A Neuro-Inspired Computational Framework for AGI: Predictive Coding, Active Inference, and Free Energy Minimisation

Alexander D. Shaw[1] and Lioba C.S. Berndt[1]

[1]Department of Psychology, University of Exeter, Exeter, UK

July 1, 2025

Abstract

This paper proposes that foundational principles from theoretical neuroscience - predictive coding, the Free Energy Principle (FEP), and variational inference - offer a biologically grounded framework for artificial general intelligence (AGI). These approaches characterise the brain as a hierarchical inference system that continuously updates beliefs and selects actions to minimise uncertainty and surprise. In contrast to conventional AI systems, which typically rely on static architectures and offline training, biological agents engage in active, generative inference within dynamic, uncertain environments. We argue that it is this inference-based architecture-not just its behavioural outputs-that underpins the adaptability, generalisation, and resilience of natural intelligence. We outline a neuro-inspired computational framework built on hierarchical generative models, scalable variational inference (e.g., Variational Laplace), and Active Inference. Finally, we contrast this approach with dominant deep learning paradigms and discuss its implications for building interpretable, adaptive, and autonomous machine intelligence.

1 Introduction

The human brain remains the only known system capable of *general intelligence (GI)* - the ability to flexibly interpret, model, and act within dynamic, uncertain, and novel environments. While formal definitions vary, GI is typically characterised by context-sensitive reasoning, adaptive learning, and goal-directed action across multiple domains [1, 2]. A growing body of work in theoretical neuroscience suggests that the key to this flexibility lies not in memorised patterns or fixed rules, but in the brain's capacity for *inference*.

Specifically, theories such as *predictive coding* [3, 4] and the *Free Energy Principle (FEP)* [5, 6] propose that perception, action, and learning arise from the hierarchical minimisation of prediction error via approximate Bayesian inference. In this view, the brain is a *generative model*: it continuously predicts sensory inputs, updates internal beliefs in response to mismatches, and selects actions that minimise expected future surprise.

This predictive loop supports both *epistemic behaviour* (reducing uncertainty through exploration) and *instrumental behaviour* (pursuing preferred outcomes), achieved by internally simulating possible futures and evaluating them according to expected free energy [7, 8]. The underlying dynamics are governed by a *variational principle*, in which the brain minimises a quantity called *variational free energy* - a bound on model evidence-providing a unifying explanation for learning, adaptation, and action selection under uncertainty. Formally, variational free energy is defined as:

$$\mathcal{F}(q,p) = \mathrm{KL}[q(z)\|p(z|x)] - \log p(x), \tag{1}$$

where $q(z)$ is the approximate posterior, $p(z|x)$ the true posterior, and $p(x)$ the model evidence. Minimising this bound approximates Bayesian inference in a tractable way.

Meanwhile, artificial intelligence (AI) has made rapid progress in areas such as *transformer-based large language models (LLMs)*, *diffusion-based generative models*, and *reinforcement learning agents* [9, 10, 11]. Despite these breakthroughs, modern AI systems often struggle to generalise beyond their training distribution, adapt online, or robustly handle uncertainty [12, 13]. They typically lack explicit generative models of the environment and rely on static architectures trained via supervised or reinforcement learning.

This paper explores the emerging hypothesis that the computational principles underlying brain function - predictive coding, dynamical systems modelling, and active inference - may provide a *biologically grounded blueprint* for constructing general-purpose agents. We argue that such agents should not merely reproduce intelligent behaviour, but embody the same self-organising, inference-based principles that underlie human cognition.

To develop this argument, we proceed as follows. First, we examine the brain as an inference engine, showing how hierarchical predictive coding and active inference give rise to perception, action, and learning. Next, we contrast this with prevailing AI paradigms-particularly transformer and diffusion-based models-highlighting their strengths and limitations. We then present the Free Energy Principle and inference methods such as *Variational Laplace* as scalable engines for neuro-inspired AI. Finally, we propose a roadmap for AGI grounded in the computational architecture of the brain.

2 The Brain as a Model of General Intelligence

Biological intelligence arises not from memorised rules or static mappings, but from the brain's ability to infer latent causes, predict sensory consequences, and adapt behaviour accordingly. Predictive coding offers a compelling computational account of how this is achieved [3, 4]. It posits that the brain maintains hierarchical generative models that continuously predict incoming sensory input and minimise the mismatch between predicted and observed signals.

This approach views perception, cognition, and action as outcomes of approximate Bayesian inference embedded in a dynamical system [14, 15]. The brain does not merely react to the world - it models it, anticipates it, and selects actions that reduce uncertainty about it.

2.1 Hierarchical Generative Models and Approximate Bayesian Inference

The brain is organised into a hierarchical architecture-spanning both anatomical and functional levels-in which each level generates predictions about the activity of the level below. Sensory signals are explained away by these top-down predictions, and only residual errors-i.e., the unexpected components of sensory input-are propagated upward. This message-passing architecture supports approximate Bayesian inference over latent variables $\mathbf{z}$, given sensory observations $\mathbf{x}$, under a joint generative model [5]:

$$p(\mathbf{x}, \mathbf{z}) = p(\mathbf{x}|\mathbf{z})p(\mathbf{z}) \tag{2}$$

Because exact Bayesian inference is intractable in realistic settings, the brain is assumed to represent a variational posterior $q(\mathbf{z})$ that approximates the true posterior $p(\mathbf{z}|\mathbf{x})$. Inference then proceeds by minimising a variational free energy functional [16]:

$$\mathcal{F}[q] = \mathrm{KL}[q(\mathbf{z})||p(\mathbf{z})] - \mathbb{E}_{q(\mathbf{z})}[\log p(\mathbf{x}|\mathbf{z})] \tag{3}$$

Minimising free energy ensures that internal beliefs are both accurate (i.e., maximise the likelihood of observed input) and parsimonious (i.e., remain close to prior expectations). This provides a principled trade-off between flexibility and generalisation, which is essential for adaptive intelligence [5, 6].

Neurobiologically, predictive coding is thought to be implemented within canonical cortical microcircuits, which exhibit a consistent layered (laminar) and column-like (columnar) structure across sensory, associative, and motor areas of the cortex [17, 18]. This architecture supports the bidirectional flow of information central to predictive coding.

In these circuits, pyramidal neurons in the superficial layers (layers 2 and 3) are believed to encode prediction errors-signals reflecting mismatches between expected and actual sensory input. These neurons send forward projections to higher cortical areas, primarily targeting layer 4 of downstream regions [17].

By contrast, deep-layer pyramidal neurons (layers 5 and 6) are thought to encode top-down predictions. They project back to lower hierarchical levels, targeting both superficial layers and layer 4 interneurons, thereby modulating the gain of ascending prediction errors.

