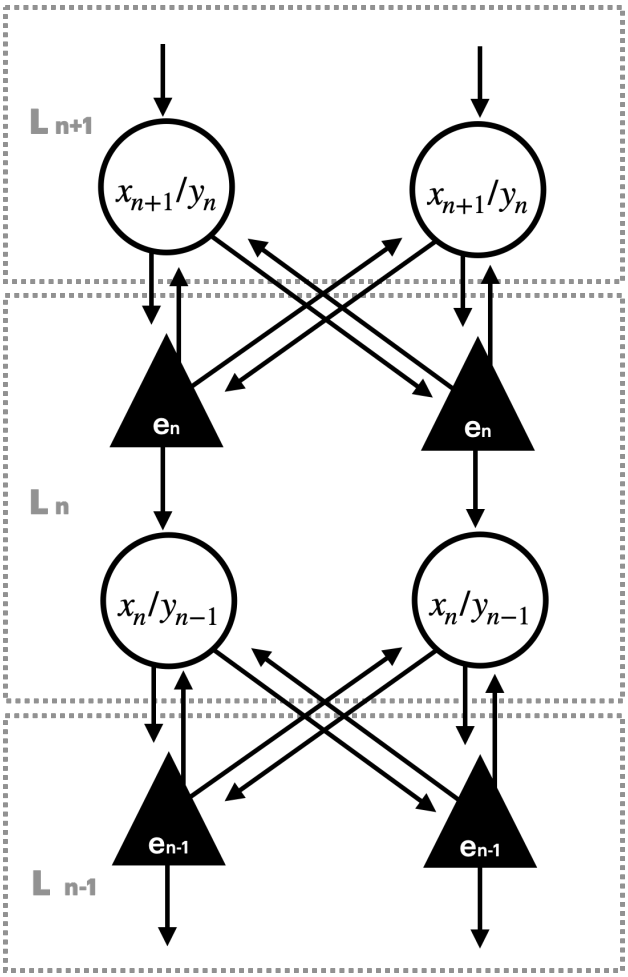

Figure 6: Hierarchical predictive coding. L: level; x: input; y: prediction; e: error. The subscripts indicate levels. The circles are prediction/input neurons and the triangles are error neurons. This figure shows three levels of inference, where the state y at a lower level serves as an input to a higher level. At each layer, the error neurons receive prediction from above and input from below and use them to compute prediction errors. It then sends the errors to the higher layer to update the prediction

.

4. Why is working memory hierarchical?

To answer the questions, we first introduce the predictive nature of WM and then walk through two examples of PC models that incorporate WM. This should familiarize readers with the necessary background knowledge on model structures and experimental settings. Then we discuss the questions one by one in Sections 5.1-5.4. We review models and experimental evidence that answer the research questions.

4.1 WM and predictions

Predictions are at the centre of PC framework. One of the most obvious links between WM and PC is how WM is involved in making predictions or maintaining predictions. We would like to make a clarification of what is meant by 'predictive' and 'prediction' in PC. In a broader context, making a prediction might be understood as making an assumption about what will happen in the future. In PC, however, predictions do not put time stamps on the inferred states. They simply mean assumptions about incoming observations based on an internal model the agent has about the world $p(o, s)$. The prefix 'pre-' can be interpreted as preceding the comparison with the real sensory input. Being 'predictive' means being dependent on the inferred states of the internal model. Under this definition of 'predictive', perception is predictive in the sense that current sensory inputs are being predicted. WM can be seen as an extension

of perception that also predicts the future, hence anticipatory [Yu and Friston, 2014], or the past, hence retrospective. Furthermore, such prediction is subject to the influence of action planning.

WM may be involved in making predictions in two ways [Trapp et al., 2021]. One is acting as a prior for sensory inputs $p(o)$. From this perspective, it can be understood as long-term memory (LTM) highlighted and retained temporarily for a task. WM can also be involved in prediction-making as the posterior belief $p(s|o)$, directly indicating the latent causes s of sensory input o.

After the prediction is made, it has to be maintained for a brief period to facilitate tasks later on. One piece of evidence that predictions are maintained in WM comes from the study performed by Zhang et al. [2019]. In the experiment, mice are motivated to explore novel environments in a T maze, which requires them to remember the arm they went to in the sampling phase and to go to the opposite arm in the trial phase. It was shown that brain oscillations indicating prediction errors appeared after a delay period in WM tasks when the mice chose an alternative route over the one they chose during the initial phase. This suggests that the prediction about the route mice chose was maintained in the WM and resulted in a prediction error when an unexpected outcome appeared.

In active inference, predictions are also made on actions in order to reach the goal (preferred observation $o|C$) in the future by minimising expected free energy G (one can refer back to Equation 9 to see how

$$G$$

is computed). This is supported by current inferred states maintained in WM in order to predict the future outcome of different actions. One can see from Figure 5 of how inference of current state affect the prediction about future state by affecting the generative model $p(o, s)$.

4.2 Prediction errors

In PC predictions of the sensory input and actual sensory inputs are compared, from which prediction errors are generated and propagated upward the information flow (Figure 6). The prediction errors help with updating beliefs about the state of the world, which is similar to what happens in WM updating (Section 2 and Figure 2). In a model for event perception, Radvansky and Zacks [2011] (see Appendix A for details) proposed that the event model in working memory is updated based on prediction errors. It is assumed that a threshold must be surpassed for this update to occur. When prediction error inputs are transient or fall below this threshold, the event model remains dominated by prior information and is not updated. The connection between prediction errors and WM is also shown by neurobiological studies where dopamine activity encodes both prediction errors and WM [Sarno et al., 2022, Yu and Friston, 2014].

5 Answering four questions about WM with PC

5.1 WM maintenance and updating

- Question: How is working memory maintained and updated?

Several studies have proposed that PC can support flexible updating of WM. That is, depending on how much a stimulus is anticipated, WM will be either maintained or updated. In the previous section, we introduced that PC can serve as an implementation of Bayesian inference. Under this setting, one can see WM as the inferred state of the world, which is maintained temporarily for predicting future states and updated as new observation is made (see section 3.1). Below we will give several examples of how this can happen.

In the generative models of event perception proposed by Kuperberg [2021], the *event model* is the sequence of already observed events containing the temporal-spatial information of the events. The event model is thought to be held in WM, providing context for predicting future events [Kuperberg, 2021, Radvansky and Zacks, 2011]. [5] The event model can be changed when prediction errors exceed a threshold, i.e. a gating mechanism, which explains WM updating. This gating mechanism helps with fighting against noise and can be realised by phasic dopamine regulation in the brain which was said to respond to only salient unexpected events [Redgrave et al., 2008].

To account for the selective updating of WM, Yu and Friston [2014] proposed that WM emerged as beliefs of the states of the immediate future. Under this assumption, WM updates can be explained by the switching of anticipatory evaluation about a set of prior beliefs. Yu and Friston [2014] performed an experiment to investigate how anticipation modulates WM update and maintenance. They designed a task which is similar to the one described in 10.2, where

[5] For those who are interested in more details of event models and there relevance to modelling WM in PC, please refer to Appendix A for a summary of the model of event perception.

subjects need to keep visual objects in WM and respond to probe cues. In particular, he added an anticipatory/ predictive cue (see Figure 2) at the beginning of the trial to indicate whether there would be an update of the initial stimuli later in the trial. This is to induce anticipation about what comes next and to test whether WM performance will be affected by the anticipation. They collected fMRI data from the participants during the task performance.

The results of Yu and Friston [2014] support the idea that WM follows hierarchical inferences in the brain. Specifically, the results demonstrate that top-down anticipation can modulate WM to maintain or update information. Sustained neural activities representing WM were found to be associated with the anticipatory set, which is regarded as the signal of predictive processing. Yu and Friston [2014] furthered his investigation to find plausible cortical message passing in WM network for implementing hierarchical inference. Using dynamic causal modelling (DCM), the authors tested hierarchical cortical organisations against fMRI data. The results are consistent with the message-passing scheme proposed by PC framework, in which anticipation is supported by top-down connections while surprise is mediated via bottom-up connections (where the authors assumed that DCM performed this test). Furthermore, the authors identified that the dopaminergic midbrain and the striatum were associated with anticipation and thus played a role in WM updating. Those results provide anatomical grounds for the implementation of PC during WM tasks in the brain.

