Trends in Cognitive Sciences

CellPress OPEN ACCESS

Opinion

Generating meaning: active inference and the scope and limits of passive AI

Giovanni Pezzulo [1,*], Thomas Parr [2], Paul Cisek [3], Andy Clark [4,5,6], and Karl Friston [7,8]

Prominent accounts of sentient behavior depict brains as generative models of organismic interaction with the world, evincing intriguing similarities with current advances in generative artificial intelligence (AI). However, because they contend with the control of purposive, life-sustaining sensorimotor interactions, the generative models of living organisms are inextricably anchored to the body and world. Unlike the passive models learned by generative AI systems, they must capture and control the sensory consequences of action. This allows embodied agents to intervene upon their worlds in ways that constantly put their best models to the test, thus providing a solid bedrock that is – we argue – essential to the development of genuine understanding. We review the resulting implications and consider future directions for generative AI.

Highlights

Generative artificial intelligence (AI) systems, such as large language models (LLMs), have achieved remarkable performance in various tasks such as text and image generation.

We discuss the foundations of generative AI systems by comparing them with our current understanding of living organisms, when seen as active inference systems.

Both generative AI and active inference are based on generative models, but they acquire and use them in fundamentally different ways.

Living organisms and active inference agents learn their generative models by engaging in purposive interactions with the environment and by predicting these interactions. This provides them with a core understanding and a sense of mattering, upon which their subsequent knowledge is grounded.

Future generative AI systems might follow the same (biomimetic) approach – and learn the affordances implicit in embodied engagement with the world before – or instead of – being trained passively.

Optimism and skepticism about generative AI

Generative AI systems (see Glossary) are taking society by storm, demonstrating impressive capabilities in domains that were previously considered the exclusive province of human cognition. **Large language models (LLMs)** such as ChatGPT[i] generate high-quality text, and text-to-image systems such as DALL-E[ii] generate (in)credible illustrations, all from simple prompts. Multimodal systems complement LLMs with visual (e.g., Flamingo [1]) and sensor data to generate planned actions (e.g., PaLM-E [2] and RT-2[iii]) and affordances [3] for robots, and are perhaps starting to bridge the apparent gap with sensorimotor integration and agency.

These and other generative AI systems – or **foundation models** [4] – are engendering excitement and intense theoretical debate. Does ChatGPT 'understand' what it talks about in the way we do, or is it an example of a 'Chinese room' [5] that transforms symbols without any real understanding? Does it have a 'grasp' on external reality, or is it a mimic that is driven by the sequential statistics of natural language? Can generative AI go beyond the data it has ingested and be creative? Ultimately, is generative AI on a path towards true artificial understanding – namely, to grasp the 'meaning' of words, percepts, and actions – or is it the dénouement of an intrinsically self-limiting approach?

The current debate vacillates between these directions (Box 1), and the development of generative AI with better capabilities – and novel emergent properties – proceeds at a fast pace, along with tools to understand what they do [6–8]. Given this, answering the above questions is perhaps premature. In this treatment we take a different approach: we offer a biophilic perspective on generative AI systems by comparing them to an **active inference** (or **predictive processing**) view of brain and cognition, which foregrounds the notion of **generative models** (or **world models**), but in a biological setting [9,10].

(Inter)action and active inference in biological systems

For any biological system to be sustainable, it must actively restrict itself to characteristic states and counter any perturbations that supplant those states. This is accomplished by physiological

[1]Institute of Cognitive Sciences and Technologies, National Research Council, Rome, Italy
[2]Nuffield Department of Clinical Neurosciences, University of Oxford
[3]Department of Neuroscience, University of Montréal, Montréal, Québec, Canada
[4]Department of Philosophy, University of Sussex, Brighton, UK
[5]Department of Informatics, University of Sussex, Brighton, UK
[6]Department of Philosophy, Macquarie University, Sydney, New South Wales, Australia

[7]Wellcome Centre for Human Neuroimaging, Queen Square Institute of Neurology, University College London, London, UK
[8]VERSES AI Research Lab, Los Angeles, CA, USA

*Correspondence:
giovanni.pezzulo@istc.cnr.it
(G. Pezzulo).

Box 1. The debate around large language models (LLMs) and other generative AI systems

There has been a general skepticism that LLMs command any type of deep understanding of reality. Such skepticism is often rooted in the unhappy experiences that some have had when questioning such systems on complex topics in which they are already expert. In addition, LLMs struggle with causal reasoning and multi-step compositional reasoning [110], and sometimes 'hallucinate' rather than reporting factual information and show 'self-delusions' (i.e., they take their predictions as evidence that the predicted circumstance is true [78]; *cf* circular inference in psychosis in [111]). This suggests a lack of causal understanding of actions and that the apparent meaningfulness of dialogues with LLMs might come from the ease with which we project our mental states and agency into these systems [55,112].

Another related type of skepticism is rooted in their apparent 'disembodiment' and lack of true causal connection to the world about which they so fluently speak [58,113]. An LLM might write movingly about the experience of eating a new breakfast cereal, but no LLM has ever eaten anything. Furthermore, some of the most advanced LLMs show only limited sensitivity to affordances compared to humans [114]. The lack of anchoring on embodied reality motivates novel foundation models for embodied intelligence that include multiple modalities [115][V] or that mimic the visual cortex, rather than starting from language [116].

By contrast, it has also been claimed that foundation models such as LLMs show some form of general intelligence [117,118] and have surprising emergent properties [4,119]. For example, they can generate meaningful answers to university problems [118], analogical reasoning problems [120], and textual descriptions of moves on a chessboard [117]. Although they are only trained with textual input, it has been claimed that they nevertheless develop models of the shape and causal structure of non-linguistic reality, including implicit models of entities mentioned in a discourse [121] and of properties such as space and direction [122], color [123], and theory of mind ([124]; *cf* [125]). Furthermore, they can be used to generate robot plans, even without or with little visual information [126]. This might be because generative AI systems are trained to extract statistical regularities from their inputs, and the regularities of texts and images implicitly distil regularities in our lived world. Under this reading, multimodal information and embodiment would not necessary be preconditions for learning about the causal structure of the world. Linguistic training could provide the same understanding. Support for this view comes from the failure of visual-and-language models (so far) to improve upon purely linguistic models in acquiring useful semantic information [127]. Furthermore, another stream of research suggests that LLMs encode conceptual information in a similar way to vision-based models, where the (structural) similarity means that word **embeddings** and image embeddings self-organize and cluster in the same way in the latent spaces of their respective (language or vision-based) models [128]. The ability of ChatGPT and similar models to engage in meaningful conversation suggests that LLMs might acquire some pragmatic ability for dialogue – and some alignment with human values – through human interaction and a fine-tuning procedure called **reinforcement learning from human feedback** [88].

control, which operates through homeostasis [11], and allostatic behavior, which extends feedback control through the environment [12]. As considered by philosophers [13–15], psychologists [16–20], neuroscientists [9,21,22], and engineers [23–25], the primary function of the brain is not to accumulate knowledge about the world but to control exchanges with the world. Crucially, specific interactions reliably change states of affairs in particular ways (e.g., eating reduces hunger, fleeing from a predator reduces danger, etc.), and we can use this reliability to our advantage. Thus, particular features of the world are meaningful to us because they specify the ways that we can act on the world – what Gibson called 'affordances' [26] – to attain characteristic states that have adaptive value. Responding to affordances is a type of sensorimotor understanding that precedes explicit knowledge of the world, both in evolution [27] and in the course of child development [16].

For many types of interaction, some (implicit or explicit) knowledge of the dynamics of the world is essential [28]. This includes the ability to predict how our actions will influence our state, and to infer the context in which such predictions apply. These are cornerstones of a prominent perspective in cognitive neuroscience called 'active inference'. A key idea here is that, in living organisms, sentient behavior – the capacity to infer states of the world and to act upon it with a sense of purpose [29] – is fundamentally predictive and rests on grounded world models that can generate predictions about the consequences of action [9,10,12,30].

Generative AI shares several commitments with active inference. Both emphasize prediction, and both rest on generative models, albeit differently (Figure 1). Generative AI systems are based on

deep (neural) networks that construct generative models of their inputs via **self-supervised learning**. For example, the training of most LLMs involves learning to predict the next word in a sentence, usually using **autoregressive models** [31] and **transformer architectures** [32]. Once trained on a large corpus of exemplars, the models learned by generative AI afford flexible prediction and can generate novel content (e.g., text or images). Furthermore, they excel in various downstream tasks, such as text summarization and question answering, and can learn from instructions and examples without additional training (i.e., in-context learning [33]). Additional fine-tuning using small, domain-specific datasets permits LLMs to address even more tasks, such as interpreting medical images [34] and writing fiction [35].

In active inference, however, generative models play a broader role that underwrites agency. During task performance, they support inference about states of the extrapersonal world and of the internal milieu, goal-directed decision-making, and planning (as predictive inference). During offline periods, such as those associated with introspection or sleep, generative models enable the simulation of counterfactual pasts and possible futures, as well as a particular form of training 'in the imagination', which together optimize generative models that – crucially – generate the policies of the agent [36–41].