Interlaminar inhibitory interneurons-such as somatostatin-positive (SST) and parvalbumin -positive (PV) cells-further refine this process by regulating the precision, or weighting, of prediction errors through modulatory control of excitatory neurons [19, 20].

This laminar organisation supports recursive message passing across hierarchical levels: prediction errors ascend to update beliefs, while predictions descend to explain away sensory input. The anatomical segregation of error and state representations across layers enables dynamic inference over multiple timescales and levels of abstraction-from basic sensation to higher-order cognition. Crucially, this architecture also provides a neurobiological substrate for attention, precision modulation, and context-sensitive gain control within predictive coding frameworks [21, 20].

2.2 Predictive Coding for Perception

In the predictive coding account of perception, the brain is cast as a generative system that infers the hidden causes of its sensory inputs [3, 4]. Each level of the cortical hierarchy encodes beliefs about latent variables in the level below, generating top-down predictions that are compared to incoming signals. When predictions fail to fully explain the input, the residuals-known as *prediction errors*-are propagated upward [15]. These errors update the internal model, refining beliefs about the underlying causes. This iterative, bidirectional message passing implements approximate Bayesian inference in real time [6].

Let s denote sensory observations and z the latent causes. A generative model $p(s, z) = p(s|z)p(z)$ defines the likelihood and prior over hidden states. Because exact inference of the posterior $p(z|s)$ is generally intractable, the brain maintains a variational approximation $q(z)$, which is updated to minimise prediction error. The prediction error can be written as:

$$\varepsilon = s - \hat{s}(z) \tag{4}$$

where $\hat{s}(z)$ is the predicted sensory input based on current estimates of z. Assuming Gaussian distributions, free energy minimisation leads to a gradient descent update of the latent causes [21]:

$$\frac{dz}{dt} \propto \frac{\partial \varepsilon}{\partial z} \cdot \Sigma^{-1} \cdot \varepsilon \tag{5}$$

Here, Σ denotes the precision (inverse variance) of the prediction error. Precision determines the influence of prediction errors on belief updating, effectively weighting the reliability of incoming sensory data [8].

At the neural level, these hierarchical generative models are thought to be implemented via canonical cortical microcircuits, in which distinct neuronal populations support bidirectional message passing and precision modulation [17, 19].

Example: Predictive Coding in Visual Perception

Consider a simple visual scene in which a person views a cup on a table. Higher-level visual areas (e.g., inferotemporal cortex) generate abstract predictions such as “object with a handle” or “familiar graspable shape” [22]. These predictions are sent down the hierarchy, shaping expectations for the patterns of activity in lower-level areas (e.g., orientation edges in V1). If the incoming sensory input matches these expectations, prediction error remains low. However, if the object is partially occluded or appears unusual-say, a broken cup or an unexpected angle-prediction errors arise in early visual areas and are sent upward. These errors are crucial: they signal a mismatch and prompt higher-level areas to update their beliefs until the model settles on the most likely explanation (e.g., “damaged cup”).

In this way, perception operates not as passive registration, but as the brain’s best guess about the hidden causes of ambiguous, noisy input, allowing for rapid adaptation to new or unexpected situations.

2.3 Action as Inference: Active Inference

Notably, predictive coding is not restricted to perception. In the framework of Active Inference, action selection is cast as an inferential process driven by the same imperative that governs perception: the minimisation of expected free energy [23]. In this view, agents do not passively infer the causes of their observations-they actively select policies that make future observations more predictable and less surprising.

The expected free energy $G(\pi)$ under a policy π can be decomposed into epistemic and instrumental components, capturing both uncertainty reduction (exploration) and goal-directed behaviour (exploitation). One common formulation expresses this as:

$$G(\pi) = \mathbb{E}_{q(\mathbf{x},\mathbf{z}|\pi)} \left[\log q(\mathbf{z}|\pi) - \log p(\mathbf{z}|\mathbf{x},\pi)\right] \tag{6}$$

This form reflects the expected divergence between posterior and likelihood-minimising it encourages the agent to seek actions that simultaneously reduce uncertainty about hidden states and achieve preferred outcomes [24, 25]. In effect, perception and action are unified under a single inferential scheme: the brain continually infers not only what is happening, but what it should do to bring about desirable and predictable states of affairs.

Unlike traditional reinforcement learning, which relies on externally defined reward signals and policy optimisation, Active Inference derives behaviour from an internal generative model and an imperative to minimise variational free energy. In this framework, goals and preferences are encoded as prior beliefs about preferred states, and behaviour emerges from the drive to realise these states while reducing uncertainty. This makes Active Inference naturally suited to continual, adaptive control in uncertain environments, offering a biologically grounded alternative to model-free or value-based methods [26].

Decomposing Expected Free Energy

Under Active Inference, agents select actions that minimise the *expected free energy* $G(\pi)$, which captures both uncertainty reduction (epistemic value) and goal fulfilment (instrumental value). A common decomposition is:

$$G(\pi) = \underbrace{-E_{q(\mathbf{x}|\pi)} \left[\mathrm{D}_{\mathrm{KL}}\big(q(\mathbf{z}|\mathbf{x},\pi) \parallel q(\mathbf{z}|\pi)\big)\right]}_{\text{Epistemic value (uncertainty reduction)}} + \underbrace{\mathbb{E}_{q(\mathbf{x}|\pi)} \left[-\log p(\mathbf{x})\right]}_{\text{Instrumental value (preference satisfaction)}}$$

Interpretation:

- The first term encourages the agent to take actions that will be informative-i.e., actions that are expected to reduce uncertainty about hidden states.
- The second term pushes the agent toward outcomes that are consistent with prior preferences-i.e., desired or rewarding outcomes.

Together, these components drive both exploration and exploitation without requiring an explicit reward signal.

2.4 Cognition as Dynamical Inference in State Space

Rather than making isolated decisions or computations, the brain operates as a dynamical system-its activity is constantly changing and evolving over time [27]. At any moment, the pattern of neural activity can be represented as a point in a high-dimensional space, where each dimension corresponds to the activity of a different neuron or neural population. As the brain processes information, this point traces out a path, or trajectory, through the space. These trajectories reflect how the brain continuously updates its beliefs about the hidden causes of sensory input. This ongoing process is driven by prediction errors-differences between expected and actual input-and is influenced by the brain's level of uncertainty about its current beliefs [21, 5].