One step forward is the model of saccadic control proposed by Parr and Friston [2017].[6] The model purposes working memory to be the change of posterior probability of the hidden state ($p(s|o, \pi)$). The dynamics of WM can therefore be understood as the process of evidence accumulation. Specifically, when new evidence shows up, the posterior is updated (Equation 2), reflecting a WM update. When a participant is shown a retro-cue of a scene that would probably be tested (see Figure 2), the uncertainty about which scene to be tested is reduced, and the posterior of the corresponding hidden state is updated where the posterior probability of the cued scene increases. When there is no new evidence, for example, during a delay period, the inference process still continues and the posterior remains the same. This allows WM to be maintained. Thus, in the context of goal-directed behaviour, PC naturally explains WM maintenance and updating such that when a state is of relevance to a given goal, predictive processing would be activated and keep the state representation up-to-date.

This WM updating mechanism differs from the event model by Kuperberg [2021] in that it did not impose a threshold for the updating. If WM is modelled with probabilistic representation, it would continuously change over even minor evidence. This seems to go against the fact that WM by its nature is robust to distractors to some degree in order to be maintained at least for a short period of time [Feng et al., 2023, Mejías and Wang, 2022, Chumbley et al., 2008] and needs further consideration. One possibility is that WM is instead represented by the posterior state estimates, which will only be updated when the accumulated evidence is big enough to change the maximum a posteriori estimate.

During WM maintenance and updating, WM also decays. In the model of saccadic control, Parr and Friston [2017] attempted to explain forgetting by the volatility of the state transition probability. They predict that the more the subject believes the environment is volatile, the easier/quicker their WM decay.

Finally, research in artificial intelligence (AI) also shows the maintenance and manipulation of WM emerging from networks incorporating PC. In a study of robotic control, actions are planned through visual simulation for a robot arm to stack three coloured boxes in a certain order [Jung et al., 2019]. WM is modelled explicitly as a separate module to preserve long-term visual information to predict future action outcomes. The WM module keeps the information about the position of boxes and is updated through its connection with other modules of the PC network. We expect that large Artificial Neural Networks may allow for investigating the role of WM in more complex tasks under PC frameworks in the future.

5.2 WM and attention

- Question: What is the relationship between attention and working memory and how do they interact?

In the context of PC, attention and WM are both related to predictions. In a thesis on selective updating of working memory, Yu and Friston [2014] pointed out that attention highlights information to make predictions about the current situation, while WM, on the other hand, maintains information based on the predictions of its relevance to the future. Thus the theory by Yu and Friston [2014] sets attention and WM apart by temporal characteristics.

The different roles of WM and attention and their interaction in PC frameworks were further clarified by Parr and Friston [2017], who proposed an MDP model with active inference (readers can refer back to Section 5 for a discussion of MDP). They divided attention into two different categories. One refers to the gains assigned to certain sensory channels, as a result of the inferred precision of the states/causes given the sensory channels [Feldman and Friston, 2010]. This is modelled as a precision parameter γ, which regulates the degree to which the EFE for each channel

[6] For those who are interested in more details of the saccadic control model and its relevance to modelling WM in PC, please refer to Appendix A for a summary of the model.

controls policy selection (γG). Another category of attention refers to salience, which refers to the probability with which an item will afford an action. In the case of saccadic control, the more salient a location is, the more likely it will be foveated [Parr and Friston, 2017]. This probability is dependent on how it can reduce the uncertainty about the predicted observations given states, which corresponds to the explorative term of EFE ($\mathrm{E}_{q(s|\pi)}[\mathrm{H}[p(o \mid s)]]$, see also Equation 10). So, salience can be seen as sampling the world that maximizes its epistemic affordance. In the case of saccadic movement, salience would be determined by how much uncertainty can be reduced by making a saccade on a location. This undercertainty can be modelled as the the belief about a policy, $p(\pi)$.

Under those definitions, WM, if modelled as the state variable in active inference, clearly has a different role from attention, yet interacts with it (Figure 7). WM is affected by attention because attention is involved in selecting policies to collect evidence (new observations) for belief updating. On the other way around, WM (inferred state posterior) affects attention (precision) by updating beliefs on policies ($p(\pi)$) [Smith et al., 2022]. This is done through the minimization of VFE and EFE. We can see that the posterior probability distribution of policies is at the centre of this reciprocal loop. In other words, action selection establishes the interaction between WM and attention.

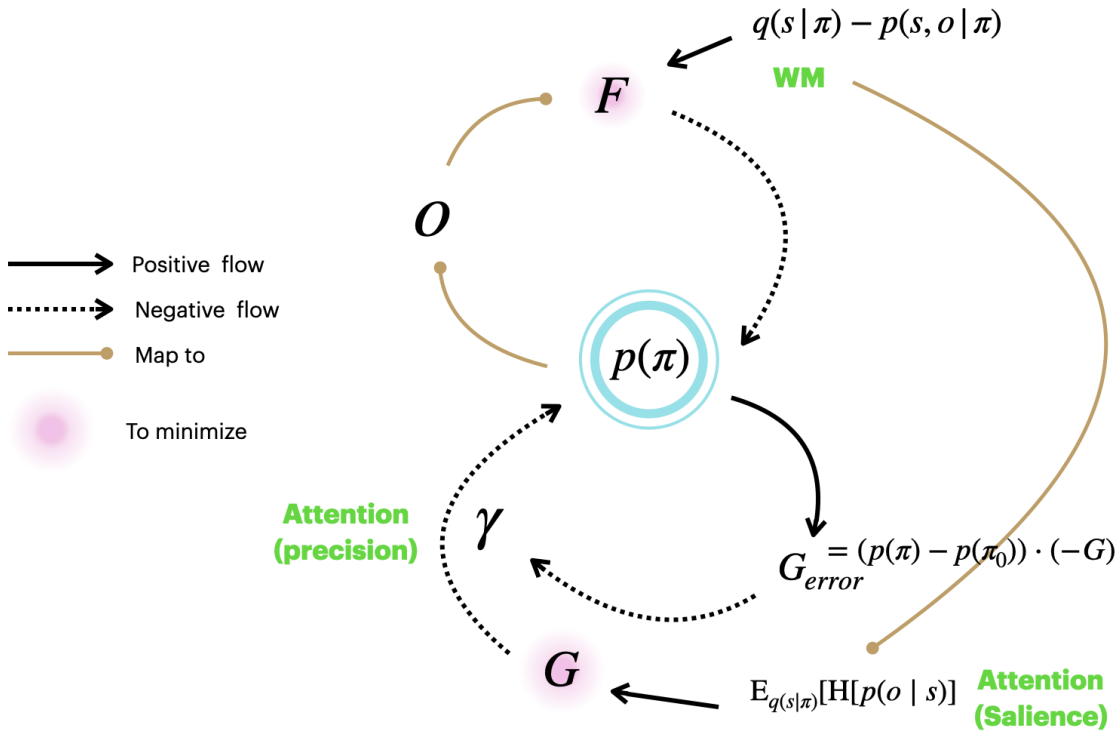

Figure 7: Interaction between WM and attention.
$p(\pi)$: belief about policy; o: observation; s: state; F: variational free energy; $q(s|\pi)$: estimated posterior given a policy; $p(s, o|\pi)$: joint distribution of s and o; G: expected free energy (EFE), G_{error}: prediction error of EFE; $p(\pi_0)$: prior belief about policy; $H[p(o|s)]$: entropy of $p(o|s)$; γ: confidence on EFE (precision). The arrows show how one quantity affect another (positively or negatively). The brown connections indicate causal map from one variable to another, e.g. a certain $p(\pi)$ will lead to a certain o. This figure shows the reciprocal relationship between WM and attention mediated by policy selection. WM is affected by attention because attention is involved in selecting policies to collect evidence (new observations) for belief updating. WM (inferred state posterior) also affects attention (precision) by updating beliefs on policies ($p(\pi)$) [Smith et al., 2022]. This is done through the minimization of VFE and EFE. We can see that the posterior probability distribution of policies is at the centre of this reciprocal loop. In other words, action selection establishes the interaction between WM and attention.

5.3 WM capacity limit

- Question: Why does working memory have limited capacity?