A key difference between the two approaches (Figure 1) is that, although generative AI learns to provide a response when prompted, active inference associates those responses with meaning that is grounded in sensorimotor experience: the words in the question and response about 'going north' or 'south' are associated with the potential for (and the prediction of) movement in physical space – and engages neuronal processes involved in guiding movement in space and predicting its multisensory and affective consequences. The discourse around spatial translations is very different in creatures capable that can move from one spatial location to another, compared to artificial systems with no capacity for movement – even if those systems can learn some aspects of the statistics of spatial translations from sentences in their training set.

What route from generative models to understanding?

The generative models of active inference – and of living organisms – can distil **latent (or hidden) variables** that abstract away from data to afford good explanations and predictions, and might underwrite concept formation. Interestingly, the studies reviewed above speak to the possibility that – in virtue of their predictive training – the latent variables of generative AI likewise come to reflect deep regularities (e.g., emergent linguistic structure in LLMs [42,43]) that might extend beyond the training domain (e.g., non-linguistic regularities for LLMs, such as the relations between looks and tastes). This may be because distilling such knowledge about the world (through language) is the best way to predict the next word. After all, the latent process that generates text rests on people who communicate to pursue their goals. Successful generative models might develop latent variables that capture aspects of this generative process, in the same way that a parrot has an implicit notion of syntax when repeating a heard phrase. Although this remains to be fully assessed, there might be important differences in the ways latent representations are installed in living organisms and generative AI.

An example will help: for humans and other creatures, interactive control exploits particular properties of the world. For instance, a table affords a place to rest a plate, a place to sit, or a place to find shelter during an earthquake. Although each of these meaningful affordances is mechanistically distinct, they are all attached to the same object in the world. Consequently, the concept of a 'table' may serve as a useful (compressed) shorthand for 'the object I can place stuff on, sit on, or hide under'. Thus, the concept is a constellation of latent constructs

Glossary

Active inference: a normative framework under which perception and action are treated as jointly optimizing a variational free energy functional.
Autoregressive model: a statistical model that predicts future data (e.g., words) based on past data (e.g., previous inputs and ensuing predictions).
Embedding: in machine learning, embedding denotes a low-dimensional representation of a discrete variable (e.g., a continuous vector representing a word or another token). An appealing feature of embeddings used in state-of-the-art models (e.g., LLMs) is that items with a similar meaning are close in embedding space.
Foundation models: large-scale generative models (e.g., language or multimodal models) that are trained on broad data at scale and are adaptable to a wide range of downstream tasks.
Generative artificial intelligence (AI) systems: we use the term to refer to AI systems that use large-scale generative models to process and generate various types of information, such as textual information (e.g., LLMs) and multimodal (e.g., text and images) information.
Generative model: a statistical model that describes how observable content is generated from unobservable (hidden or latent) causes; for example, how a visual object generates an image on the retina. It can be used to generate novel, synthetic content (also inverted, to infer causes from observables). Technically, it encodes the joint probability distribution of observables and hidden causes.
Intervention: in the field of causal inference, it refers to the purposeful modification of the state of the world (e.g., by acting) to disclose its causal structure.
Large language models (LLMs): generative models that generate natural (textual) language, usually via self-supervised learning. Some famous examples are bidirectional encoder representations from transformers (BERT) and generatively pretrained transformer (GPT) models.
Latent (or hidden) variable: an internal variable of a generative model. It is called 'latent' or 'hidden' because it cannot be observed, and instead much be inferred.
Precision: the inverse of variance or standard error. Precision weighting is a mechanism in active inference that

that link an object to its action-dependent consequences [44,45]. This perspective is in keeping with embodied cognition studies showing that living organisms learn about objects through sensorimotor experience, and their abstract concepts – such as 'weight', 'size', and 'throwability' – are grounded in modal information [19]. As to the 'grounding' relation itself, we remain agnostic, except to observe that a necessary condition for grounding will usually (perhaps with the exception of a few highly theoretical scientific types) be realized by learning what to expect: namely, through an appreciation of the sensorimotor and interoceptive consequences of self-initiated actions with respect to the object, event, or state of affairs in question.

Language competence itself – comprising semantic and pragmatic abilities – is built on top of knowledge grounded in the sensory modalities [46] and a non-linguistic 'interaction engine' which capitalizes on nonverbal joint actions [47,48] such as moving a table around a tight corner. This competence is bootstrapped during development through collaborative sense-making and child–adult interactions situated in the physical world [49]. However, the question is not (only) how the symbols of language can be connected to non-symbolic processes [46], but rather where the symbols themselves even come from [50,51]. As the example above shows, the sensorimotor interaction comes first, long before symbols appear in both phylogeny [52] and ontogeny [16]. What then is the origin of the symbols, and how can they become detached from the sensorimotor knowledge that grounds their meaning [50]? One simple answer follows from considering the nature of communication.

From an embodied perspective, communication is a type of sensorimotor interaction, albeit one that extends to other creatures in our environment [52]. Consider a human infant that cannot accomplish much on its own. Fortunately, in the niche of helpless human infants there is something called a parent, which has the handy properties of being incredibly complex but also very easy for the infant to control. The baby cries and the parent rushes over to figure out and fix the problem, whether this involves procuring milk or driving at high speed to a hospital. With time, the baby can learn to make distinct noises to produce different outcomes via the parent, and the parent will deliberately help the baby to learn which noises make the parent bring her food versus water versus changing the diaper, and so on. Throughout, the real purpose of making noises is not to convey knowledge but to persuade. Animals do this all the time, from the threat postures of crayfish, to monkeys baring their teeth, to humans uttering 'back off!'. Importantly, the meaning of the communiqué is not in the acoustics or syntax of a given utterance and instead lies in the interaction that the utterance is predicted to induce in those who speak the same language, and the desired consequence of that exchange. The words themselves are only shorthand notation for the meaningful interactions, and they are compact and 'symbolic'. For example, when the baby cries to engage its parent, the noise it makes need not specify the path or necessary foot placements – the parent will take care of all that. For this reason, the interaction between agents is naturally symbolic and purposeful.

Human linguistic communication takes this to extremes of abstraction, but is still grounded by the fundamental context of interactive control. These examples illustrate the fact that we learn the meaning of linguistic symbols as part of pragmatically rich interactions with our conspecifics. The meanings of words supervene on a more primitive understanding of the world that we acquire by interacting with it. Current efforts to model grounded language acquisition in cognitive robotics follow a similar (albeit simplified) approach which involves training models to develop linguistic and symbolic abilities in the context of goal-directed actions [53] and in interactive settings [54]. This contrasts with the approach taken by current LLMs and other generative AI, systems which learn passively from large sets of textual multimodal (e.g., text and video) data.

determines how much weight or impact sensory observations have on belief updating.

Predictive processing: a theoretical approach to the study of living organisms and cognition, based on the idea that the brain is fundamentally a 'prediction machine'. It is sometimes used as a suitcase word to refer to predictive coding, active inference, and other related theories.

Reinforcement learning from human feedback: a methodology to align the outputs of LLMs (e.g., ChatGPT) to human preferences by using feedback from human raters.

Self-attention: see Transformer architecture.

Self-supervised learning: an algorithmic procedure usually adopted to learn generative models without any need for supervision, annotated data, or reward. A typical self-supervised task used in LLMs is learning to predict the next word in a sentence.

Transformer architecture: a machine learning (feedforward) architecture that is particularly effective in training LLMs and other generative AI systems. One of its peculiarities that it uses an attention (or self-attention) mechanism to afford a greater weight (i.e., precision) to the most informative inputs when predicting outputs.

World model: an internal model of the world and its dynamics. In this context, we use it synonymously with 'generative model', but it could be used in different ways.