Such recurrent neural dynamics-whether oscillatory, chaotic, or metastable-can be formalised as solutions to differential equations that govern the temporal evolution of beliefs. In continuous time, this is captured by:

$$\frac{d\mu}{dt} = f(\mu, \Sigma, \mathbf{x}) - \nabla_{\mu}\mathcal{F} \tag{7}$$

The first term $f(\mu, \Sigma, \mathbf{x})$ describes how beliefs would evolve over time based on prior expectations or natural dynamics-essentially, how the brain expects things to change even without new sensory input. The second term acts as a corrective force, adjusting beliefs to reduce prediction errors and improve the fit between the brain's model and actual sensory input [28].

Importantly, these neural dynamics are not mere computational artefacts-they have clear biological substrates. Cortical microcircuits implement the message passing required for hierarchical inference; thalamocortical loops regulate the gain and precision of these messages through rhythmic synchronisation; and neuromodulatory systems (such as noradrenaline and dopamine) modulate uncertainty by adjusting the precision weighting of prediction errors [29, 17].

This perspective challenges the classical view of cognition as computation in the traditional Turing sense-that is, as discrete, symbolic manipulation or pattern classification. Instead, cognition emerges as a continuous, embodied, and context-sensitive process of inference. The brain functions as a dynamical generative system, continuously inferring both its own internal state and the state of the world by traversing a landscape of beliefs shaped by free energy gradients.

2.5 From Neural Dynamics to General Intelligence

The implication is profound: general intelligence in biological agents does not arise from learning specific input–output mappings or storing task-specific solutions. Instead, it emerges from a capacity for continual, structured inference-dynamically estimating latent causes, updating beliefs over time, and selecting actions to minimise future uncertainty [30].

From this perspective, the brain's solution to general intelligence is not a fixed algorithm, but a dynamical system for approximate inference in deep, uncertain, and non-stationary environments. Hierarchical generative models, precision-weighted prediction errors, and temporally extended action selection together form a biologically grounded architecture for flexible cognition [31].

Crucially, predictive coding and Active Inference provide more than a descriptive theory of neural computation. They offer a mechanistic template for constructing artificial systems that can reason, plan, and adapt in the same situated, embodied manner as human agents. If general intelligence requires the ability to model hidden causes, update beliefs online, and act to shape future outcomes, then these frameworks may serve as blueprints for AGI.

In the next section, we examine how current AI systems fall short of these principles-and what may be gained by shifting toward inference-based architectures inspired by the brain.

3 Predictive Coding and the Limits of Deep Learning

The recent successes of deep learning-particularly convolutional neural networks (CNNs) in vision [32] and large language models (LLMs) in language modelling [9, 11]-have driven remarkable advances in artificial intelligence. These architectures excel at extracting statistical regularities from large, static datasets, enabling superhuman performance on benchmark tasks such as image classification, text completion, and game playing [33].

Despite these achievements, deep learning systems remain fundamentally constrained in several respects. They generalise poorly beyond their training distribution [34], struggle to represent uncertainty in a principled way [35], and lack the capacity for continual, online adaptation [36]. Most notably, they are reactive rather than proactive: they do not infer hidden causes, construct internal models of the environment, or select actions based on long-term expectations [37].

In this section, we briefly review the core operating principles of CNNs and LLMs and contrast them with those of predictive coding and Active Inference architectures [5, 6]. Whereas conventional deep learning systems rely on static feedforward mappings optimised via backpropagation, predictive coding systems perform ongoing, hierarchical inference over latent causes, with uncertainty estimation and action selection embedded in the same inferential loop.

Recent developments in deep learning-such as recurrent architectures, uncertainty-aware models, and online fine-tuning-have begun to blur these boundaries. Nevertheless, predictive coding and Active Inference provide a unified, biologically grounded framework in which inference, learning, and behaviour emerge from a single underlying principle: the minimisation of variational free energy.

This contrast highlights key missing ingredients in current AI systems-and motivates a shift toward inference-based, neurobiologically inspired architectures with the potential for more general, adaptive intelligence.

It should be noted that, despite their conceptual and computational differences, predictive coding and Active Inference architectures share some limitations with deep learning approaches-most notably, the reliance on fixed model structures and predefined state spaces. Furthermore, although these inference-based models offer theoretical advantages in uncertainty handling, continual adaptation, and the integration of perception and action, they have yet to demonstrate consistent superiority over deep learning on large-scale, real-world tasks. At present, their strengths are most evident in neuroscience settings and constrained simulations rather than in mainstream AI benchmarks.

3.1 Convolutional Neural Networks: Feedforward Feature Extractors

Convolutional neural networks (CNNs) are designed to exploit the spatial regularities of image data through a cascade of local filtering operations [32]. Each convolutional layer applies a set of learned filters to its input, producing activation maps that represent increasingly abstract features-ranging from edges and textures to object parts and global shapes. These activations are typically followed by nonlinearities (e.g., ReLU) and pooling

Table 1: Contrasting deep learning systems with predictive coding and Active Inference architectures.

Aspect	**Deep Learning (CNNs / LLMs)**	**Predictive Coding / Active Inference**
Core Principle	Pattern recognition from data	Inference over latent causes
Architecture	Feedforward, static layers	Hierarchical, recurrent, dynamic
Learning Mechanism	Gradient descent on loss functions (e.g. cross-entropy)	Minimisation of variational free energy
Uncertainty Handling	Often implicit or approximate (e.g. dropout)	Explicit via precision-weighted errors and variational posteriors
Adaptivity	Retraining or fine-tuning required	Online inference over fixed model structure; adaptation without retraining
Action Selection	Typically via separate reinforcement learning module	Unified with perception via expected free energy
World Model	Discriminative, input-to-output mapping	Generative, simulates causes and consequences
Generalisation	Often brittle under distribution shift	Potentially more robust via structured inference and uncertainty modelling
Biological Plausibility	Loosely inspired (e.g., convolutional hierarchy)	Explicitly mapped to cortical microcircuits and neuromodulation

layers that introduce spatial invariance, culminating in a fully connected classification or regression output layer.

Mathematically, the activity at layer l is computed as:

$$h^{(l)} = \sigma(W^{(l)} * h^{(l-1)} + b^{(l)}) \tag{8}$$

where $*$ denotes convolution, $W^{(l)}$ and $b^{(l)}$ are trainable weights and biases, and σ is a pointwise nonlinearity.

While CNNs achieve strong performance on benchmark tasks such as object recognition and segmentation, they are fundamentally feedforward and static [38]. They do not maintain internal beliefs about latent causes, nor do they revise those beliefs in light of new observations. In contrast to predictive coding systems, CNNs lack recurrent dynamics, uncertainty representation, or generative models capable of simulating sensory input [39, 40].

As a result, standard feedforward CNNs are brittle under occlusion, adversarial noise, or distributional shift [41, 34]. Their mappings are primarily discriminative-optimised to assign labels to inputs rather than to infer their underlying causes. They typically lack mechanisms for top-down feedback, hypothesis testing, or self-correction in the face of sensory mismatch. In short, standard CNNs can classify what they have seen, but not infer why they have seen it.