A functional explanation of WM capacity limit given the predictive nature of WM was recently proposed by Trapp et al. [2021]. The authors argued that there is a trade-off between predictive accuracy and capacity/complexity, and specifically the number of items to retain has to be limited to achieve a certain level of accuracy. This limitation coincides with the setting of four to six temporal steps into the future for the simulations in deep temporal models from their study. It was also mentioned that the generalized coordinates of motions (the derivatives of motions, that is position as the 1st order derivative, speed as the 2nd order derivative etc.) are usually truncated at about the fourth order for the same reason.

The prediction of the future can be modelled as a random process with the probability of the future event given the current event being a constant ($p(\text{future event}|\text{current event}) = c$). It seems intuitive that when we predict further into the future, the probability decreases, meaning the accuracy of the prediction will also decrease. This set a limitation on how many steps to take for prediction. Trapp et al. [2021] argue that if the sequential representation plays a dual role in both retrospective and prospective memory (more discussion see 6.2), the limitation on prediction accuracy will be reflected in the number of steps that can be stored for the past and the future.

The assumption provides an elegant explanation to account for both capacity limitations of past and future. Though it does not seem to be necessary to hold onto the assumption of a MDP to have only one single sequential representation playing a dual role. If there are two copies of the representation, there should not be a limitation for WM capacity for past events. This is because the uncertainties about the future have been resolved when observations were made. Also, the theory of accuracy-complexity trade-off seems to only work for sequential items, but since the current model is designed only with temporal depth, it does not explain why the capacity limit also applies to parallel items.

To further validate the hypothesis of Trapp et al. [2021], researchers should test the link between predictive and memory capacity experimentally, e.g. by behavioural tests on both predictive and WM tasks. Researchers would then draw inference as to whether stronger predictive capacity is accompanied by stronger WM capacity. [7]

It should also be noted that the WM capacity limit measured in experiments might be a result of a strong violation of the experimental setting against prior beliefs [Orhan et al., 2014]. For example, participants' prior belief might be that the items presented in different trials are dependent when, in reality, the items across trials are actually independent. This mismatch leads to biases and inefficiencies in memory encoding and retrieval, thus potentially limiting observed WM capacity.

Indeed, if WM is seen as a result of Bayesian inference, it will not reflect the observations if there is a strong belief against the state itself ($p(s)$), or about the state of the world that it can hardly generate such observations (the likelihood $p(o|s)$). The effect can happen more easily in an experimental setting because it can be unnatural, and the experimental setting is counter to prior beliefs the agent developed in their life. In active inference framework, this may again be explained as a trade-off between complexity/energy and accuracy (Equation 6), where a large change of belief needed to account for an observation would require compromises in encoding accuracy.

In summary, PC can explain WM capacity limit in terms of the trade-off between the complexity and accuracy of the information being encoded. This provides an alternative theory to the slot model[Cowan et al., 2013], the resource model [Bays et al., 2009] and the interference model [Oberauer and Lin, 2017] to account for the limited capacity of WM.

5.4 Hierarchical WM

- Question: Why is working memory hierarchical?

Hierarchical information processing relies on top-bottom interactions. The interaction between top-down and bottom-up information flow in PC allows comparisons between predictions and sensory inputs and allows for updates of the internal model. The top-bottom influence of brain oscillations has also been discovered during WM updating, which coincides with those found for PC [Miller et al., 2018, Zhang et al., 2019].

In a task of visual WM of orientation Berlot and de Lange [2022], it was found that the moving dot and the grating stimuli are both re-coded as line-like representations. This suggests that the stimuli were efficiently coded into a format that serves the goal of the task, which reflects how PC infers the state of the world from top-bottom interactions (in this case, the orientation).

[7] As a side note, more evidence is needed to show that prospective and retrospective WM memories are indeed represented by the same neural component.

Framework	Task	computational roles	Variables	Neurobiological signatures	References
Active inference	Explore-exploit	Belief updating	$p(s\|o,\pi)$	Neural activation	[Smith et al., 2022]
Predictive coding	Event perception	Event model	Not formally defined. Can be thought of as p(s,o),	N.A.	[Kuperberg, 2021]
Generalised predictive coding	WM updating task (with anticipatory cues)	Beliefs about the immediate future	Not formally defined. Can be thought of as $p(s\|o,\pi)$ and its generalised motions	Dopamine	[Yu and Friston, 2014]
Active inference	WM updating task	Evidence accumulation/beliefs	Not formally defined. Can be thought of as $p(s\|o,\pi)$	context: hippocampus; locations: parietal cortex	[Berk Mirza et al., 2016]
Active inference	WM updating task	Evidence accumulation/beliefs	Not formally defined. Can be thought of as $p(s\|o,\pi)$	context: hippocampus; locations: parietal cortex	[Parr and Friston, 2017]
Active inference	Undefined	Bayesian model averaging	$p(s\|o,\pi)$	Matrix thalamocortical circuits	[Friston et al., 2018]
Generalised predictive coding	Undefined	Belief trajectories	$p(s\|o,\pi)$ and its generalised motions	N.A.	[Friston, 2008]

Table 1: Summary of WM modelling in PC frameworks

Studies show that higher-order information bias the inference of the stimulus feature, where top-down conditioning may play a role in maintaining or restoring WM over noise. In a psychophysical study by Brady and Alvarez [2011], the memory of the size of individual circles presented was found affected by the mean size of all the presented circles, indicating multiple levels of abstraction. In another psychophysical experiment, Luu and Stocker [2021] showed that categorical and detailed feature information are both stored in WM, which can be manipulated separately.

Hasson et al. [2015] proposed a hierarchical WM framework where the temporal receptive window increases with hierarchy, and thus higher areas process information at a larger timescale. There is also some neuroscientific evidence for the existence of this framework such that there is the hierarchical topology of WM according to the measurement of fMRI, electrocorticography and single-unit recording during auditory and visual processing of temporal information [Honey et al., 2012, Lerner et al., 2011, Hasson et al., 2008].

Those different levels of WM representations can naturally fit into hierarchical PC, with WM at each level being the result of state inference at that level. It can explain the effect of top-down conditioning on WM by the influence of hidden states at a higher level to the hidden states at a lower level.

6 Modelling WM in PC

This section is aimed at providing guidance and stimulating discussions on modelling WM in PC. We first introduce WM as posterior beliefs and summarize the ingredients that are important in PC frameworks to model WM. Then we discuss a few open questions on modelling choices, together with a comparison to some mechanistic models of WM. We then discuss how modelling WM in PC frameworks can contribute to applied topics such as better schizophrenia treatment and more robust AI. Finally, we talk about the general limitations of PC frameworks.

6.1 Three ingredients

Table 1 contains modeling strategies to account for WM using PC, conceptually or formally. In summary, WM is often modelled as part of the existing PC, for example, using state variables. This means that the maintenance of WM during a delay period is not a cause of active retention but is a side product of inference processes. This may seem counterintuitive because researchers often presume that the representation of the sensory input will fade away right after receiving a stimulus, and thus we need a mechanism to actively support WM maintenance. However, under PC framework, inference itself is an active endeavour which continuously happens. In such a framework, not only is a changing stimulus an input, but a unchanged stimulus also is. When a un-changed stimulus is presented, the belief remains the same after the inference, therefore manifested by a persistent activity.

If we see perception as an inference of a moment, what we need to distinguish WM from perception is that WM is a lasting inference process that is driven by a goal. This is in contrast with some biophysical WM models using a dedicated canonical neural circuit to explain persistent activity [Curtis and Sprague, 2021]. The passiveness of WM maintenance is consistent with the emergent theory of WM, where WM is an emergent property of other functions of the brain.

Fig 8 contains a synthesized example of the PC framework that can be used to model WM, where three ingredients are included, which are temporal depth, goals and hierarchy. The temporal depth is modelled as chain of states s_n; the goals are modelled in the expected free energy G; the hierarchy is modelled as two levels of MDPs where the top one guide the bottom ones.