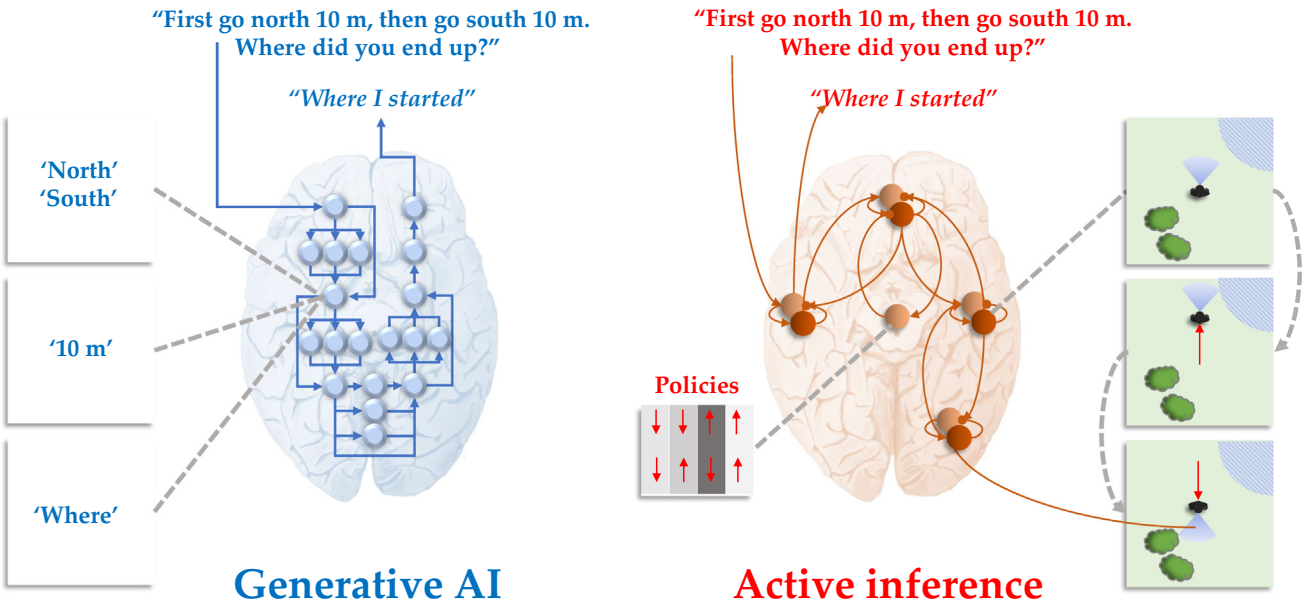

Figure 1. Generative models in generative artificial intelligence (AI) and active inference. This figure highlights the conceptual differences between the ways that generative models support the solution of the same problem: predicting a travel destination. The left schematic is designed to resemble that of a series of transformer networks [32]. These are feedforward architectures based upon a repeated motif with a (multi-head) 'self-attention' structure. This structure allows interactions between different parts of a sequence such that particular elements (e.g., specific words, shown in the boxes) in the sequence are emphasized relative to other elements – effectively picking out salient information that predicts the output. The active inference architecture [9], on the right, illustrates a network of neuronal systems with reciprocal connectivity – of the type found in the brain – supporting recurrent dynamics [142]. The hierarchical structure is evident in the asymmetrical connectivity patterns. Specifically, the 'descending' connections between brain areas are shown with round arrowheads to imply an inhibitory connection, as if we subtract some prediction from a higher level – from the 'ascending' inputs to that region – to compute a prediction error. The 'ascending' connections are shown with a pointed arrowhead to suggest an excitatory connection in which prediction errors drive belief updating and learning. Crucially, in the active inference hierarchy, predictions based upon the policy we might pursue – shown as combinations of 'north' (upwards arrow) and 'south' (downwards arrow) actions – influence hidden states of the world (e.g., my location in allocentric space), which themselves predict both the words we might hear and speak, and the views we might encounter. These inferred hidden states – including where we as a physical agent are in the world, and where we plan to go – are central to biological systems that engage in active inference. In generative AI, a prompt is the input for which there is a desired output. Conversely, in biological exchanges with the world, inputs depend upon action, namely how the world is sampled. Hearing the question shown at the top of the figure updates our beliefs about the sequence of actions we might take (or imagine ourselves taking), which updates predictions about the sequence of locations we will visit (and the visual scenes we will encounter), itself updating our predictions about the next words we will speak to answer the question; an example in a simple navigation setting can be found in [139]. In the brain, the generative models for spatial navigation entail distributed cortical and subcortical (e.g., hippocampal) networks, and achieving advanced machine autonomy might benefit from reproducing the functional properties of these networks [40,143–147].

In sum, our grasp of the meaning of linguistic symbols does not originate from our ability to process natural language but from the more foundational understanding of the lived world that we accumulate by sampling and interacting with it. Although it is possible that the latent variables of generative AI likewise come to reflect statistical regularities of the world (that are inherent in our language and art), these regularities are accessed by skipping the above scaffolding processes – by distilling world knowledge from curated sets of text- or image-based content. Because this content is the product of human communication, generative AI inherits the structure of the meaningful interactions that humans express (e.g., causes precede effects, paragraphs stay on topic, and some phrases are repeated in specific contexts). In the case of an LLM, for example, the meaning to which the words refer is understood by the humans who produced the training text, as well as by those who read the transformed text, but the transformer of the text itself was never provided with any connection to the interactions that lent the words their meaning. Thus, it remains to be seen to what extent generative AI systems trained on human-generated content – that is imbued with meaning for us in virtue of the fact that it was produced by human exchanges – inherit the semantics of that content or whether they merely mimic its

statistical structure [55]. In this respect, the efforts reported above (Box 1) – to assess whether the latent variables of generative AI reflect meaningful color or distance representations – might not be sufficiently diagnostic. These should be complemented by efforts to understand whether these are meaningful for the generative AI systems that use them and not only for us as collocutors. The problem here is that it is not clear what type of analysis would offer a fair test (could we replace driving instructors with LLMs – and would you let them drive your car?). It is as if we encountered an alien species whose window on reality was through our descriptions of the world (Box 2).

In the AI community, models are usually judged based on performance metrics, but good performance on a task done well by humans does not imply that they employ similar processes. For example, despite initial excitement about deep convolutional networks – as models of the primate visual object recognition system [56] – empirical evidence suggests that their operation bears little resemblance to established psychophysical phenomena [57]. Similarity to brains may not be relevant for many engineering applications, but it may foreground viable paths toward general AI. Although it is too early to tell, an analogous lesson may await LLMs and other generative AI systems: will they overcome apparent limitations when given still more data, or is their capacity for understanding inherently limited? An answer would require novel benchmarks that measure the biomimetic ability of generative AI (e.g., embodied in robots) not merely to answer questions but to achieve open-ended goals in the environment [58][iv].

A complementary approach to this question – pursued below – compares how generative AI and active inference acquire generative models, to draw conclusions about what sort of 'grip' on reality these generative models might afford.

Generative model acquisition in generative AI and active inference

> The child does not 'learn', but builds his knowledge through experience and relationships with his surroundings – (Maria Montessori)

Box 2. Word-world: a thought experiment

Imagine an alien lifeform whose only contact with some underlying reality is via a huge stream of words: items that bear real but complex and sometimes imprecise relations to that hidden reality. The hidden reality is our human world populated with cats, pastors, economic depressions, LLMs, elections, and more. Think of this being's access to the stream of words as itself a type of modality, a sensory channel. During its youth, our alien being (let us call it Wordy) found itself driven to try to predict the next item in that sensory stream, inferring underlying patterns that enabled it to do that job surprisingly well. This was good for Wordy's survival.

Wordy has only a single sensory channel. Even so, that single channel bears rich indirect traces of our own much more varied forms of sensory access. Wordy is, however, oddly separate from its own underlying world. When we humans act in our world, we are regulated by the very world we are attempting to describe and engage. When I try to pick up the cup I see in front of me, there is a possibility that I will fail. The true location of the cup in space relative to me, an embodied organism, constantly holds my visuomotor action routine to account. Other humans also hold me to account, and there too learning from my mistakes is possible. We are constantly answerable to a web of regulative interactions that anchor us, both individually and collectively, to the world.

Wordy, by contrast, is only very indirectly regulated by the world that the texts it was trained on describe. It is held to account only by a successor relation defined over words. Even within its own domain of action (outputting more words), Wordy was never in the business of needing to estimate the consequences of its actions. Nor could it learn from failures to correctly estimate those consequences, or select actions designed to test or improve its own state of information.

This lack of anchoring marks a real difference both from biological organisms and the active inference systems discussed in the text. Could we alter Wordy's survival niche in some way that remedies this shortfall? Perhaps. However, as things stand, the type of generative model that Wordy commands remains unlike that of an embodied organism whose actions constantly expose them to the very world they are attempting to model.

Generative AI and living organisms (seen from the active inference perspective) acquire their generative models through different training regimes. Although both systems might learn about the same concepts (e.g., what it means to go north versus to go south), they do it differently (Figure 2).

Living organisms (and active inference systems) acquire their generative models by engaging in sensorimotor exchanges with the world – and conspecifics – and learning the statistical regularities of such interactions. These interactions enable sensorimotor predictions that shape and structure perception of the world – and other agents – and afford our causal understanding of action and effects. Empirical and cognitive robotics studies have shown the importance not only of sensory-based predictions that do not demand action *per se* [59–63] but also of active engagement and moving in the world as a means to develop generative models and specific forms of understanding within them. We move from our inception, and it is possible that grounding concrete and even abstract concepts requires action and action prediction [19,45,52,64,65] in the spirit of active sensing and learning. It is by moving that we acquire representations of affordances, space, object, scene, and a sense of self and agency [10,64,66]. For example,