Recent research has begun to address these limitations by augmenting CNNs with recurrent connections, generative components, or Bayesian uncertainty modeling (e.g., [39, 35, 42]). However, these architectures are not yet widely adopted in practice, in part due to increased computational demands and limited gains over conventional models in many applied settings [43].

3.2 Large Language Models: Amortised, Static Predictors

Large Language Models (LLMs) such as GPT and PaLM are built on the transformer architecture, which models relationships between input tokens via 'self-attention' [9]. During training, these models learn to predict the next token in a sequence-effectively approximating the conditional distribution $p(x_t \mid x_{<t})$ across large-scale datasets of natural language.

The core of each transformer layer is a self-attention mechanism:

$$\text{attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right) V \tag{9}$$

where queries Q, keys K, and values V are learned linear projections of token embeddings. These are processed through stacked layers of attention and feedforward transformations, producing rich contextual representations for each token.

Importantly, inference in LLMs is amortised: all model parameters are learned during training and remain fixed at inference time. When generating text, LLMs sample from the learned distribution without updating their internal parameters or explicitly representing uncertainty or prediction errors [13]. While they can incorporate contextual information from preceding tokens within a prompt, they do not revise their underlying beliefs or adapt in real time based on new observations.

Although transformers excel at capturing statistical patterns, syntax, and factual knowledge from training data, they remain fundamentally static predictors. They lack mechanisms for continual online adaptation, explicit uncertainty estimation, or model-based reasoning about hidden causes [44]. Unlike biological systems, LLMs cannot actively test hypotheses or update their internal models to reduce uncertainty about their environment. Their strength lies in mimicking linguistic structure and knowledge, rather than in performing inference or goal-directed action.

3.3 Deep Learning vs. Dynamical Inference

While deep learning models such as convolutional neural networks and large language models have achieved impressive feats in pattern recognition and data generation, they lack a foundational principle that characterises biological intelligence: dynamical inference. That is, the continual updating of internal beliefs over time in response to ongoing sensory input and uncertainty.

Most deep networks rely on a large set of static parameters, optimised offline via gradient descent. Once trained, these parameters remain fixed; inference involves applying a learned mapping rather than updating latent beliefs in light of new observations [1, 45]. There is no mechanism for real-time hypothesis testing, belief revision, or uncertainty minimisation.

While models such as LLMs, VAEs, GANs, and Diffusion Models are generative in the statistical sense, they do not maintain causal generative models of the environment. They typically do not infer hidden states that evolve over time or simulate the consequences of possible actions. There is no formal encoding of prediction errors, no top-down modulation of beliefs, and no probabilistic inference over hidden causes grounded in sensory feedback.

Furthermore, action selection-when included-is usually appended via external reinforcement learning modules. These operate separately from perception, lacking the tight integration found in systems that select actions to reduce uncertainty or fulfil prior preferences [46].

By contrast, biological systems continuously engage in closed-loop inference: they predict, perceive, and act in a recursive cycle aimed at reducing uncertainty and maintaining homeostasis. Predictive coding and Active Inference offer a principled account of such intelligence, where beliefs and actions emerge from a unified dynamical process. From this perspective, conventional deep learning falls short-not due to lack of complexity, but due to a lack of inference over time.

Deep Learning vs. Predictive Coding: A Summary of Missing Ingredients

- **Learning**: Deep networks have fixed parameters post-training; predictive coding systems learn online via continual belief updates.
- **Uncertainty**: Deep learning lacks epistemic introspection; Active Inference explicitly models and minimises uncertainty.
- **Causality**: CNNs and LLMs map inputs to outputs; the brain infers latent causes via generative models.
- **Feedback**: Conventional deep models are feedforward; predictive coding employs hierarchical top-down feedback.
- **Action**: In deep learning, action is bolted on; in Active Inference, action emerges from inference itself.

3.4 Advantages of Predictive Coding Architectures

Predictive coding architectures offer a number of advantages over conventional deep learning systems, particularly when applied to general intelligence in dynamic, uncertain environments [31, 4, 6]:

- **Uncertainty-aware inference**: Precision weighting allows the system to balance sensory evidence against prior expectations, enabling context-sensitive and uncertainty-modulated decision-making [47].
- **Sparse error signalling**: Only mismatches between predictions and observations-i.e., prediction errors-are propagated upward, reducing redundant signalling. While

this supports efficient processing at the level of error transmission, generating top-down predictions still requires sustained computational activity [21].

- **Biological plausibility**: Predictive coding aligns closely with known features of cortical organisation, including hierarchical structure, laminar microcircuits, and extensive feedback pathways [17].
- **Unified perception–action loop**: In Active Inference, perception and action arise from the same inferential process-actions are selected to fulfil predictions, completing a closed-loop system for autonomous, goal-directed behaviour [23].

Taken together, these properties position predictive coding not only as a leading theory of brain function, but also as a biologically grounded design principle for building artificial systems with human-like adaptability, efficiency, and autonomy.

In the next section, we explore how the Free Energy Principle and variational inference formalise these insights into a unified computational engine for artificial general intelligence.

4 The Free Energy Principle and Variational Laplace as Inference Engines

The Free Energy Principle (FEP) provides a unifying theoretical framework for understanding perception, action, and learning in biological systems [5, 16, 6]. At its core, the FEP posits that any adaptive agent-biological or artificial-must minimise a quantity known as variational free energy in order to resist entropy-that is, the natural tendency for physical systems to become disordered-and remain viable within its environment.

This free energy serves as a tractable upper bound on surprise (negative log model evidence), which is otherwise intractable to compute directly [5]. Minimising free energy ensures that the agent maintains a generative model of the world that remains consistent with incoming sensory data. This minimisation occurs both through updating internal beliefs (perception) and through selecting actions that lead to more predictable and preferred outcomes (action) [48].