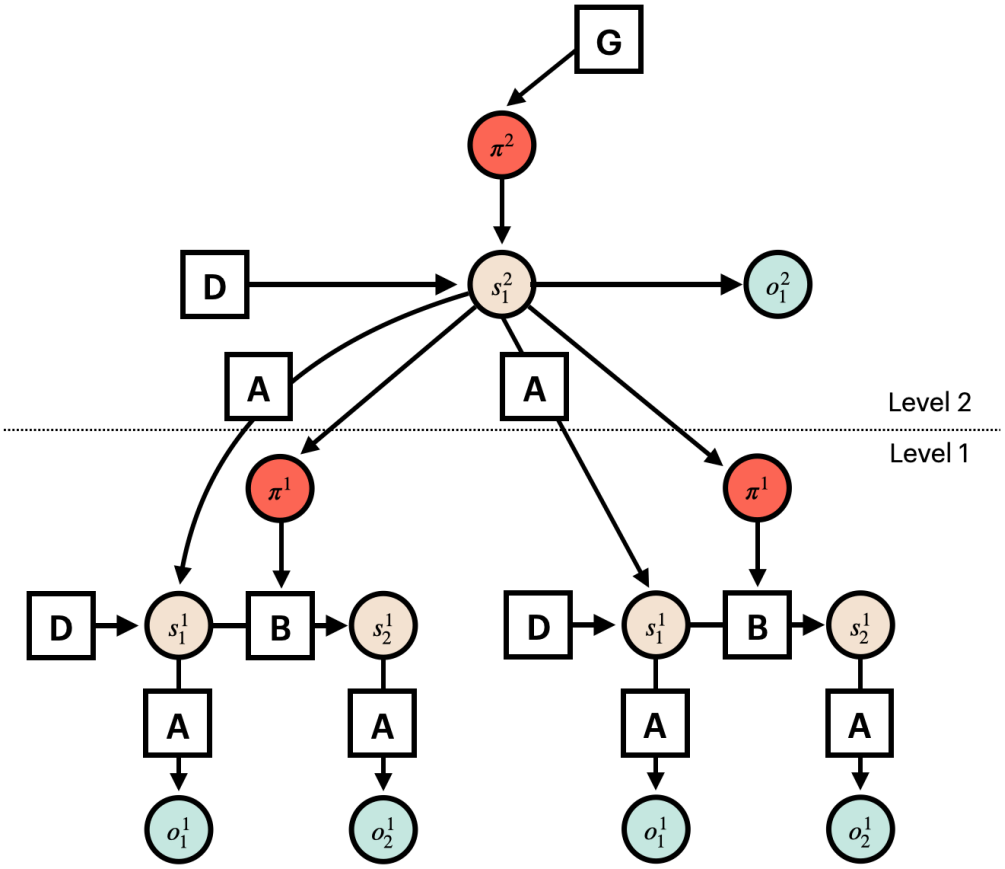

Figure 8: A deep temporal MDP framework to model WM in PC. Two levels of inference are shown in the graph. G: expected free energy, where preference is integrated (see Equation 9); π: policy; D: prior; s: state posterior; B: transition probability; A: likelihood; o: evidence. The superscripts indicate the level of inference. The subscripts indicate the time points.

.

6.2 Ingredient 1: Temporal depth

One of the most important ingredients in this framework is to accommodate temporal depth. As an example, this can be done by using a Markov chain to model the accumulation of evidence. Modelling temporal depth also enables the modelling of both retrospective and prospective influence on WM.

By modelling WM in PC frameworks, both beliefs about the past and future are relevant. Perception is only about inferring the state given the current observation. Working memory, on the other hand, is goal-directed, which requires inferring the states about either past or the future given the current state. To make this more concrete, imagine a situation where you came to a new place at night and stayed until the morning. When the sun rose, you saw that there was a tree in front of you. Now you might infer that there was probably a tree the night before and there will probably be a tree in the future. Depending on whether the task is to keep a diary about yesterday or climb the tree, the belief about the past or the future will be maintained in your working memory. One can refer to the saccadic control task from Parr and Friston [2017] as an experiment on the belief about the past and the explore-exploit task in [Smith et al., 2022] as an experiment on the belief about the future. In a formal model, this means working memory at $t = 0$ can be the state variable at $t = 1$ or $t = -1$. In generalized PC, this is pushed one step forward, where what is inferred is not a state at a single time point, but its trajectory.

6.3 Ingredient 2: Goal-directed behaviour

Working memory (WM) plays a crucial role in enabling goal-directed behaviour by retaining information that is relevant for selecting future actions. In the context of active inference, goals are formalised as preferences over outcomes, and the minimisation of expected free energy (EFE) drives action selection to satisfy these preferences. WM facilitates this process by maintaining beliefs about the current and future states of the world that are relevant for policy evaluation.

In this framework, WM is not merely a passive buffer, but an active inference mechanism supporting decision-making. Specifically, WM retains beliefs about the hidden causes of sensory input (posterior over states), which in turn determine the expected outcomes of different policies. Policies that minimise EFE are selected based on both epistemic value (uncertainty reduction) and pragmatic value (goal satisfaction) (see Equation 9).

This formulation implies that WM content is modulated by task demands. For instance, when a preferred outcome changes, the belief state held in WM needs to be updated accordingly. This dynamic was illustrated in the WM updating models discussed in Section 5.1, where WM content changes based on predictive cues indicating potential updates.

Furthermore, dopaminergic signalling, modelled as precision over EFE, plays a key role in modulating the salience of goals and thus the content of WM [Yu and Friston, 2014, Smith et al., 2022]. This establishes a link between the motivational significance of an outcome and the selective retention of information in WM, enabling adaptive goal-directed behaviour in uncertain environments.

Thus, in a PC framework, WM is inherently teleological: it encodes beliefs in service of expected outcomes, constantly shaped by the inferred value of information relative to future goals.

6.4 Ingredient 3: Hierarchical structure

A defining feature of WM is its ability to represent information at multiple levels of abstraction. This hierarchical organisation aligns naturally with the hierarchical structure of predictive coding (PC), where inference is distributed across multiple layers corresponding to increasing temporal and conceptual abstraction.

In PC, higher levels encode slower dynamics and more abstract features, while lower levels encode faster, more concrete sensory details [Friston, 2008, Hasson et al., 2015]. WM representations can emerge at any level of this hierarchy, depending on the task. For example, in visual WM tasks, early visual cortex may maintain low-level sensory details, while prefrontal regions may encode categorical or semantic information relevant for decision-making [Luu and Stocker, 2021, Berlot and de Lange, 2022].

This architecture allows WM to flexibly transform between different representational formats, such as between visual and verbal codes, depending on contextual demands. Moreover, top-down influences from higher levels can stabilise lower-level representations, helping to maintain WM in the presence of noise or distractors [Zhang et al., 2019, Brady and Alvarez, 2011].

Experimental findings showing that WM contents are shaped by higher-order knowledge—such as ensemble statistics [Brady and Alvarez, 2011], categorical context [Luu and Stocker, 2021], or task goals [Berlot and de Lange, 2022]—can be accounted for by hierarchical message passing in PC. Under this view, WM is not a fixed capacity store, but a dynamic inference process where representations at different levels of the hierarchy constrain and refine each other.

Importantly, this hierarchical structure also supports generalisation and abstraction. For instance, learning a schema at a higher level can facilitate WM performance on related but novel tasks by reusing abstract predictive structures.

Therefore, modelling WM within hierarchical PC not only explains behavioural and neurophysiological findings, but also provides a principled account of how WM can bridge perception and action across multiple timescales and levels of abstraction.

6.5 WM as posterior belief in active inference

To model WM in PC, another important question we might ask is what variable it represents in a PC framework. In most of the models we review here, WM is modelled as the posterior belief about states of the world ($p(s|o)$). This modelling choice can reproduce behavioural results and neural activities during the tasks, as mentioned previously [Honey et al., 2012, Lerner et al., 2011, Hasson et al., 2008]. It also explains how WM is tightly incorporated into goal-directed behaviours. That is, in order to reach a goal (preferred observation), an agent would need to first gain some information about the state of the world. For example, if one needs to grasp something in a dark room, one needs to first turn on the light to know where the item is (the state of the world). Thus WM emerges from a process of evidence accumulation to reduce uncertainties about the state of the world.

In the case of saccadic control, eye movements are controlled to reduce the uncertainty about what was initially presented, e.g. to allow the participants to choose the correct cue at the end of an experimental trial. The beliefs about hidden states of initial scenes are updated following the observations after selected saccades to reduce the uncertainty about the states.

7 Open questions

Although many studies discussed the relationship between WM and PC, formal models are not often presented, leaving challenges and opportunities for future research.