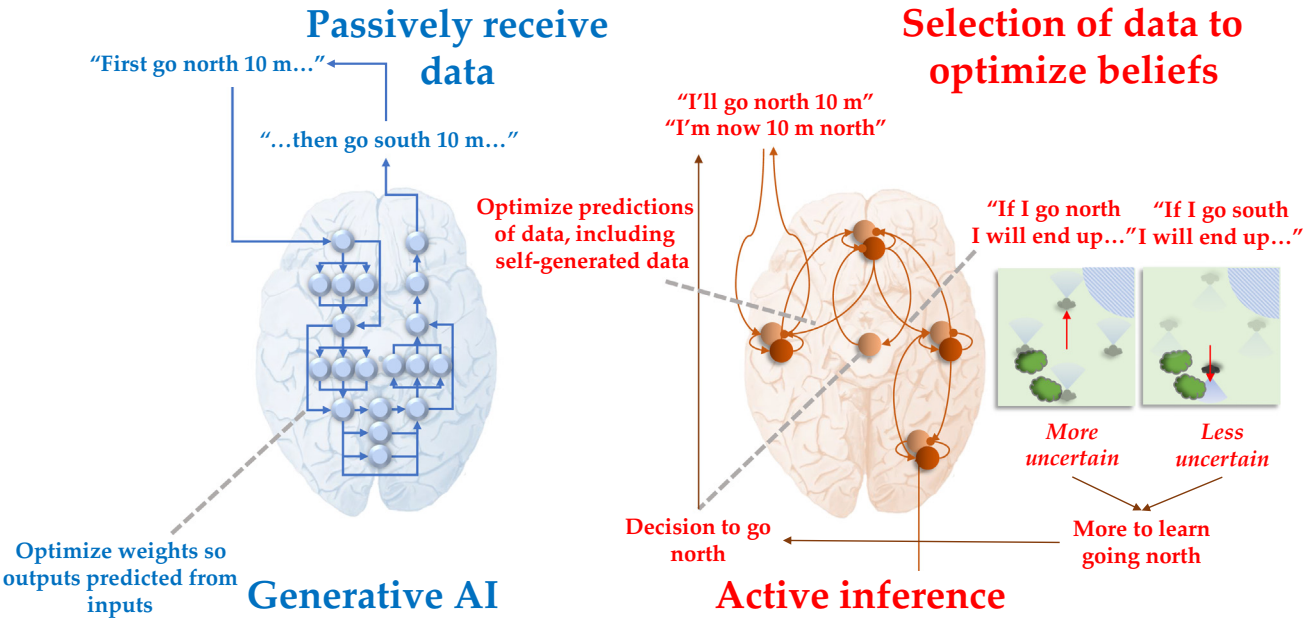

Figure 2. How generative artificial intelligence (AI) and biological systems might learn generative models to solve the wayfinding task of Figure 1. (Left) Cartoon of the pretraining process for generative AI systems in which they are passively presented with (large quantities) of data. The weights of the network are then optimized such that their outputs are more probable given the inputs. State-of-the-art models often include subsequent fine-tuning in a (semi)supervised manner [88]; however, this still relies upon passive presentation of labeled data or self-generated outputs paired with rewards. (Right) By contrast, the generative models that underwrite active inference [148] involve reciprocal interactions with the world. This means that our current beliefs about the world can be used to select those data that have 'epistemic affordance' – in other words they are most useful to resolve our uncertainty about the data-generating process. In the process of learning what it means to go north or south, we may be more or less certain about the location we will end up in under each of these actions (shown here with a relatively high confidence of ending up in the southern position if going south, but more uncertainty in going north). By choosing to go north (and observing being 10 m north from our starting location), we are now in a better position to resolve our uncertainty and optimize our predictions. Beliefs about the causes of our data are an important part of this process of curiosity, exploration, or information seeking [80]. However, these beliefs may easily be neglected in the process of function approximation used in current generative AI systems, where all that matters is the desired output. The neuroanatomical diagrams in this figure are intended purely for illustrative purposes and are not to be taken seriously as anatomical hypotheses – which would distract from the focus of this paper on AI. However, process theories have been developed from active inference frameworks (e.g., [93,103,149]) to which we direct interested readers. Broadly, we might expect planning and policy selection to rely upon networks involving cortical and subcortical regions (e.g., cortico-basal-ganglia-thalamo-cortical loops) in which asymmetrical neuronal connectivity patterns between different cortical regions reflect communication between different hierarchical levels.

various studies show that the hippocampal formation and the entorhinal cortex develop spatial codes – and possibly codes for more abstract conceptual spaces [67] – by path-integrating self-motion information [68]. Likewise, studies of frontoparietal cortex suggest that it contains specialized circuits for detecting affordances and using them to guide specific types of movements [69–71]. In living organisms, these (and other) circuits support a core sensorimotor understanding of reality – an embodied intelligence – that grounds our knowledge and world models, thus providing a foundation for conceptual and abstract thought. In addition, it grounds our ability to generalize to novel tasks without the extensive retraining that is required by current AI systems [72–76].

By contrast, LLMs such as ChatGPT learn by passively ingesting large corpora and by performing self-supervised tasks (e.g., predicting words). Other generative AI systems use the same approach, albeit with other data formats such as pictures and sometimes robot sensor data [2]. The 'understanding' of current generative AI systems is not action-based and is essentially passive – it reflects statistical (rather than causal) regularities evidenced within large datasets of curated data (e.g., text, images, code, videos): they generate content from content, not from causes. Without the capability to actively select their observations – and to make **interventions** during training – generative AI may be unable to develop causal models of the contingencies between actions and effects, or of the distinction between predictions and observations [77,78].

Without a core understanding of reality (or a 'common sense'), current AI systems are brittle: they can learn specific tasks but often fail when presented with close variants of the same tasks because they learn inessential features that do not generalize[iv]. Technically, this type of overfitting reflects a focus on predictive accuracy at the expense of model complexity (Box 3). This may limit the types of learning that are possible using LLMs and generative AI. This is a matter of debate because some believe that autonomous machine intelligence will emerge by enriching and scaling internal models, letting them learn as much as possible from textual knowledge or by passive video observation. However, this type of 'scaling up' might be intrinsically limited. For example, it has been proven that learning a context-sensitive programming language is not possible using any finite set of exemplar code, and it is likely that the challenge is even greater for inferring meaning from natural languages [79]. Similarly, we believe that pursuing an exclusively passive methodology to learn from specific samples of text or videos is unlikely to lead to a core understanding of the real-world causes and effects that are responsible for producing those samples. A more promising path – to artificial general intelligence – combines real-world interactions with sensorimotor predictions.

Given their different training regimes, generative AI systems and active inference agents have different ways to determine what is salient, and what to attend to. In the transformer architectures used in generative AI, attention (or **self-attention**) refers to a mechanism that assigns greater or lower weight to their (extremely long) inputs, thereby filtering them. In active inference, attention encompasses both this filtering role (by varying the **precision** of predictions and sensory information) and the active selection of salient data from the environment that resolves uncertainty. Active inference systems can perform 'experiments' and elicit information that is expected to maximize information gain. This curiosity is ubiquitous in living organisms, but is more challenging to obtain with passive learning [80].

A key aspect of natural intelligence is embodiment. Creatures acquire their generative models under the selective (evolutionary) pressure of adaptive control that serves metabolic needs and survival [27,81]. It has been speculated that this grounding engenders our emotions by reflecting a sense of 'mattering to me' that structures and informs the ways we process information

Box 3. Tradeoffs in active inference between complexity and accuracy, and between exploration and exploitation

The imperative to maximize the evidence (also known as marginal likelihood) for generative (i.e., world) models of how observations are caused has been an essential feature of recent trends in theoretical neurobiology, machine learning, and AI. Evidence-maximization explains both sense-making and decision-making in self-organizing systems from cells [129] to cultures [130]. This imperative can be expressed as minimizing an evidence bound, termed 'variational free energy' [131], that comprises complexity and accuracy:

Free energy = model complexity–model accuracy [1]

Accuracy measures goodness of fit, whereas complexity measures the divergence between prior beliefs (before seeing outcomes) and posterior beliefs (afterwards). More intuitively, complexity scores the information gain or (informational and thermodynamic) cost of changing one's mind. This means that evidence-maximization is about finding an accurate explanation that is minimally complex (*cf* Occam's principle). Importantly, in the context of generative and generalized AI, it implies optimizing generative models such that they explain data more parsimoniously, with fewer parameters [38].

In an enactive setting – apt for explaining decision-making – beliefs about 'which plan to commit to' are based on the expected free energy under a plausible plan. This implicit planning as inference can be expressed as minimizing the expected free energy [9,132]:

Expected free energy = risk (expected complexity) + ambiguity (expected inaccuracy) [2]

Risk is the divergence between probabilistic predictions about outcomes, given a plan, relative to prior preferences. Ambiguity is the expected inaccuracy. An alternative decomposition is:

Expected free energy = expected cost − expected information gain [3]

The expected information gain underlies the principles of optimal Bayesian design [80], whereas the expected cost underlies Bayesian decision theory [133]. In short, active inference appeals to two types of Bayes optimality and subsumes information- and preference-seeking behavior under a single objective.

Free-energy minimization operates both during task performance and during offline periods, such as when the brain is at rest. Minimizing free energy during offline periods optimizes the generative model for future use, even in the absence of data; for example, reducing model complexity by pruning irrelevant parameters or self-generating data through 'generative replay' can go beyond experienced data to encompass counterfactual (but plausible) events [36–40]. Finally, during evolution, free-energy minimization could endow animal brains with prior structure encoded in species-specific circuitry [27,96,134].

[21,82,83], and that imbues our world models with meaning and purpose. Active inference models this aspect of agency by using the construct of 'interoceptive prediction' [84–87]. This provides a firm ground to evaluate the courses of action that increase or decrease the viability of an organism, and ultimately to determine what matters and what does not. Importantly, interoceptive prediction, exteroceptive prediction, and proprioceptive (action-guiding) prediction are all co-computed as living organisms go about the task of living. In this way, active inference may naturally scale up in ways that do not seem to have clear analogs in the sessile, data-fed methods used by generative AI, in which learning and fine-tuning are implemented sequentially [88].