4.1 Free Energy as a Unified Objective

Let $\mathbf{x}$ denote sensory input and $\mathbf{z}$ the latent states of the environment. In the Free Energy framework, the brain (or *agent*) maintains a generative model of the form:

$$p(\mathbf{x}, \mathbf{z}) = p(\mathbf{x}|\mathbf{z})p(\mathbf{z}) \tag{10}$$

Inference proceeds by approximating the posterior $p(\mathbf{z}|\mathbf{x})$ with a variational distribution $q(\mathbf{z})$, optimised by minimising the variational free energy:

$$\mathcal{F}[q] = \mathrm{KL}[q(\mathbf{z})\|p(\mathbf{z})] - \mathbb{E}_{q(\mathbf{z})}[\log p(\mathbf{x}|\mathbf{z})] \tag{11}$$

This objective balances two terms:

- **Accuracy**: How well the model explains observed sensory data
- **Complexity**: The divergence between posterior and prior beliefs

Importantly, this formulation closely resembles the Evidence Lower Bound (ELBO) used in machine learning, both serving as tractable objectives for approximate Bayesian inference [8] [49]. Rearranging terms gives:

$$\log p(\mathbf{x}) \geq \mathbb{E}_{q(\mathbf{z})}[\log p(\mathbf{x}|\mathbf{z})] - \mathrm{KL}[q(\mathbf{z})\|p(\mathbf{z})] = -\mathcal{F}[q] \tag{12}$$

Thus, minimising free energy is equivalent to maximising the evidence lower bound (ELBO). Both provide a tractable bound on the marginal likelihood and guide approximate Bayesian inference. While machine learning typically uses the term ELBO, the Free Energy Principle adopts the language of variational free energy to emphasise its roots in thermodynamics and its application to biological systems as self-organising, entropy-resisting agents. This equivalence reveals a deep formal connection between predictive coding in neuroscience and variational inference in artificial intelligence [8].

4.2 Variational Laplace: A Scalable Engine for Dynamical Inference

While the Free Energy Principle defines the objective—minimising variational free energy—an agent still requires a concrete algorithm to achieve this in practice. Standard optimisation methods like gradient descent can be slow or biologically implausible. A more efficient and neurally inspired alternative is **Variational Laplace**, an inference scheme that underpins neuroimaging tools such as Dynamic Causal Modelling (DCM) [50, 51].

Variational Laplace operates under a key simplifying assumption: that the approximate posterior distribution over hidden states is Gaussian. In this formulation, an agent's beliefs are represented by a mean vector μ and a precision matrix Σ^{-1}, encoding both what is believed and how certain those beliefs are:

$$q(\mathbf{z}) = \mathcal{N}(\mathbf{z}; \mu, \Sigma)$$

The method efficiently updates these beliefs by approximating the local curvature of the free energy landscape using a second-order Taylor expansion. This allows it to compute both the direction and step size for updates—akin to the Newton–Raphson method—enabling faster and more stable convergence than simple gradient descent.

In continuous time, this yields a gradient flow over beliefs:

$$\frac{d\mu}{dt} = -\nabla_\mu \mathcal{F}(\mu)$$

This expresses a central insight of the Free Energy Principle: inference is not a static computation, but a dynamical process unfolding over time. The agent continuously refines its beliefs by traversing a trajectory through state space, driven by prediction errors and modulated by uncertainty.

Variational Laplace provides a scalable and interpretable inference engine that naturally supports this kind of online adaptation—making it well-suited for modelling biological intelligence as well as building adaptive artificial agents. As we show in later sections, this approach can be applied directly to active perception and control tasks, including our toy Active Inference agent for Pong.

4.3 Learning and Adaptation under the Free Energy Principle

While Section 2.3 outlined how action selection can be framed as inference through the minimisation of expected free energy, the Free Energy Principle extends further-offering a unified account of perception, action, and learning as complementary aspects of the same optimisation process.

In this framework, perception corresponds to inferring the current state of the world, action corresponds to selecting policies that minimise future expected surprise, and learning involves updating the parameters of the generative model to improve future predictions [52, 7]. Over time, these model updates reduce long-term free energy by enabling more accurate and efficient inference across changing environments.

This deep integration means that agents can adaptively tune both their beliefs and their internal models in response to uncertainty and experience. Rather than relying on task-specific reward signals or episodic retraining, the agent maintains a coherent internal model that evolves through ongoing experience-supporting continual, goal-directed behaviour.

Taken together, the Free Energy Principle provides more than just a theory of brain function: it offers a principled framework for building autonomous systems that perceive, act, and learn in a dynamically structured world.

4.4 From Principle to Practice: Active Inference in 2D Pong

We implemented a minimal Active Inference agent to control a paddle in a 2D Pong environment (Figure 1). At each time step, the agent receives a visual observation of the current ball and paddle positions and uses Variational Laplace to infer the most likely hidden states of the environment-such as the ball's velocity-under a generative model [50, 6]. This generative model was hand-specified and includes simplified physical dynamics that govern how the ball and paddle evolve over time.

The agent evaluates a discrete set of candidate actions by simulating their consequences under the generative model and computing the expected free energy of each future trajectory. It then selects the action that minimises expected free energy and updates the paddle's position accordingly [7, 8].

Unlike reinforcement learning approaches, which typically learn policies from external reward signals via trial-and-error, this agent uses a model-based, inference-driven strategy. Its behaviour emerges from the minimisation of variational and expected free energy, enabling online adaptation and uncertainty-aware decision making without task-specific supervision or reward shaping.

The full perception–action loop is summarised by four core equations (Table 2), which formalise observation, state inference, action selection, and state transition.

Table 2: Core equations governing the 2D Pong Active Inference loop.

Component	**Equation**
Observation model	$y(t) = g(u(t)) + \omega_t$
State inference (VL)	$\mu(t) \approx \arg\min_\mu \mathcal{F}(\mu) = \frac{1}{2}\|y(t) - g(\mu)\|^2 + \mathrm{KL}[q(\mu)\|p(\mu)]$
Action selection	$a(t) = \arg\min_a \mathbb{E}[G(\pi)]$
State transition model	$u(t+1) = f(u(t), a(t)) + \zeta_t$

5 Toward a Neuro-Inspired AGI Architecture

The theoretical framework developed in this paper-centered on predictive coding, Active Inference, and the minimisation of variational free energy-offers a guiding framework for the development of artificial general intelligence (AGI). Rather than focusing solely on behavioural benchmarks, this approach aims to replicate the underlying computational principles of biological intelligence: inference-driven, uncertainty-sensitive, and dynamically adaptive [5, 8, 53].

However, important gaps remain. Most notably, while the framework assumes access to a generative model, it does not yet provide a general solution for autonomously discovering latent states, learning model structure, or generating suitable priors for new tasks. These are major open challenges for scaling Active Inference to realistic AGI applications.

Thus, while the architecture outlined here is conceptually unified and biologically grounded, it remains aspirational. Turning these principles into engineering tools that can support scalable, general-purpose intelligence will require substantial further research-particularly in the domains of structure learning, planning, and continual adaptation.

5.1 Key Design Principles

A neuro-inspired AGI system would be governed by the following key principles, implemented as an iterative loop of inference, action, and learning:

- **Generative models**: Maintain internal probabilistic models $p(x, z)$ that generate sensory inputs x from latent states z, supporting simulation, counterfactual reasoning, and belief updates [3].