7.1 Is WM a separate component?

The first remaining question is whether to model WM as a component separated from the inference process. This question has raised a debate between the modular and emergent view of WM [Postle, 2006]. Most of the models we previously reviewed do not separate WM from the rest of the components of the PC. Very often, it is modelled as the traces of the state variables. Some of the other probabilistic models of human WM incorporating Bayesian inference, instead, model it as a separate module [Sims et al., 2012]. Models in artificial intelligence also tend to model WM as a separate component [Jung et al., 2019]. As a long debating question, this deserves further attention. For a review see Hasson et al. [2015].

7.2 Which variable represents WM?

Assuming WM is not modelled separately, the next question is what variable represents WM in PC exactly. The answer might sit in between two candidates: posterior probabilities or predictions. A fundamental difference between classical predictive coding and active inference is that classical predictive coding estimates predictions using point estimators (see Equation 1), whereas active inference estimates full posterior distributions over latent states (see equations in Section 3.2).

If working memory (WM) is represented by posterior distributions, its updates are continuous, dynamically reflecting graded changes in sensory input. This is compatible with the continuous nature of synaptic plasticity [Miller et al., 2018]. In contrast, if WM is represented by discrete point estimates, updates may be more categorical—occurring only when environmental changes exceed a certain threshold. This thresholded updating is consistent with bifurcations observed in attractor models of WM, where state transitions occur abruptly once critical inputs are reached [Feng et al., 2023].

Another possibility is that WM is represented by prediction errors, the differences between sensory inputs and predictions. For future work, researchers could compare the activity patterns of the three variables (point estimator predictions, posterior distributions and prediction errors) during a task using existing models and evaluate their differences more systematically. That work could be followed with a more intense review of current experimental measurements of behavioural and neuronal activities during WM, so that researchers can compare the simulation results from PC models with the experimental measurements to establish a map between the variables in PC and different measurements. Researchers might also want to investigate this question from the perspective of neural coding (a review see Ma and Jazayeri [2014]) to see which assumption is more biologically realistic.

In the end, it is possible that WM is a multifaceted entity represented by all the above-mentioned variables (point estimator predictions, posterior distributions and prediction errors). For example, the change of prediction may underlie the sparse spikes observed in experimental data and the continuous state updates may underlie the trace of synaptic weights. PC may provide a solution to resolve the conflict between different theories of WM by integrating different neural and bio-chemical representations observed during WM task (Table 2).

Finally, although most of the literature reviewed here only pays attention to the beliefs about the state of the world, we have to remind ourselves that all variables being inferred can be part of WM. For example, although the volatility of the world often serves as a hyper-parameter, it can also be inferred and updated during a task, which can be maintained in WM.

7.3 Where in the hierarchy?

Another question that requires further attention is where in the hierarchy is WM maintained. Currently, different studies show different results. For example, Luu and Stocker [2021] show that WM is represented at multiple levels while Berlot and de Lange [2022] show WM represented at only the early sensory area. Since WM facilitates action planning,

model	WM representation	Literature	PC account	Tensions	Further integration
classical attractor model	population persistent steady-state firing rates	[Wang, 2001]	state variables	attractor WM has no prediction error layer	
bump attractors	cross-neuron bell-shaped steady-state firing rate	[Wimmer et al., 2014]	state variables	attractor WM has no prediction error layer	continuous states
distributed attractor model	population persistent steady-state firing rates across cortical hierarchy	[Mejías and Wang, 2022]	state variables; hierarchies	attractor WM has no prediction error layer	
silent WM model	brief bursts of spikes and temporary change of synaptic weight	[Mongillo et al., 2008, Trübutschek et al., 2017]	prediction updates and state variables	PC cannot account for refreshing of synaptic weights; NOWM has no PDE	
silent WM network oscillation model	brief bursts of spikes and temporary change of synaptic weight	[Miller et al., 2018]	prediction updates and state variables; top	PC cannot account for refreshing of synaptic weights; NOWM has no PDE	
linear WM or "models without feedback"	population firing rates	[Goldman, 2009]		LWM has no feedback and PDE	
models with cortico-thalamic interactions	population firing rates	[Jaramillo et al., 2019]	attentional regulation (cortical-thalamic pathways)		

Table 2: PC-WM model and mechanistic/biophysical models of WM

it is possible that the level of WM representation is dependent on the length of planning. This implies that in tasks requiring planning over a longer period of time, WM is represented higher in the hierarchy so that the memory can be more robust and thus survive noise for a longer time period. The current PC framework models neuronal activities at all levels. For future theoretical work, researchers might need to envision a mechanism to allow neural representation at selected levels.

7.4 What are the applications for WM in PC? Brain disorders and more robust artificial neural networks?

Because of the important role of WM in both human cognition and artificial intelligence, we expect the efforts of formalizing the link between WM and PC would also help with developing tools to understand and tackle WM-related brain disorders and lead to future AI that can be more human-like or/and be more efficient in solving engineering problems.

There are many brain disorders accompanied by working memory deficits, for example, schizophrenia [Eryilmaz et al., 2016] and ADHD [Ortega et al., 2020]. Since PC can model a variety of cognitive and motor functions, if we model WM with PC, we can get a more holistic understanding of the cognitive mechanism behind brain disorders with WM deficits. This may help with developing better psychological interventions. With PC's specialization in representing hierarchical inference of the brain, it can be more capable of modelling complex tasks than some pure mechanistic models based on first principles, such as the ones mentioned in Table. 2. We can therefore using this advantage to identify more psychological traits of the brain disorders to help with diagnosis and monitoring of the disorder. Another advantage of PC is that it can be easily mapped to neuronal implementations. By establishing neuronal message-passing schemes of PC models, we can map critical parameters and connections in the models to brain chemicals and structures, to help with developing medical or surgical treatment.

In the field of artificial intelligence (AI), deep neural networks have been proven a powerful tool in visual object recognition. However, deep neural networks suffer from a sensitivity to (a small amount of) noise in the inputs [Zheng et al., 2016]. The idea of maintaining a stable WM through a prediction-error-gated mechanism might help AI to achieve more robust performance. Since PC is designed to make inferences in a noisy environment, it should help with this issue. Modelling WM in PC would potentially help AI to deal with uncertainties that bear temporal complexity.

8 Limitations with describing WM with PC

In this review, we present possible explanations of WM provided by PC. However, one needs to be reminded that PC theory itself has its own limitations.

To begin with, whether prediction errors can be represented by neural systems is yet to be concluded, although there is increasing evidence of this phenomenon [Mikulasch et al., 2023]. Since the exact neural implementation is not the focus of this paper, readers interested in the schematics of neural implementation can refer to work by Rao and Ballard [1999], Smith et al. [2022], Kogo and Trengove [2015], van de Laar and de Vries [2019]. For further thoughts, one might look

into probabilistic neural coding theories by Ma and Jazayeri [2014]. To the present authors, it is also doubtful how *all* hidden states can be represented by neuronal activities, considering there can be a large amount of hidden states.

There may also be limitations in terms of the scope that PC can apply to in explaining a cognitive phenomenon. For example, Ransom and Fazelpour [2015] listed three problems of PC accounts for attention. Specifically, they argued that although PC can explain spatial attention, it cannot explain feature-based attention. They also pointed out that PC may not explain non-perceptual attention such as attention to one's own thoughts. Lastly, they mentioned PC cannot explain affective salience or attention in high-cost situations. Those problems likely also hold for how PC accounts for WM as well. Since we only inspected a few examples of working memory tasks, it is possible that some aspects of WM cannot be covered by PC.

9 Conclusion

In this literature thesis, we reviewed how working memory (WM) can fit within predictive coding (PC) frameworks and demonstrated how this arrangement can help with answering four specific research questions on WM:

1. How is working memory maintained and updated?
2. Why does working memory have limited capacity?
3. What is the relationship between attention and working memory and how do they interact?
4. Why is working memory hierarchical?

To answer the questions, we introduced the Markov decision process with active inference as a PC framework for implementing variational Bayesian inference over time. Both theoretical analysis and experimental evidence surrounding this idea have provided important insights for answering those questions about WM.

It is shown that by treating WM as posterior beliefs, we can explain WM maintenance and updating by how beliefs are updated in variational Bayesian inference over time. When WM is seen as predictions/inferred states, the balance between stability and flexibility can be explained by the gating of evidence accumulation on the change of predictions, where predictions are only updated when the posterior probability of one state is pushed above another.