A related point is that, to maintain bodily viability and pursue their goals, living organisms cannot passively wait for the next input but need to proactively engage in purposeful (and sometimes risky) interactions with the world. This requires generative models that ensure behavioral flexibility in the face of careful tradeoffs; for example, between exploratory and exploitative behavior, stay-or-leave decisions, and so on. Furthermore, generalizability requires generative models that are not merely accurate but are also parsimonious (and thereby energy-efficient). Depending on the ecological niche, this tradeoff might favor sophisticated (e.g., temporally and hierarchically deep) generative models that encompass a hierarchy of timescales in action and perception [81], versus minimalistic generative models that afford accurate control without forming rich

representations of the environment [89–91] such as the generative models (e.g., central pattern generators) for action cycles in simple organisms [92]. In active inference, the tradeoffs between exploratory and exploitative behavior – and between the efficiency and accuracy of generative models – are all gracefully resolved by pursuing the imperative of free-energy minimization (Box 3). Solving these tradeoffs evinces flexible forms of control that balance the cost-benefits of low- to high-level goals [93] and of habitual versus goal-directed policies [94]. Context-sensitive, flexible control of this type is not yet enabled in generative AI, where there is generally only one fixed form of inference or 'response', and has a fixed budget[iv].

Finally, there is a key difference regarding the phylogenetic trajectories (or training curricula) that living organisms and generative AI systems follow. In organisms like us, abstract thought and linguistic knowledge are grounded in the circuits that supported sensorimotor predictions and purposive control in our evolutionary ancestors [22,27,95,96]. In other words, linguistic abilities develop on top of grounded concepts, even if they can – to some extent – become 'detached' from the sensorimotor context [50].

We believe that the confluence of these many factors (anchored, multi-timescale predictions of the sensory consequences of our own actions and those of others, in constant dialogue with predictions of our own internal physiological states) is responsible for authentic understanding. The literature on 'meaning' encompasses multiple phenomena. However, what all varieties of (authentic) meaning have in common, we suggest, is that they are grounded in, or built upon, a basic grasp of the sensorimotor and interoceptive consequences of our own actions. Meaningful activity patterns (or pragmatic representations) emerge from the capacity to predict and control simple behavioral strategies. Some organisms endogenously generate these representations, and detach them from the context in which they were initially developed and from the mandatory link to specific sensory inputs, action execution, and behavioral state. These detached representations retain their grounding but also afford advanced cognitive capacities such as planning, imagination, and communication about 'what is not there'. For example, grounded representations of food-related affordances could be endogenously generated when talking about food, remembering it, or selecting restaurants, in the absence of food-related cues or hunger. Sophisticated mental life [16] might originate with this capacity for detachment, marking a shift from the functions of pragmatic representations, such as action selection, to those of semantic or descriptive representations, such as planning, imagination, communication, and contemplation. In turn, this shift enriches meaning and understanding through social interaction and the capacity to engage in more sophisticated world interactions – planned behaviors and the prediction of distal action consequences – that engage temporally deep generative models [9]. In this respect, authentic understanding cannot be separated from agentic understanding (i.e., 'agency') and a sense of prediction and authoring of our sensorium, ranging from immediate to distal, counterfactual, and detached (Box 4).

Current generative AI is following a path that differs fundamentally from the phylogenetic trajectories of living organisms described above: they are following an 'inverse phylogeny' that starts from acquiring knowledge directly from text, alone or with other modalities. This approach is motivated by technological considerations, such as the availability of large textual corpora and the effectiveness of transformer architectures on textual learning and prediction. An interesting question arises here: will further scaling up of generative AI move in the opposite direction to natural intelligence and active inference – that foregrounds statistical and thermodynamic efficiency?

What way forward?

Given the above discussion, one might ask: what are the most promising future directions for generative AI? One might imagine future developments along several lines. One axis is a

Box 4. A bright line between generative AI and active Inference?

What is the bright line between generative AI and active inference? The answer is straightforward: generative models of active inference endow generative AI with agency because they include the consequences of action and equip artificial (and natural) intelligence with the ability to plan [103,106,107]. The notion of agency rests upon generative models that have a broader scope than is usually considered in AI, and that provide causal understanding at multiple levels – from the sensory observations that one gathers by acting (e.g., the sensations associated with drinking fresh water) to the type of things that are usually associated with 'intuitive theories' of physics and psychology [135,136], such as the consequences of acting upon physical objects (e.g., squeezing a plastic bottle of water) and of interacting with other people (e.g., asking a friend for a bottle of water), up to a level that considers what matters for an organism (e.g., a prediction of the physiological consequences of drinking, such as the expected resolution of thirst [84–87]). Living organisms acquire a sense of 'mattering' because they learn generative models under selective pressure to satisfy metabolic needs and remain within viable states. Their 'authentic' understanding of reality is – we argue – grounded in their agentive, purposeful interactions with the embodied world, including other agents: interactions that enable agents to become 'authors' of their sensorium. This embodied intelligence – and the early connection to sensorimotor reality – provides a common ground for conceptual and linguistic knowledge [72,73].

Similarly, active inference agents generate content by acting on – or intervening in – the world in which they operate. Figure 2 offers an example of this: it shows an active inference agent that selects navigation actions to resolve its uncertainty about its location: an epistemic imperative that is often a precondition for the pragmatic imperative of reaching a goal destination [137]. During linguistic exchanges, instead of generating content that sounds like a question, an agent would ask questions that resolve uncertainty about some state of affairs or achieve pragmatic goals. In short, active inference is purposeful. The consequences of behavior have meaning for that agent, in exchange with her world. This type of modeling is fundamentally different from LLM and can be used to model dyadic interactions where agents have epistemic skin in the game [138] and agents that can explain themselves [139]. This early work is at a small scale, and the development of active inference agents that successfully operate in the real world will require solutions to various conceptual and technical challenges, such as developing more efficient methods to plan ahead, and building grounded world models that support embodied interactions [105,140,141]. However, we believe that this early work exemplifies a promising path to artificial understanding, via agency.

The field of generative AI is increasingly moving towards multimodal and embodied settings, for example by learning from egocentric videos that show sensorimotor actions [116] and by providing sensorimotor streams (e.g., visual inputs and robot controls) along with linguistic streams to transformers [2,3]. Although it is possible to learn a lot by predicting videos and by coupling controls and linguistic inputs, the ensuing agents would have no control over their sensorimotor experiences and cannot engage in purposive exchanges with the environment – or in useful interventions that scaffold our causal understanding of the world. Living organisms start with agency from the beginning [64,66]; it remains to be seen whether bootstrapping learning with limited or no agency, and then adding agency at a later stage, is sufficient to build authentic intelligence in future AI systems (see Outstanding questions).

continuum between simpler and more complex models. The complexity here reflects the number of model parameters and their training data. A second issue concerns the type of inputs used for training (e.g., textual, visual, multimodal), perhaps to exploit their complementarities and synergies. A third axis is the addition of extra capabilities, as exemplified by generative agents that engage in simulated dialogue in virtual environments [97] and by commonsense reasoning systems [98]. A fourth axis regards various training and 'engagement' regimes that range from passive ingestion of curated data versus active selection of data through embodied interactions with the world (and others), and which include the pursuit of intrinsic (i.e., epistemic) goals while learning about the world (note that the notions of action and interaction generalize beyond the movements of the physical body; see Outstanding questions).

Current efforts to scale up generative AI systems focus on increasing complexity but with little emphasis on actively selecting their training corpus; in other words, by selecting 'smart' data that optimize active learning and inference. We believe this is a missed opportunity. The 'meaningful anchoring' characteristic of natural intelligence might rest on instantiating an (implicit) generative model of the sensory consequences of the agent's own actions, namely the epistemic and instrumental affordances implicit in an embodied interaction with the world [12]. The resulting 'core understanding' of concepts such as effort, resistance, weight, inertia, and cause and effect might then later be leveraged using essentially passive (LLM-style) resources trained on huge

datasets to deliver something closer to a (super)human understanding of the lived world, perhaps even surpassing our capabilities for flexible behavior and abstract thought. This approach would therefore not simply recapitulate the ways in which living organisms evolve but would exploit the unprecedented possibilities of generative AI systems to learn from large corpora. In our opinion, this synergy is not likely to be achieved by first building up larger LLMs and then connecting them to the world, and instead could be better realized using an interaction-first, LLM-style-last method. Of course, such a strategy has not yet been systematically investigated, and it remains to be established whether it will lead to more advanced and general AI.

Finally, one issue we have deliberately not addressed is the role, if any, of qualitative conscious experience in the generation of (what we are calling) authentic meaning. It remains possible that conscious experience of this type (also known as the experience of 'qualia' or 'phenomenal consciousness' [99]) is a further necessary condition for the appreciation of true meaning. However, it is also possible – and we think more likely – that it is the other way around. That is, within systems that generate meaning, qualitative experience might in some cases occur whenever the right types of (temporally deep, self-model involving) generative models are used to predict and select sensorimotor interactions (including interactions with others). Some developments of these ideas can be found in [29,82,100,101]. Fortunately, nothing in the present discussion requires an answer to these questions.