- **Hierarchical structure**: Organise internal representations into layers, where high-level beliefs generate top-down predictions and low-level sensory data propagate bottom-up prediction errors [4].

- **Online variational inference**: Use real-time algorithms such as Variational Laplace to continuously update beliefs $\mu(t)$ about hidden causes, in response to new sensory data [50, 6].

- **Precision weighting**: Dynamically estimate and apply precision (inverse variance) to modulate the influence of prediction errors, enabling adaptive attention, robust inference, and uncertainty-aware behaviour [24].

- **Active inference loop**: Select actions $a(t)$ that minimise expected free energy $G(\pi)$, unifying exploration (epistemic value) and goal pursuit (instrumental value) in a principled way [7]. However, planning via expected free energy minimisation remains computationally demanding, and developing scalable approximations is an open area of research.

- **Continual learning and plasticity**: Adapt the parameters θ of the generative model over time to minimise long-term free energy. Unlike belief updating, this form of structural learning is non-trivial and remains an active research frontier [54].

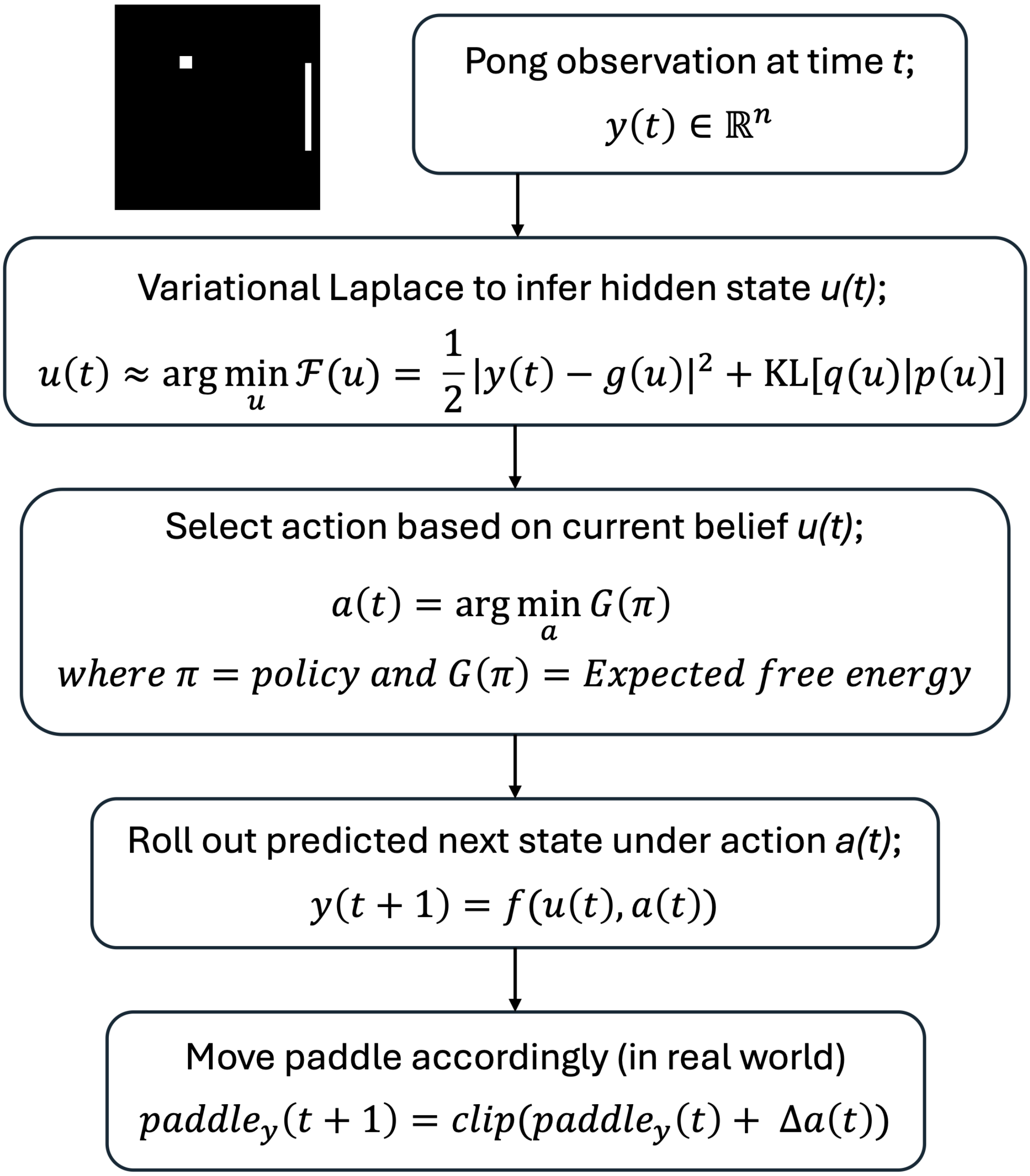
Pong observation at time *t*;
$y(t) \in \mathbb{R}^n$
Variational Laplace to infer hidden state *u(t)*;
$u(t) \approx \arg\min_u \mathcal{F}(u) = \frac{1}{2}|y(t) - g(u)|^2 + \mathrm{KL}[q(u)|p(u)]$
Select action based on current belief *u(t)*;
$a(t) = \arg\min_a G(\pi)$
$where\ \pi = policy\ and\ G(\pi) = Expected\ free\ energy$
Roll out predicted next state under action *a(t)*;
$y(t+1) = f(u(t), a(t))$
Move paddle accordingly (in real world)
$paddle_y(t+1) = clip(paddle_y(t) + \Delta a(t))$

Figure 1: This schematic illustrates the core steps of an Active Inference loop applied to a simple Pong environment. At each time step t, the agent observes the sensory input $y(t)$, which includes ball and paddle positions. Using a generative model $g(\mu)$, the agent performs Bayesian state estimation via Variational Laplace, updating its belief $\mu(t)$ about hidden environmental variables (e.g., ball velocity). Based on this belief, it evaluates multiple candidate policies π and selects the action $a(t)$ that minimises expected free energy $G(\pi)$, balancing uncertainty reduction and goal satisfaction. The chosen action is used to roll forward predictions, update the paddle position, and interact with the environment. This generates a new observation $y(t+1)$, closing the perception–action loop. This process allows the agent to infer, plan, and act continuously in a dynamic, uncertain setting.

While these components provide a compelling blueprint for adaptive intelligence, many of them-especially those involving structural learning and long-horizon planning-face serious scalability challenges in complex environments. Addressing these limitations is central to the research roadmap we outline in the supplementary materials.