PC explains the relationship between WM and attention through their interactions with action selection. Actions selected by an agent take effect on the environment, leading to evidence which is used to update WM as posterior beliefs about states. With regard to attention, the prediction errors of selected actions regulate the precision of the sensory channels. Meanwhile, the evidence resulting from selected actions updates the likelihood distribution, which changes the salience e.g. of the locations in a visual field. Reciprocally, WM and attention influence action selection through VFE and EFE respectively, therefore, forming a cycle where WM and attention affect each other via the process of action selection.

The capacity limit of WM can be explained by the trade-off between accuracy and complexity in predicting the future. That is, given the evidence, the trade-off between how well the belief predicts the sensory input and the amount of energy required in belief updating to achieve such accuracy. In the case of a MDP, this corresponds to the limitation to the number of time steps one can include to ensure a certain level of prediction accuracy.

Finally, hierarchical PC naturally explains the hierarchical organisation of WM by its hierarchical inference architecture, where WM at different levels of abstraction maps to posterior beliefs at different levels of inference.

To summarize, WM can be modelled in an active inference framework which incorporates temporal depth, goals and hierarchy. With those ingredients, WM supports goal-directed behaviour through the minimization of variational free energy and expected free energy. In this framework, WM is part of the inference architecture, instead of being a separate component. There are several candidate variables in the framework to represent WM. For example, WM can be seen as belief updating/evidence accumulation about the hidden state of the world or the predictions of the incoming observations in the past and future. Activity patterns of different variables accommodate and account for different neural and bio-chemical representations observed during the WM task, which may resolve the conflict between different mechanistic theories of WM. Although the evidence shown here explains WM quite well, it should be noted that we only reviewed a few specific WM tasks, such as event perception and saccadic control. We purposefully only looked at these tasks to keep the review constrained to a few specific examples. For future work, a diversity of WM tasks needs to be taken into consideration in order to generalise the conclusions made here and refine the proposed framework.

Apart from providing scientific insights about WM, modelling WM in PC frameworks also provides potential developments of new computational models of human and artificial working memory, leading to better treatment of brain disorders with WM deficits and AI which is more robust to noises.

References

Long Luu and Alan A. Stocker. Categorical judgments do not modify sensory representations in working memory. *PLoS Computational Biology*, 17(6):1–28, 2021. ISSN 15537358. doi:10.1371/journal.pcbi.1008968.

Uri Hasson, Janice Chen, and Christopher J. Honey. Hierarchical process memory: Memory as an integral component of information processing. *Trends in Cognitive Sciences*, 19(6):304–313, 2015. ISSN 1879307X. doi:10.1016/j.tics.2015.04.006. URL `http://dx.doi.org/10.1016/j.tics.2015.04.006`.

Christopher J. Honey, Thomas Thesen, Tobias H. Donner, Lauren J. Silbert, Chad E. Carlson, Orrin Devinsky, Werner K. Doyle, Nava Rubin, David J. Heeger, and Uri Hasson. Slow cortical dynamics and the accumulation of information over long timescales. *Neuron*, 76(2):423–434, 2012.

Timothy F. Brady and George A. Alvarez. Hierarchical encoding in visual working memory: Ensemble statistics bias memory for individual items. *Psychological Science*, 22(3):384–392, 2011. ISSN 09567976. doi:10.1177/0956797610397956.

Yulia Lerner, Christopher J. Honey, Lauren J. Silbert, and Uri Hasson. Topographic mapping of a hierarchy of temporal receptive windows using a narrated story. *Journal of Neuroscience*, 31(8):2906–2915, 2011. ISSN 02706474. doi:10.1523/JNEUROSCI.3684-10.2011.

Uri Hasson, Eunice Yang, Ignacio Vallines, David J. Heeger, and Nava Rubin. A hierarchy of temporal receptive windows in human cortex. *Journal of Neuroscience*, 28(10):2539–2550, 2008. ISSN 02706474. doi:10.1523/JNEUROSCI.5487-07.2008.

Beren Millidge, Anil Seth, and Christopher L Buckley. Predictive Coding: a Theoretical and Experimental Review. *arXiv*, pages 1–56, 2021a. URL `http://arxiv.org/abs/2107.12979`.

Ryan Smith, Karl J. Friston, and Christopher J. Whyte. A step-by-step tutorial on active inference and its application to empirical data. *Journal of Mathematical Psychology*, 107:102632, 2022. ISSN 10960880. doi:10.1016/j.jmp.2021.102632. URL `https://doi.org/10.1016/j.jmp.2021.102632`.

Yen Yu and Karl J. Friston. *The selective updating of working memory : a predictive coding account*. PhD thesis, 2014.

Nelson Cowan. What are the differences between long-term, short-term, and working memory? *Prog Brain Res.*, 6123 (07):323–338, 2009. ISSN 0079-6123. doi:10.1016/S0079-6123(07)00020-9.What.

Klaus Oberauer and Hsuan Yu Lin. An interference model of visual working memory. *Psychological Review*, 124(1): 21–59, 2017. ISSN 0033295X. doi:10.1037/rev0000044.

Earl K. Miller, Mikael Lundqvist, and André M. Bastos. Working Memory 2.0. *Neuron*, 100(2):463–475, 2018. ISSN 10974199. doi:10.1016/j.neuron.2018.09.023.

Klaus Oberauer. Working memory and attention - A conceptual analysis and review. *Journal of Cognition*, 2(1):1–23, 2019. ISSN 25144820. doi:10.5334/joc.58.

Sabrina Trapp, Thomas Parr, Karl Friston, and Erich Schröger. The Predictive Brain Must Have a Limitation in Short-Term Memory Capacity. *Current Directions in Psychological Science*, 30(5):384–390, 2021. ISSN 14678721. doi:10.1177/09637214211029977.

Naomi P. Friedman and Trevor W. Robbins. The role of prefrontal cortex in cognitive control and executive function. *Neuropsychopharmacology*, 47(1):72–89, 2022. ISSN 1740634X. doi:10.1038/s41386-021-01132-0.

Alan Baddeley. Working memory: Looking back and looking forward. *Nature Reviews Neuroscience*, 4(10):829–839, 2003. ISSN 14710048. doi:10.1038/nrn1201.

Nelson Cowan. Working Memory Underpins Cognitive Development, Learning, and Education. *Educationnal Psychological Review*, 26(2):197–223, 2014. doi:10.1007/s10648-013-9246-y.Working.

Rodrigo Ortega, Vladimir López, Ximena Carrasco, María Josefina Escobar, Adolfo M. García, Mario A. Parra, and Francisco Aboitiz. Neurocognitive mechanisms underlying working memory encoding and retrieval in Attention-Deficit/Hyperactivity Disorder. *Scientific Reports*, 10(1):1–13, 2020. ISSN 20452322. doi:10.1038/s41598-020-64678-x.

Hamdi Eryilmaz, Alexandra S. Tanner, New Fei Ho, Adam Z. Nitenson, Noah J. Silverstein, Liana J. Petruzzi, Donald C. Goff, Dara S. Manoach, and Joshua L. Roffman. Disrupted working memory circuitry in schizophrenia: Disentangling fMRI markers of core pathology vs other aspects of impaired performance. *Neuropsychopharmacology*, 41(9): 2411–2420, 2016. ISSN 1740634X. doi:10.1038/npp.2016.55.

Holger Jahn. Memory loss in Alzheimer's disease. *Dialogues Clin Neurosci*, pages 1–5, 2013. doi:10.4324/9780429314384-1.

Stephan Zheng, Yang Song, Thomas Leung, and Ian Goodfellow. Improving the robustness of deep neural networks via stability training. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016-December:4480–4488, 2016. ISSN 10636919. doi:10.1109/CVPR.2016.485.

Eva Berlot and Floris P. de Lange. Flexible recoding of visual input for memory storage. *Neuron*, 110(11):1747–1749, 2022. ISSN 10974199. doi:10.1016/j.neuron.2022.04.023. URL https://doi.org/10.1016/j.neuron.2022.04.023.

Thomas Parr and Karl J. Friston. Working memory, attention, and salience in active inference. *Scientific Reports*, 7(1): 1–21, 2017. ISSN 20452322. doi:10.1038/s41598-017-15249-0.

Mengli Feng, Abhirup Bandyopadhyay, and Jorge F. Mejias. Emergence of distributed working memory in a human brain network model. *bioRxiv*, page 2023.01.26.525779, 2023. URL https://www.biorxiv.org/content/10.1101/2023.01.26.525779v1%0Ahttps://www.biorxiv.org/content/10.1101/2023.01.26.525779v1.abstract.