Concluding remarks

A practical consideration – that inherits directly from an enactivist perspective – is the distinction between generative AI and generalized AI which involves active inference and learning. Both rest upon the implicit or explicit use of generative or world models [102–105]. However, generative AI is limited to generating content (images, code, or text) of the type that we would generate given the same prompt or context. Conversely, active inference is in the game of generating the causes of content in the service of action selection, also known as 'planning as inference' [106–108]. This has several foundational implications. First, planning entails agency, in the sense that only agents are equipped with a generative model of the consequences of their actions. Second, it means that generative models need to be learned through sensorimotor experience via exchange with a world that is actionable – in other words they are grounded world models. In short, generalized AI needs to experience the consequences of its actions. This provides agents with information that directly (and efficiently) reveals the causal structure of the world, relative to information gleaned from a corpus of data that only implicitly reflects that structure. The implicit learning of affordances is fundamentally different from learning the statistical regularities in data or content generated by others. Practically, this means that generative AI is not necessarily the best technology that could be deployed in autonomous robots or vehicles. Furthermore, because it has no notion of epistemic affordance, it will not be apt for active learning or applications that rest upon artificial curiosity or insight [38,104]. Addressing these limitations requires better models of embodied intelligence [58][iv].

Despite these differences, the current wave of generative AI systems can impact on our ecosystems in interesting ways. They do not simply throw our own understandings back at us (although they do that, for obvious reasons). They also package and repackage those understandings and can, with mixed results, suggest bridges between distant parts of the world-model we have uploaded into our various data streams. This positions them to play a role in something that we believe to be crucial but under-theorized – the way in which we humans repeatedly externalize our thoughts and ideas, thus creating new structured objects for critical scrutiny [109]. Generative AI, by finding faint and distant patterns – ones we may have missed in our own material trails and then repackaging them according to arbitrary prompts – offers a golden opportunity to take this

Outstanding questions

Given that generative AI systems are in the public eye, how do we provide a veridical assessment of their capabilities and answer people's questions? How do we avoid the 'Eliza effect' – the tendency to anthropomorphize the behavior of advanced AI systems?

When evaluating generative AI systems, can we trust our intuitions about human understanding, or do we need a more nuanced notion of 'understanding' that goes beyond classical dichotomies and speaks to the diverse capabilities of living and artificial systems?

How important is the agent–environment interaction in bootstrapping meaning – and in forming the grounding that other cognitive faculties (e.g., linguistic learning) rest upon? How much of this meaning-generating interaction is necessary for LLMs? Is the role of agent–environment interaction smaller than has been assumed by embodied cognition theories, given the volume of information that has already been uploaded into the word matrix?

Is the 'inverse phylogeny' evinced by generative AI systems sufficient to acquire meaning and an authentic understanding – without an initial grounding in sensorimotor exchanges with the world?

What types of actions are necessary for autonomous systems to acquire a grounded understanding of reality? Embodied and action-based theories of cognition assign importance to perception–action loops to ground knowledge. However, action is not only physical movement. There are all manner of actions, including communication, which have meaningful consequences.

Is it possible that processes that align generative AI systems with human values (e.g., reinforcement learning from human feedback) also imbue them with a form of 'mattering' – and of 'prior preferences' similar to those of active inference systems?

If we develop novel AI systems that – like living organisms – select and pursue their goals autonomously, how do we ensure that their goals are aligned

distinctively human form of epistemic self-engineering to a whole new level, thereby allowing us to materialize and engage hitherto hidden aspects of our cumulative world-model.

It could be argued that generative AI is one of the most beautiful and important inventions of the century – a 21st-century 'mirror' in which we can see ourselves in a new and revealing light. However, when we look behind the mirror, there is nobody there.

with human values? In parallel with the development of more efficient AI systems, ongoing vigilance of the ethical implications of these advances is imperative.

Acknowledgments

This research received funding from the EU Horizon 2020 Framework Programme for Research and Innovation under grant agreements 945539 (Human Brain Project SGA3) to G.P. and K.F., and 952215 (TAILOR) to G.P.; the European Research Council under the agreements 820213 (ThinkAhead) to G.P. and 951631 (XSCAPE: Material Minds) to A.C.; the Natural Sciences and Engineering Research Council of Canada (RGPIN/05345) to P.C., the Wellcome Centre for Human Neuroimaging (205103/Z/16/Z) to K.F., a Canada–UK Artificial Intelligence Initiative (ES/T01279X/1) to K.F., and Recovery and Resilience Plan (PNRR) MUR projects PE0000013-FAIR and IR0000011–EBRAINS-Italy to G.P.

Declaration of interests

K.F. holds a chief scientific adviser position at VERSES AI. The other authors declare no conflicts of interest.

Resources

[i]https://openai.com/blog/chatgpt

[ii]https://openai.com/product/dall-e-2

[iii]https://robotics-transformer2.github.io/

[iv]https://openreview.net/pdf?id=BZ5a1r-kVsf

References

1. Alayrac, J.-B. *et al.* (2022) Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems* (Vol. 35), pp. 23716–23736, NeurIPS
2. Driess, D. *et al.* (2023) PaLM-E: an embodied multimodal language model. *ArXiv* Published online March 6, 2023. https://doi.org/10.48550/arXiv.2303.03378
3. Ahn, M. *et al.* (2022) Do as I can, not as I say: grounding language in robotic affordances. *ArXiv* Published online April 4, 2022. https://doi.org/10.48550/arXiv.2204.01691
4. Bommasani, R. *et al.* (2021) On the opportunities and risks of foundation models. *ArXiv* Published online August 16, 2021. https://doi.org/10.48550/arXiv.2108.07258
5. Searle, J.R. (1980) Minds, brains, and programs. *Behav. Brain Sci.* 3, 417–424
6. Lampinen, A.K. *et al.* (2022) Can language models learn from explanations in context? *ArXiv* Published online April 5, 2022. https://doi.org/10.48550/arXiv.2204.02329
7. Binz, M. and Schulz, E. (2023) Using cognitive psychology to understand GPT-3. *Proc. Natl. Acad. Sci. U. S. A.* 120, e2218523120
8. Srivastava, A. *et al.* (2022) Beyond the imitation game: quantifying and extrapolating the capabilities of language models. *ArXiv* Published online June 9, 2022. https://doi.org/10.48550/arXiv.2206.04615
9. Parr, T. *et al.* (2022) *Active Inference: The Free Energy Principle in Mind, Brain, and Behavior*, MIT Press
10. Clark, A. (2015) *Surfing Uncertainty: Prediction, Action, and the Embodied Mind*, Oxford University Press
11. Cannon, W.B. (1929) Organization for physiological homeostasis. *Physiol. Rev.* 9, 399–431
12. Pezzulo, G. and Cisek, P. (2016) Navigating the affordance landscape: feedback control as a process model of behavior and cognition. *Trends Cogn. Sci.* 20, 414–424
13. Dewey, J. (1896) The reflex arc concept in psychology. *Psychol. Rev.* 3, 357–370
14. Clark, A. (1998) Embodied, situated, and distributed cognition. In *A Companion to Cognitive Science* (Bechtel, W. and Graham, G., eds), pp. 506–517, Blackwell
15. Merleau-Ponty, M. (1945) *Phènomènologie de la Perception*, Gallimard
16. Piaget, J. (1954) *The Construction of Reality in the Child*, Ballentine
17. Gibson, J.J. (1979) *The Ecological Approach to Visual Perception*, Lawrence Erlbaum Associates Inc
18. Powers, W.T. (1973) *Behavior: The Control of Perception*, Aldine
19. Barsalou, L.W. (2008) Grounded cognition. *Annu. Rev. Psychol.* 59, 617–645
20. Glenberg, A.M. (2010) Embodiment as a unifying perspective for psychology. *Wiley Interdiscip. Rev.* 1, 586–596
21. Quigley, K.S. *et al.* (2021) Functions of interoception: from energy regulation to experience of the self. *Trends Neurosci.* 44, 29–38
22. Cisek, P. and Kalaska, J.F. (2010) Neural mechanisms for interacting with a world full of action choices. *Annu. Rev. Neurosci.* 33, 269–298
23. Wiener, N. (1948) *Cybernetics: or Control and Communication in the Animal and the Machine*, MIT Press
24. Ashby, W.R. (1952) *Design for a Brain*, Wiley
25. Brooks, R.A. (1991) Intelligence without representation. *Artif. Intell.* 47, 139–159
26. Gibson, J.J. (1977) The theory of affordances. In *Perceiving, Acting, and Knowing: Toward an Ecological Psychology* (Shaw, R. and Bransford, J., eds), pp. 67–82, Routledge
27. Cisek, P. (2019) Resynthesizing behavior through phylogenetic refinement. *Atten. Percept. Psychophys.* 81, 2265–2287
28. Conant, R.C. and Ashby, W.R. (1970) Every good regulator of a system must be a model of that system. *Int. J. Syst. Sci.* 1, 89–97
29. Friston, K. *et al.* (2023) Path integrals, particular kinds, and strange things. *Phys Life Rev* 47, 35–62
30. Hohwy, J. (2013) *The Predictive Mind*, Oxford University Press
31. Brown, T. *et al.* (2020) Language models are few-shot learners. In *Advances in Neural Information Processing Systems* (Vol. 33), pp. 1877–1901, NeurIPS
32. Vaswani, A. *et al.* (2017) Attention is all you need. In *Advances in Neural Information Processing Systems* (Vol. 30), pp. 5998–6008, NeurIPS