5.2 Architecture Sketch

The pseudocode below captures the operational logic of a neuro-inspired AGI system based on predictive coding and Active Inference. At its core, the architecture is an iterative loop that continuously refines beliefs, selects actions, and updates its generative model through the minimisation of variational and expected free energy [55].

The system begins by maintaining internal beliefs μ about hidden causes in the environment, and a generative model $p(x, z; \theta)$ parameterised by θ. At each time step, it receives new sensory input y_t, updates its beliefs via Variational Laplace inference, and uses precision estimates Σ_t to modulate the influence of prediction errors. It then simulates a set of candidate actions (or policies π) and evaluates their expected free energy $G(\pi)$. The optimal action is selected and executed, and model parameters θ are updated to improve future inference and prediction. If a hierarchical structure is present, prediction errors are propagated across layers to coordinate updates across abstraction levels.

This architecture implements perception, action, and learning as deeply intertwined processes-each grounded in a single unifying principle: the minimisation of variational free energy.

Pseudocode: Active Inference Loop for Neuro-Inspired AGI

```
Initialize:
$\theta \leftarrow$ generative model parameters
$\mu \leftarrow$ initial belief over hidden states
$\Sigma \leftarrow$ initial precision (inverse covariance)
Loop over time $t = 1, 2, \ldots$:
  1. Observe sensory input:
       $y_t \leftarrow$ sensory observation from environment
  2. Infer hidden states via Variational Laplace:
       $\mu_t \approx \arg\min_\mu \mathcal{F}(\mu)$
       $\mathcal{F}(\mu) = \frac{1}{2}\|y_t - g(\mu)\|^2 + \mathrm{KL}[q(\mu)\|p(\mu)]$
  3. Update precision:
       $\Sigma_t \leftarrow$ function of current residual and belief uncertainty
  4. Evaluate actions via expected free energy:
       $G(\pi) = \mathbb{E}_q[\log q(\mathbf{z}|\pi) - \log p(\mathbf{z}|\mathbf{x}, \pi)]$
  5. Select and execute action:
       $a_t = \arg\min_a \mathbb{E}[G(\pi)]$
       Execute $a_t$ in environment
  6. Update generative model parameters:
       $\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{F}(\theta)$
  7. Propagate hierarchical updates (if applicable):
       Update higher and lower levels via prediction errors
```

Note: This pseudocode is intended as a conceptual sketch rather than a directly implementable algorithm. Several steps-such as evaluating the expected free energy over all possible future trajectories, or performing stable online updates to the generative model's parameters-pose significant computational and theoretical challenges. In practice, these operations would require substantial approximations, heuristics, or architectural constraints to become tractable. Nevertheless, this schematic illustrates the integrated nature of perception, inference, learning, and action in Active Inference frameworks, grounded in the minimisation of variational free energy.

5.3 Comparison with Existing AI Systems

The neuro-inspired AGI architecture outlined here diverges fundamentally from current deep learning approaches. Conventional AI models-such as convolutional neural networks and large language models-learn static mappings from inputs to outputs using large offline datasets. While effective in narrow domains, these systems typically lack mechanisms for online adaptation, uncertainty representation, and causal reasoning.

By contrast, the architecture proposed here functions as an inference-driven agent: it maintains a generative model of the environment, continuously updates beliefs via variational inference, and selects actions that minimise expected free energy. This enables flexible, context-sensitive behaviour and robust generalisation to novel or nonstationary environments.

This paradigm aligns with a growing class of inference-based AI systems. Notably, Ha and Schmidhuber's *World Models* framework [56] demonstrated that agents equipped with internal generative models can learn compact representations and plan actions in latent space. Similarly, DeepMind's work on Active Inference agents [57, 55] has shown that variational free energy can be used as a unifying objective for perception, action, and learning in deep neural architectures.

However, most of these systems retain amortised or offline components and stop short of full hierarchical, online variational inference. The architecture presented here emphasises biologically plausible, dynamic updates across multiple levels of abstraction-mirroring the continual inference processes of the brain.

In short, whereas conventional AI systems *imitate* intelligent behaviour by replaying learned responses, a neuro-inspired AGI aims to *compute* intelligence through continual inference, adaptation, and goal-directed interaction with the world.

6 Conclusion

Theoretical neuroscience provides not just a metaphorical lens, but a computational blueprint for general intelligence. By casting perception, action, and learning as processes of variational inference, the frameworks of predictive coding and the Free Energy Principle (FEP) offer implementable strategies that unify information processing across cognitive domains [5, 58].

Unlike traditional AI approaches that rely on static architectures trained on fixed datasets, neuro-inspired systems grounded in Active Inference dynamically infer causes, forecast consequences, and adaptively select actions to minimise uncertainty and achieve goals [59, 60]. These agents are not simply reactive; they are generative, predictive, and epistemically driven.

The vision presented here is one of convergence: bridging the gap between modelling the mind and building machines that think. By integrating insights from cortical computation, hierarchical generative models, and variational optimisation, we can move toward artificial systems that share the hallmarks of biological intelligence-flexibility, resilience, and adaptability in uncertain environments.

Ultimately, the path to artificial general intelligence may lie not in mimicking outputs of intelligent behaviour, but in replicating the inferential machinery that underwrites it.

References

[1] Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. Building machines that learn and think like people. *Behavioral and brain sciences*, 2017.

[2] Shane Legg and Marcus Hutter. Universal intelligence: A definition of machine intelligence. 2007.

[3] Rajesh PN Rao and Dana H Ballard. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, 1999.

[4] Karl Friston. A theory of cortical responses. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 2005.

[5] Karl Friston. The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, 2010.

[6] Rafal Bogacz. A tutorial on the free-energy framework for modelling perception and learning. *Journal of Mathematical Psychology*, 2017.

[7] Karl Friston, Thomas FitzGerald, Francesco Rigoli, Philipp Schwartenbeck, and Giovanni Pezzulo. Active inference, agency and anxiety. *Neural Computation*, 2017.

[8] Thomas Parr, Giovanni Pezzulo, and Karl J Friston. *Active inference: The free energy principle in mind, brain, and behavior*. 2022.

[9] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. `https://papers.nips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html`.

[10] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020. URL `https://arxiv.org/abs/2006.11239`.

[11] Tom B Brown, Benjamin Mann, Nick Ryder, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 2020.

[12] Robert Geirhos et al. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2020.

[13] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Michael von Arx, Michael S. Bernstein, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[14] David C Knill and Alexandre Pouget. Bayesian approaches to sensory and motor systems. *Trends in Neurosciences*, 2004.

[15] Karl Friston. The free-energy principle: a rough guide to the brain? *Trends in Cognitive Sciences*, 2009.