Jutta S. Mayer, Robert A. Bittner, Danko Nikolić, Christoph Bledowski, Rainer Goebel, and David E.J. Linden. Common neural substrates for visual working memory and attention. *NeuroImage*, 36(2):441–453, 2007. ISSN 10538119. doi:10.1016/j.neuroimage.2007.03.007.

Eric I. Knudsen. Fundamental components of attention. *Annual Review of Neuroscience*, 30:57–78, 2007. ISSN 0147006X. doi:10.1146/annurev.neuro.30.051606.094256.

Harriet Feldman and Karl J. Friston. Attention, uncertainty, and free-energy. *Frontiers in Human Neuroscience*, 4 (December):1–23, 2010. ISSN 16625161. doi:10.3389/fnhum.2010.00215.

Joseph R. Pauszek. *A predictive coding account of attention control*. PhD thesis, 2019.

Jakob Hohwy. Attention and conscious perception in the hypothesis testing brain. *Frontiers in Psychology*, 3(APR): 1–14, 2012. ISSN 16641078. doi:10.3389/fpsyg.2012.00096.

Madeleine Ransom, Sina Fazelpour, and Christopher Mole. Attention in the predictive mind. *Consciousness and Cognition*, pages 1–14, 2016. ISSN 1053-8100. doi:10.1016/j.concog.2016.06.011. URL http://dx.doi.org/10.1016/j.concog.2016.06.011.

Nelson Cowan. . Introduction to the problem of mental storage capacity. *Behavioral and brain sciences*, (4):87–185, 2000. ISSN 0140-525X.

Pierre Barrouillet and Valérie Camos. Interference: unique source of forgetting in working memory? *Trends in Cognitive Sciences*, 13(4):145–146, 2009. ISSN 13646613. doi:10.1016/j.tics.2009.01.002.

John E. Lisman and Marco A.P. Idiart. Storage of 7 ± 2 short-term memories in oscillatory subcycles. *Science*, 267 (5203):1512–1515, 1995. ISSN 00368075. doi:10.1126/science.7878473.

John Palmer. Attentional Limits on the Perception and Memory of Visual Information. *Journal of Experimental Psychology: Human Perception and Performance*, 16(2):332–350, 1990. ISSN 00961523. doi:10.1037/0096-1523.16.2.332.

Roshan Cools and Mark D'Esposito. Inverted-U-shaped dopamine actions on human working memory and cognitive control. *Biological Psychiatry*, 69(12), 2011. ISSN 00063223. doi:10.1016/j.biopsych.2011.03.028.

Laurence Aitchison and Máté Lengyel. With or without you: predictive coding and Bayesian inference in the brain. *Current Opinion in Neurobiology*, 46:219–227, 2017. ISSN 18736882. doi:10.1016/j.conb.2017.08.010.

Rajesh P.N. Rao and Dana H. Ballard. Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, 2(1):79–87, 1999. ISSN 10976256. doi:10.1038/4580.

Andy Clark. Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, 36(3):181–204, 2013. ISSN 14691825. doi:10.1017/S0140525X12000477.

Karl Friston. Active inferece: a process theory. *Neural Comput.*, 2733(March):2709–2733, 2017. doi:10.1162/NECO. URL http://arxiv.org/abs/1803.01446.

Maell Cullen. *This electronic thesis or dissertation has been downloaded from Explore Bristol Research , Author :*. PhD thesis, 2020.

Takuya Isomura, Hideaki Shimazaki, and Karl J. Friston. Canonical neural networks perform active inference. *Communications Biology*, 5(1):1–15, 2022. ISSN 23993642. doi:10.1038/s42003-021-02994-2.

Charlotte Caucheteux, Alexandre Gramfort, and Jean Rémi King. Evidence of a predictive coding hierarchy in the human brain listening to speech. *Nature Human Behaviour*, 7(March), 2023. ISSN 23973374. doi:10.1038/s41562-022-01516-2.

Fabian A. Mikulasch, Lucas Rudelt, Michael Wibral, and Viola Priesemann. Where is the error? Hierarchical predictive coding through dendritic error computation. *Trends in Neurosciences*, 46(1):45–59, 2023. ISSN 1878108X. doi:10.1016/j.tins.2022.09.007. URL https://doi.org/10.1016/j.tins.2022.09.007.

M. W. Spratling. A review of predictive coding algorithms. *Brain and Cognition*, 112:92–97, 2017. ISSN 10902147. doi:10.1016/j.bandc.2015.11.003.

Rebecca V. Zhang, Robert E. Featherstone, Olya Melynchenko, Raymond Gifford, Rachel Weger, Yuling Liang, and Steven J. Siegel. High-beta/low-gamma frequency activity reflects top-down predictive coding during a spatial working memory test. *Experimental Brain Research*, 237(7):1881–1888, 2019. ISSN 14321106. doi:10.1007/s00221-019-05558-3. URL https://doi.org/10.1007/s00221-019-05558-3.

Karl Friston. Hierarchical models in the brain. *PLoS Computational Biology*, 4(11), 2008. ISSN 15537358. doi:10.1371/journal.pcbi.1000211.

Beren Millidge, Alexander Tschantz, and Christopher L. Buckley. Whence the expected free energy? *Neural Computation*, 33(2):447–482, 2021b. ISSN 1530888X. doi:10.1162/neco_a_01354.

L. Paletta, C. Seifert, and G. Fritz. Saccadic object recognition by a markov decision process in a cascaded framework. In Robert Schwartz, editor, *Perception*, pages 126–126. Malden Ma: Blackwell, 2004.

Gina R. Kuperberg. Tea With Milk? A Hierarchical Generative Framework of Sequential Event Comprehension. *Topics in Cognitive Science*, 13(1):256–298, 2021. ISSN 17568765. doi:10.1111/tops.12518.

I L Hijne. *Generalised Motions in Active Inference by finite differences*. PhD thesis, 2020. URL http://resolver.tudelft.nl/uuid:9102f269-ca73-4281-99e0-ea911282859e.

Karl Friston, Queen Square, and James Kilner. Europe PMC Funders Group Action understanding and active inference. *Biological cybernetics*, 104(1-2):137–160, 2011. doi:10.1007/s00422-011-0424-z.Action.

Shirin Dora, Sander M. Bohte, and Cyriel M.A. Pennartz. Deep Gated Hebbian Predictive Coding Accounts for Emergence of Complex Neural Response Properties Along the Visual Cortical Hierarchy. *Frontiers in Computational Neuroscience*, 15(July):1–20, 2021. ISSN 16625188. doi:10.3389/fncom.2021.666131.

Gabriel A. Radvansky and Jeffrey M. Zacks. Event perception. *Wiley Interdisciplinary Reviews: Cognitive Science*, 2 (6):608–620, 2011. ISSN 19395078. doi:10.1002/wcs.133.

Stefania Sarno, Manuel Beiran, Joan Falco-Roget, Gabriel Diaz-DeLeon, Roman Rossi-Pool, Ranulfo Romo, and Nestor Parga. Dopamine firing plays a dual role in coding reward prediction errors and signaling motivation in a working memory task. *Proceedings of the National Academy of Sciences of the United States of America*, 119(2), 2022. ISSN 10916490. doi:10.1073/pnas.2113311119.

Peter Redgrave, Kevin Gurney, and John Reynolds. What is reinforced by phasic dopamine signals? *Brain Research Reviews*, 58(2):322–339, 2008. ISSN 01650173. doi:10.1016/j.brainresrev.2007.10.007.

Jorge F. Mejías and Xiao Jing Wang. Mechanisms of distributed working memory in a large-scale network of macaque neocortex. *eLife*, 11:1–33, 2022. ISSN 2050084X. doi:10.7554/ELIFE.72136.

Justin R. Chumbley, Raymond J. Dolan, and Karl J. Friston. Attractor models of working memory and their modulation by reward. *Biological Cybernetics*, 98(1):11–18, 2008. ISSN 03401200. doi:10.1007/s00422-007-0202-0.

Minju Jung, Takazumi Matsumoto, and Jun Tani. Goal-Directed Behavior under Variational Predictive Coding: Dynamic organization of Visual Attention and Working Memory. *IEEE International Conference on Intelligent Robots and Systems*, pages 1040–1047, 2019. ISSN 21530866. doi:10.1109/IROS40897.2019.8968597.