33. Liu, J. *et al.* (2021) What makes good in-context examples for GPT-3? *ArXiv* Published online January 17, 2021. https://doi.org/10.48550/arXiv.2101.06804
34. Chambon, P. *et al.* (2022) Adapting pretrained vision-language foundational models to medical imaging domains. *ArXiv* Published online October 9, 2022. https://doi.org/10.48550/arXiv.2210.04133
35. Yuan, A. *et al.* (2022) Wordcraft: story writing with large language models. In *27th International Conference on Intelligent User Interfaces*, pp. 841–852, Association for Computing Machinery
36. Pezzulo, G. *et al.* (2021) The secret life of predictive brains: what's spontaneous activity for? *Trends Cogn. Sci.* 25, 730–743
37. Hinton, G.E. *et al.* (1995) The 'wake-sleep' algorithm for unsupervised neural networks. *Science* 268, 1158–1161
38. Friston, K. *et al.* (2017) Active inference, curiosity and insight. *Neural Comput.* 29, 2633–2683
39. Barron, H.C. *et al.* (2020) Prediction and memory: a predictive coding account. *Prog. Neurobiol.* 192, 101821
40. Stoianov, I. *et al.* (2022) The hippocampal formation as a hierarchical generative model supporting generative replay and continual learning. *Prog. Neurobiol.* 217, 102329
41. Pezzulo, G. *et al.* (2019) Planning at decision time and in the background during spatial navigation. *Curr. Opin. Behav. Sci.* 29, 69–76
42. Manning, C.D. *et al.* (2020) Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proc. Natl. Acad. Sci.* 117, 30046–30054
43. Goldstein, A. *et al.* (2022) Shared computational principles for language processing in humans and deep language models. *Nat. Neurosci.* 25, 369–380
44. Bastos, A.M. *et al.* (2012) Canonical microcircuits for predictive coding. *Neuron* 76, 695–711
45. Borghi, A.M. *et al.* (2019) Words as social tools: language, sociality and inner grounding in abstract concepts. *Phys Life Rev* 29, 120–153
46. Harnad, S. (1990) The symbol grounding problem. *Phys. D Nonlinear Phenom.* 42, 335–346
47. Pezzulo, G. (2011) The 'interaction engine': a common pragmatic competence across linguistic and nonlinguistic interactions. *IEEE Trans. Auton. Ment. Dev.* 4, 105–123
48. Levinson, S.C. (2006) On the human 'interaction engine'. In *Roots of Human Sociality: Culture, Cognition and Interaction* (Enfield, N.J. and Levinson, S.C., eds), Routledge
49. Di Paolo, E.A. *et al.* (2022) Laying down a forking path: tensions between enaction and the free energy principle. *Philos. Mind Sci.* 3, 2
50. Pezzulo, G. and Castelfranchi, C. (2007) The symbol detachment problem. *Cogn. Process.* 8, 115–131
51. Cisek, P. (2021) An evolutionary perspective on embodiment. In *Handbook of Embodied Psychology: Thinking, Feeling, and Acting* (Robinson, M.D. and Thomas, L.E., eds), pp. 547–572, Springer
52. Cisek, P. (1999) Beyond the computer metaphor: behaviour as interaction. *J. Conscious. Stud.* 6, 11–12
53. Sugita, Y. and Tani, J. (2005) Learning semantic combinatoriality from the interaction between linguistic and behavioral processes. *Adapt. Behav.* 13, 33–52
54. Steels, L. (1995) A self-organizing spatial vocabulary. *Artif. Life* 2, 319–332
55. Bender, E.M. *et al.* (2021) On the dangers of stochastic parrots: can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 610–623, Association for Computing Machinery
56. Yamins, D.L.K. and DiCarlo, J.J. (2016) Using goal-driven deep learning models to understand sensory cortex. *Nat. Neurosci.* 19, 356–365
57. Bowers, J.S. *et al.* (2022) Deep problems with neural network models of human vision. *Behav. Brain Sci.* Published online December 1, 2022. https://doi.org/10.1017/S0140525X22002813
58. Zador, A. *et al.* (2023) Catalyzing next-generation artificial intelligence through NeuroAI. *Nat. Commun.* 14, 1597
59. Aru, J. *et al.* (2020) Cellular mechanisms of conscious processing. *Trends Cogn. Sci.* 24, 814–825
60. Block, N. (2005) Two neural correlates of consciousness. *Trends Cogn. Sci.* 9, 46–52
61. Rao, R.P. and Ballard, D.H. (1999) Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nat. Neurosci.* 2, 79–87
62. Laureys, S. (2005) The neural correlate of (un)awareness: lessons from the vegetative state. *Trends Cogn. Sci.* 9, 556–559
63. Pennartz, C.M.A. (2022) What is neurorepresentationalism? From neural activity and predictive processing to multi-level representations and consciousness. *Behav. Brain Res.* 432, 113969
64. Buzsáki, G. (2019) *The Brain from Inside Out*, Oxford University Press
65. Kontra, C. *et al.* (2015) Physical experience enhances science learning. *Psychol. Sci.* 26, 737–749
66. Sloan, A.T. *et al.* (2023) Meaning from movement and stillness: signatures of coordination dynamics reveal infant agency. *Proc. Natl. Acad. Sci.* 120, e2306732120
67. Theves, S. *et al.* (2020) The hippocampus maps concept space, not feature space. *J. Neurosci.* 40, 7318–7325
68. McNaughton, B.L. *et al.* (2006) Path integration and the neural basis of the 'cognitive map'. *Nat. Rev. Neurosci.* 7, 663–678
69. Cisek, P. (2007) Cortical mechanisms of action selection: the affordance competition hypothesis. *Phil. Trans. R. Soc. B.* 362, 1585–1599
70. Fadiga, L. *et al.* (2000) Visuomotor neurons: ambiguity of the discharge or 'motor' perception? *Int. J. Psychophysiol.* 35, 165–177
71. Graziano, M.S. (2016) Ethological action maps: a paradigm shift for the motor cortex. *Trends Cogn. Sci.* 20, 121–132
72. Johnson, M. (1987) *The Body in the Mind: The Bodily Basis of Meaning, Imagination and Reason*, University of Chicago Press
73. Pezzulo, G. *et al.* (2013) Computational grounded cognition: a new alliance between grounded cognition and computational modeling. *Front. Psychol.* 3, 612
74. Buzsáki, G. and Moser, E.I. (2013) Memory, navigation and theta rhythm in the hippocampal-entorhinal system. *Nat. Neurosci.* 16, 130–138
75. Buzsáki, G. *et al.* (2015) Emergence of cognition from action. *Cold Spring Harb. Symp. Quant. Biol.* 79, 41–50
76. Safron, A. (2021) The radically embodied conscious cybernetic Bayesian brain: from free energy to free will and back again. *Entropy* 23, 783
77. Pearl, J. and Mackenzie, D. (2018) *The Book of Why: The New Science of Cause and Effect*, Basic Books
78. Ortega, P.A. *et al.* (2021) Shaking the foundations: delusions in sequence models for interaction and control. *ArXiv* Published online October 20, 2021. https://doi.org/10.48550/arXiv.2110.10819
79. Merrill, W. (2019) Sequential neural networks as automata. *ArXiv* Published online June 4, 2019. https://doi.org/10.48550/arXiv.1906.01615
80. Lindley, D.V. (1956) On a measure of the information provided by an experiment. *Ann. Math. Stat.* 27, 986–1005
81. Pezzulo, G. *et al.* (2015) Active inference, homeostatic regulation and adaptive behavioural control. *Prog. Neurobiol.* 136, 17–35
82. Clark, A. (2019) Consciousness as generative entanglement. *J. Philos.* 116, 645–662
83. Seth, A.K. (2013) Interoceptive inference, emotion, and the embodied self. *Trends Cogn. Sci.* 17, 565–573
84. Seth, A.K. and Friston, K.J. (2016) Active interoceptive inference and the emotional brain. *Phil. Trans. R. Soc. B* 371, 20160007
85. Barrett, L.F. and Simmons, W.K. (2015) Interoceptive predictions in the brain. *Nat. Rev. Neurosci.* 16, 419–429
86. Kleckner, I.R. *et al.* (2017) Evidence for a large-scale brain system supporting allostasis and interoception in humans. *Nat. Hum. Behav.* 1, 00069
87. Pezzulo, G. (2014) Why do you fear the bogeyman? An embodied predictive coding model of perceptual inference. *Cogn. Affect. Behav. Neurosci.* 14, 902–911
88. Ouyang, L. *et al.* (2022) Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems* (Vol. 35), pp. 27730–27744, NeurIPS