[16] Karl J Friston and Klaas E Stephan. Free-energy and the brain. *Synthese*, 159(3): 417–458, 2007. doi: 10.1007/s11229-007-9237-y. URL `https://doi.org/10.1007/s11229-007-9237-y`. PMID: 19325932, PMCID: PMC2660582.

[17] Andre M Bastos, W Martin Usrey, Rick A Adams, George R Mangun, Pascal Fries, and Karl J Friston. Canonical microcircuits for predictive coding. *Neuron*, 2012.

[18] Rodney J Douglas and Kevan AC Martin. Neuronal circuits of the neocortex. *Annual Review of Neuroscience*, 2004.

[19] Seppo P Ahlfors, Samuel R Jones, and Matti S Hämäläinen. Laminar analysis of 7t meg and fmri reveals dissociable prediction error signals in sensory and motor cortices. *Journal of Neuroscience*, 2018.

[20] Stewart Shipp. Neural elements for predictive coding. *Frontiers in Psychology*, 2016.

[21] Karl Friston. Hierarchical models in the brain. *PLoS Computational Biology*, 2008.

[22] Lars Muckli and et al. Contextual feedback to superficial layers of v1. *Current Biology*, 2015.

[23] Karl Friston, Jean Daunizeau, and Stefan J Kiebel. Action and behavior: a free-energy formulation. *Biological Cybernetics*, 2010.

[24] Thomas Parr, Giovanni Pezzulo, and Karl Friston. Generalised free energy and active inference. *Biological Cybernetics*, 2019.

[25] Lancelot da Costa, Thomas Parr, Noor Sajid, Srdjan Veselic, Victor Neacsu, and Karl Friston. The relationship between dynamic programming and active inference: the discrete, finite case. *Neural Computation*, 2021.

[26] Beren Millidge, Alexander Tschantz, Christopher L Buckley, and Karl J Friston. Whence the expected free energy? *Neural Computation*, 2021.

[27] Mikhail I Rabinovich, Pablo Varona, Allen I Selverston, and Henry DI Abarbanel. Dynamical principles in neuroscience. *Reviews of Modern Physics*, 2006.

[28] Thomas Parr and Karl J Friston. Neuronal message passing using mean-field, bethe, and marginal approximations. *Scientific Reports*, 2019.

[29] Christopher D Fiorillo, Philippe N Tobler, and Wolfram Schultz. Discrete coding of reward probability and uncertainty by dopamine neurons. *Science*, 2003.

[30] Markovic. Empirical evidence for predictive coding in the face of uncertainty: A review. *Frontiers in Human Neuroscience*, 2020.

[31] Andy Clark. Whatever next? predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, 2013.

[32] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 2015.

[33] David Silver, Aja Huang, Chris J Maddison, et al. Mastering the game of go with deep neural networks and tree search. *Nature*, 2016.

[34] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? 2019.

[35] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 2017.

[36] German I Parisi, Ronald Kemker, Jose L Part, Christopher Kanan, and Stefan Wermter. Continual lifelong learning with neural networks: A review. *Neural Networks*, 2019.

[37] James C R Whittington, Sophie Deneve, Mikhail Belkin, and Blake A Richards. Disentangling with biological constraints: A theory of functional cell types. *Neuron*, 2022.

[38] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. 2012.

[39] William Lotter, Gabriel Kreiman, and David Cox. Deep predictive coding networks for video prediction and unsupervised learning. 2016.

[40] James C. R. Whittington and Rafal Bogacz. Approximation and learning in predictive coding networks. *PLoS Computational Biology*, 2017.

[41] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[42] Vladislav Ayzenberg and Stella M Lourenco. Recurrent convolutional neural networks approximate predictive processing. *Nature Communications*, 2023.

[43] Andrew Gordon Wilson and Pavel Izmailov. Bayesian deep learning and a probabilistic perspective of generalization. *Advances in Neural Information Processing Systems*, 2020.

[44] Marvin Binz, Eric Schulz, David Krueger, and Brenden M. Lake. Do large language models learn causal representations? *arXiv preprint arXiv:2305.14969*, 2023.

[45] Adam H Marblestone, Greg Wayne, and Konrad P Kording. Toward an integration of deep learning and neuroscience. *Frontiers in Computational Neuroscience*, 2016.

[46] Matthew Botvinick, Sam Ritter, Jane X Wang, Zeb Kurth-Nelson, Charles Blundell, and Demis Hassabis. Deep reinforcement learning and its neuroscientific implications. *Neuron*, 2020.

[47] Heidi Feldman and Karl Friston. Attention, uncertainty, and free-energy. *Frontiers in Human Neuroscience*, 2010.

[48] Karl Friston, Jean Daunizeau, James Kilner, and Stefan Kiebel. Action, perception and free energy. *Biological Cybernetics*, 2010.

[49] David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 2017.

[50] Karl Friston, Jérémie Mattout, Nelson Trujillo-Barreto, John Ashburner, and Will Penny. Variational free energy and the laplace approximation. *NeuroImage*, 2007.

[51] Karl Friston, Nelson Trujillo-Barreto, and Jean Daunizeau. Variational filtering. *NeuroImage*, 2008.

[52] Karl Friston, Jean Daunizeau, and Stefan Kiebel. Reinforcement learning or active inference? *PLoS One*, 2009.

[53] Christopher L Buckley, Chang Sub Kim, Simon McGregor, and Anil K Seth. The free energy principle for action and perception: A mathematical review. *Journal of Mathematical Psychology*, 2017.

[54] Beren Millidge, Alexander Tschantz, Anil K Seth, and Christopher L Buckley. Predictive coding: a theoretical and experimental review. *arXiv preprint arXiv:2210.00577*, 2022.

[55] Beren Millidge, Alexander Tschantz, and Christopher L Buckley. Deep active inference as variational policy gradients. *Journal of Mathematical Psychology*, 2020.

[56] David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018. URL `https://arxiv.org/abs/1803.10122`.

[57] Karl Friston, Richard Rosch, Thomas Parr, Cathy Price, and Howard Bowman. Deep temporal models and active inference. *Neuroscience & Biobehavioral Reviews*, 2018.

[58] Karl J Friston, Thomas Parr, and Richard Rosch. Deep inference: From theoretical neuroscience to the foundations of ai. *Philosophical Transactions of the Royal Society A*, 2022.

[59] Lancelot Da Costa, Thomas Parr, Noor Sajid, Stefan Veselic, Victor Neacsu, and Karl J Friston. Active inference: From variational to deep learning. *Neural Computation*, 2021.

[60] Beren Millidge, Alexander Tschantz, Christopher L Buckley, and Karl J Friston. Predictive coding: a theoretical and experimental review. *Frontiers in Computational Neuroscience*, 2021.