A. Emin Orhan, Chris R. Sims, Robert A. Jacobs, and David C. Knill. The Adaptive Nature of Visual Working Memory. *Current Directions in Psychological Science*, 23(3):164–170, 2014. ISSN 14678721. doi:10.1177/0963721414529144.

Nelson Cowan, Christopher L. Blume, and J. Scott Saults. Attention to attributes and objects in working memory. *Journal of Experimental Psychology: Learning Memory and Cognition*, 39(3):731–747, 2013. ISSN 02787393. doi:10.1037/a0029687.

Paul M. Bays, Raquel F.G. Catalao, and Masud Husain. The precision of visual working memory is set by allocation of a shared resource. *Journal of Vision*, 9(10):1–11, 2009. ISSN 15347362. doi:10.1167/9.10.7.

Clayton E. Curtis and Thomas C. Sprague. Persistent Activity During Working Memory From Front to Back. *Frontiers in Neural Circuits*, 15(July):1–17, 2021. ISSN 16625110. doi:10.3389/fncir.2021.696060.

M. Berk Mirza, Rick A. Adams, Christoph D. Mathys, and Karl J. Friston. Scene construction, visual foraging, and active inference. *Frontiers in Computational Neuroscience*, 10(JUN), 2016. ISSN 16625188. doi:10.3389/fncom.2016.00056.

Karl J. Friston, Richard Rosch, Thomas Parr, Cathy Price, and Howard Bowman. Erratum: Deep temporal models and active inference (Neuroscience Biobehavioral Reviews (2017) 77 (388–402) (S0149763416307096) (10.1016/j.neubiorev.2017.04.009)). *Neuroscience and Biobehavioral Reviews*, 90(May):486–501, 2018. ISSN 18737528. doi:10.1016/j.neubiorev.2018.04.004. URL https://doi.org/10.1016/j.neubiorev.2018.04.004.

Bradley R. Postle. Working Memory as an Emergent Property of the Mind and Brain. *Neuroscience*, 139(1):1–7, 2006. ISSN 15378276.

Chris R. Sims, Rober A. Jacobs, and David C. Knil. An ideal observer analysis of visual working memory. *Psychological Review*, 90(2):133–154, 2012. ISSN 15378276. doi:10.1037/a0029856.An. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3624763/pdf/nihms412728.pdf.

Wei Ji Ma and Mehrdad Jazayeri. Neural coding of uncertainty and probability. *Annual Review of Neuroscience*, 37: 205–220, 2014. ISSN 15454126. doi:10.1146/annurev-neuro-071013-014017.

Xiao Jing Wang. Synaptic reverberation underlying mnemonic persistent activity. *Trends in Neurosciences*, 24(8): 455–463, 2001.

Klaus Wimmer, Duane Q. Nykamp, Christos Constantinidis, and Albert Compte. Bump attractor dynamics in prefrontal cortex explains behavioral precision in spatial working memory. *Nature Neuroscience*, 17(3):431–439, 2014. ISSN 10976256. doi:10.1038/nn.3645. URL http://dx.doi.org/10.1038/nn.3645.

Gianluigi Mongillo, Omri Barak, and Misha Tsodyks. Synaptic Theory of Working Memory. *Science*, 319(5869): 1543–1546, 2008. ISSN 10959203. doi:10.1126/science.1150769.

Darinka Trübutschek, Sébastien Marti, Andrés Ojeda, Jean Rémi King, Yuanyuan Mi, Misha Tsodyks, and Stanislas Dehaene. A theory of working memory without consciousness or sustained activity. *eLife*, 6, 2017. ISSN 2050084X. doi:10.7554/eLife.23871.

Mark S. Goldman. Memory without Feedback in a Neural Network. *Neuron*, 61(4):621–634, 2009. ISSN 08966273. doi:10.1016/j.neuron.2008.12.012.

Jorge Jaramillo, Jorge F. Mejias, and Xiao Jing Wang. Engagement of Pulvino-cortical Feedforward and Feedback Pathways in Cognitive Computations. *Neuron*, 101(2):321–336.e9, 2019. ISSN 10974199. doi:10.1016/j.neuron.2018.11.023. URL https://doi.org/10.1016/j.neuron.2018.11.023.

Naoki Kogo and Chris Trengove. Is predictive coding theory articulated enough to be testable? *Frontiers in Computational Neuroscience*, 9(SEP):1–4, 2015. ISSN 16625188. doi:10.3389/fncom.2015.00111.

Thijs W. van de Laar and Bert de Vries. Simulating active inference processes by message passing. *Frontiers Robotics AI*, 6(MAR), 2019. ISSN 22969144. doi:10.3389/frobt.2019.00020.

Madeleine Ransom and Sina Fazelpour. Three Problems for the Predictive Coding Theory of Attention Madeleine, 2015. URL https://mindsonline.philosophyofbrains.com/2015/session4/three-problems-for-the-predictive-coding-theory-of-attention/.

10 Appendix A Two models of WM with PC

There are two PC models explicitly incorporating WM, which we use to explain how PC explain WM. Readers who would like to gain a more holistic understanding of the relevant background can refer to this appendix.

10.1 A model of event perception

When a set of visual events unfolds in front of our eyes, being able to predict what happens next is very useful to comprehend the events with speed and accuracy. For example, when a woman pours water into a cup, we can infer that she may want to make tea and therefore predict the next event, such as adding milk. Being able to predict sensory input allows us to quickly change our behavior.

Kuperberg [2021] proposed a hierarchical generative framework with three levels to model event perception. WM is represented in the second level of the model containing the *event model* for making predictions of future states. The *event model* represents a series of events that have happened to provide the context for making predictions about the next event [Radvansky and Zacks, 2011].

The model assumes that schema-relevant knowledge are inferred through the goals at the top/third level to constrain the resultant predictions. For example, when the higher level infers that the subject is making tea $p(\text{making tea}|o)$ is highest, then WM receives tea-relevant event clusters, from which single events at the bottom/first level are generated.

The three hierarchies are only conceptual and there is no one-to-one map to brain structure. If needed, such a model can be expanded to more hierarchies.

10.2 A model of saccadic eye movement

Saccadic control can also be modelled with active inference. This model assumes that during saccadic eye movement, an agent selects a saccadic target based on the predictions about the outcomes of their movement. This then supports or denies their beliefs about hidden states of the world given an observation.

Parr and Friston [2017] proposed a model of saccadic control based on a MDP, similar to the framework shown in Figure 8. This model assumes the dependency of the variables as the following: the outcome of each movement is dependent on a hidden state in the present time point, that hidden state depends on the policy and the hidden state at the previous time point. In such a MDP, the agent needs to minimize free energy over time. In order to do that, they need to choose a policy which minimizes their expected free energy.

The authors simulated the model on a specific WM task. In this task, participants are initially shown three pairs of scenes and instructed to remember one specific pair. After this initial encoding phase, a retrocue is presented, indicating which of the initial scenes is most likely to be tested. At the end of the trial, participants are shown two scenes and must identify which one was presented at the beginning. Throughout the trial, saccadic eye movements play a critical role in reducing uncertainty about the initially presented scenes, thereby improving performance. The authors model this task by treating the set of initial scenes as a hidden state and interpret the retrocue as an opportunity for belief updating. This theoretical framing allows the model to capture how participants integrate prior information and sensory evidence to guide memory-based decision making.

In this model, WM is interpreted as the accumulated evidence in the hidden state unit.The posterior probability of the presented initial scenes (the hidden state), changes with the accumulation of evidence. It is updated when the retrocue appears and maintained the same value during the delay period when there is no further input. This corresponds well with the update and maintenance of WM. The authors developed a neuronal message-passing scheme for the model [Parr and Friston, 2017, Friston et al., 2018], grounded in the mathematical formulation of variational free energy minimisation. This scheme maps belief updating operations—such as inference over states and policy selection—onto biologically plausible circuit motifs, including forward (bottom-up) and backwards (top-down) message passing between hierarchical levels. Crucially, the message-passing architecture incorporates cortical and subcortical structures thought to underlie WM processes, such as the prefrontal cortex and basal ganglia loops. The temporal dynamics of the simulated hidden state units, which represent beliefs about the initial scenes, closely resembled delay-period activity observed in electrophysiological recordings from monkey prefrontal cortex during WM tasks. This qualitative match supports the model's potential as a mechanistic account of WM in the brain.