89. Tschantz, A. *et al.* (2020) Learning action-oriented models through active inference. *PLoS Comput. Biol.* 16, e1007805
90. Mannella, F. *et al.* (2021) Active inference through whiskers. *Neural Netw.* 144, 428–437
91. Pezzulo, G. *et al.* (2017) Model-based approaches to active perception and control. *Entropy* 19, 266
92. Kato, S. *et al.* (2015) Global brain dynamics embed the motor command sequence of *Caenorhabditis elegans*. *Cell* 163, 656–669
93. Pezzulo, G. *et al.* (2018) Hierarchical active inference: a theory of motivated control. *Trends Cogn. Sci.* 22, 294–306
94. Parr, T. *et al.* (2023) Cognitive effort and active inference. *Neuropsychologia* 184, 108562
95. Pezzulo, G. and Castelfranchi, C. (2009) Thinking as the control of imagination: a conceptual framework for goal-directed systems. *Psychol. Res.* 73, 559–577
96. Pezzulo, G. *et al.* (2022) The evolution of brain architectures for predictive coding and active inference. *Philos. Trans. R. Soc. B* 377, 20200531
97. Park, J.S. *et al.* (2023) Generative agents: interactive simulacra of human behavior. *ArXiv* Published online April 7, 2023. https://doi.org/10.48550/arXiv.2304.03442
98. Wong, L. *et al.* (2023) From word Models to world models: translating from natural language to the probabilistic language of thought. *ArXiv* Published online June 22, 2023. https://doi.org/10.48550/arXiv.2306.12672
99. Chalmers, D. (1996) *The Conscious Mind*, Oxford University Press
100. Clark, A. *et al.* (2019) Bayesing qualia: consciousness as inference, not raw datum. *J. Conscious. Stud.* 26, 19–33
101. Safron, A. (2022) Integrated world modeling theory expanded: implications for the future of consciousness. *Front. Comput. Neurosci.* 16, 642397
102. Dayan, P. *et al.* (1995) The Helmholtz machine. *Neural Comput.* 7, 889–904
103. Parr, T. and Friston, K.J. (2018) The anatomy of inference: generative models and brain structure. *Front. Comput. Neurosci.* 12, 90
104. Schmidhuber, J. (2006) Developmental robotics, optimal artificial curiosity, creativity, music, and the fine arts. *Connect. Sci.* 18, 173–187
105. Taniguchi, T. *et al.* (2023) World models and predictive coding for cognitive and developmental robotics: frontiers and challenges. *Adv. Robot.* 37, 780–786
106. Attias, H. (2003) Planning by probabilistic inference. In *Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics* (Bishop, C.M. and Frey, B.J., eds), pp. 9–16, Society for Artificial Intelligence and Statistics
107. Botvinick, M. and Toussaint, M. (2012) Planning as inference. *Trends Cogn. Sci.* 16, 485–488
108. Lanillos, P. *et al.* (2021) Active inference in robotics and artificial agents: survey and challenges. *ArXiv* Published online December 3, 2021. https://doi.org/10.48550/arXiv.2112.01871
109. Clark, A. (2001) *Mindware: An Introduction to the Philosophy of Cognitive Science*, Oxford University Press
110. Dziri, N. *et al.* (2023) Faith and fate: limits of transformers on compositionality. *ArXiv* Published online May 29, 2023. https://doi.org/10.48550/arXiv.2305.18654
111. Jardri, R. and Denève, S. (2013) Circular inferences in schizophrenia. *Brain* 136, 3227–3241
112. Sejnowski, T.J. (2023) Large language models and the reverse turing test. *Neural Comput.* 35, 309–342
113. Aru, J. *et al.* (2023) The feasibility of artificial consciousness through the lens of neuroscience. *ArXiv* Published online June 1, 2023. https://doi.org/10.48550/arXiv.2306.00915
114. Jones, C.R. *et al.* (2022) Distrubutional semantics still can't account for affordances. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 44), Cognitive Science Society.
115. Huang, S. *et al.* (2023) Language is not all you need: aligning perception with language models. *ArXiv* Published online February 27, 2023. https://doi.org/10.48550/arXiv.2302.14045
116. Majumdar, A. *et al.* (2023) Where are we in the search for an artificial visual cortex for embodied intelligence? *ArXiv* Published online March 31, 2023. https://doi.org/10.48550/arXiv.2303.18240
117. Noever, D. *et al.* (2020) The chess transformer: mastering play using generative language models. *ArXiv* Published online Month 1, 2020. https://doi.org/10.48550/arXiv.2008.04057
118. Katz, D.M. *et al.* (2023) Gpt-4 passes the bar exam. *SSRN* Published online March 15, 2023. https://doi.org/10.2139/ssrn.4389233
119. Wei, J. *et al.* (2022) Emergent abilities of large language models. *ArXiv* Published online June 15, 2022. https://doi.org/10.48550/arXiv.2206.07682
120. Webb, T. *et al.* (2023) Emergent analogical reasoning in large language models. *Nat. Hum. Behav.* 7, 1526–1541
121. Li, B.Z. *et al.* (2021) Implicit representations of meaning in neural language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing* (Vol. 1), pp. 1813–1827, Association for Computational Linguistics
122. Patel, R. and Pavlick, E. (2022) Mapping language models to grounded conceptual spaces. In *Proceedings of the International Conference on Learning Representations, virtual conference*
123. Abdou, M. *et al.* (2021) Can language models encode perceptual structure without grounding? a case study in color. *ArXiv* Published online September 13, 2021. https://doi.org/10.48550/arXiv.2109.06129
124. Kosinski, M. (2023) Theory of mind may have spontaneously emerged in large language models. *ArXiv* Published online February 4, 2023. https://doi.org/10.48550/arXiv.2302.02083
125. Shapira, N. *et al.* (2023) Clever Hans or neural theory of mind? Stress testing social reasoning in large language models. *ArXiv* Published online May 24, 2023. http://dx.doi.org
126. Jansen, P. (2020) Visually-grounded planning without vision: language models infer detailed plans from high-level instructions. In *Findings of the Association for Computational Linguistics*, pp. 4412–4417, Association for Computational Linguistics
127. Yun, T. *et al.* (2021) Does vision-and-language pretraining improve lexical grounding? *ArXiv* Published online September 21, 2021. https://doi.org/10.48550/arXiv.2109.10246
128. Merullo, J. *et al.* (2022) Linearly mapping from image to text space. *ArXiv* Published online September 30, 2022. https://doi.org/10.48550/arXiv.2209.15162
129. Friston, K. *et al.* (2015) Knowing one's place: a free-energy approach to pattern regulation. *J. R. Soc. Interface* 12, 20141383
130. Veissière, S.P. *et al.* (2020) Thinking through other minds: a variational approach to cognition and culture. *Behav. Brain Sci.* 43, e90
131. Winn, J. *et al.* (2005) Variational message passing. *J. Mach. Learn. Res.* 6, 661–694
132. Friston, K. (2010) The free-energy principle: a unified brain theory? *Nat. Rev. Neurosci.* 11, 127–138
133. Berger, J.O. (2013) *Statistical Decision theory and Bayesian Analysis*, Springer Science & Business Media
134. Mitra, P.P. (2021) Fitting elephants in modern machine learning by statistically consistent interpolation. *Nat. Mach. Intell.* 3, 378–386
135. Lake, B.M. *et al.* (2017) Building machines that learn and think like people. *Behav. Brain Sci.* 40, e253
136. Battaglia, P.W. *et al.* (2013) Simulation as an engine of physical scene understanding. *PNAS* 110, 18327–18332
137. Friston, K. *et al.* (2017) Active inference: a process theory. *Neural Comput.* 29, 1–49
138. Friston, K.J. *et al.* (2020) Generative models, linguistic communication and active inference. *Neurosci. Biobehav. Rev.* 118, 42–64
139. Parr, T. and Pezzulo, G. (2021) Understanding, explanation, and active inference. *Front. Syst. Neurosci.* 15, 772641
140. Tschantz, A. *et al.* (2020) Scaling active inference. In *Proceedings of the 2020 International Joint Conference on Neural Networks*
141. Maisto, D. *et al.* (2021) Active inference tree search in large POMDPs. *ArXiv* Published online March 25, 2021. https://doi.org/10.48550/arXiv.2103.13860
142. Singer, W. (2021) Recurrent dynamics in the cerebral cortex: integration of sensory evidence with stored knowledge. *Proc. Natl. Acad. Sci.* 118, e2101043118
143. Safron, A. *et al.* (2022) Generalized simultaneous localization and mapping (G-SLAM) as unification framework for natural and artificial intelligences: towards reverse engineering the hippocampal/entorhinal system and principles of high-level cognition. *Front. Syst. Neurosci.* 16, 787659

144. Stoianov, I.P. *et al.* (2018) Model-based spatial navigation in the hippocampus-ventral striatum circuit: a computational analysis. *PLoS Comput. Biol.* 14, e1006316
145. Eslami, S.M.A. *et al.* (2018) Neural scene representation and rendering. *Science* 360, 1204–1210
146. Pezzulo, G. *et al.* (2014) Internally generated sequences in learning and executing goal-directed behavior. *Trends Cogn. Sci.* 18, 647–657
147. George, D. *et al.* (2021) Clone-structured graph representations enable flexible learning and vicarious evaluation of cognitive maps. *Nat. Commun.* 12, 2392
148. Mirza, M.B. *et al.* (2018) Human visual exploration reduces uncertainty about the sensed world. *PLoS ONE* 13, e0190429
149. Shipp, S. (2016) Neural elements for predictive coding. *Front. Psychol.* 7, 1792