Designing explainable artificial intelligence with active inference: A framework for transparent introspection and decision-making

Mahault Albarracin[1,2], Inês Hipólito[1,3,4], Safae Essafi Tremblay[1,5], Jason G. Fox[1], Gabriel René[1], Karl Friston[1,6], and Maxwell J. D. Ramstead[1,6]

[1] *VERSES Research Lab, Los Angeles, CA 90016, USA*
[2] *Département d'informatique, Université du Québec à Montréal, 201, Avenue du Président-Kennedy, Montréal, H2X 3Y7*
[3] *Berlin School of Mind & Brain, Humboldt-Universität zu Berlin, Berlin, Germany*
[4] *Department of Philosophy, Macquarie University, Sydney, New South Wales, Australia*
[5] *Département de philosophie, Université du Québec à Montréal, 455, Boulevard René-Lévesque Est, Montréal, H2L 4Y2*
[6] *Wellcome Centre for Human Neuroimaging, University College London, London WC1N 3AR, UK*

6th June 2023

Abstract

This paper investigates the prospect of developing human-interpretable, explainable artificial intelligence (AI) systems based on active inference and the free energy principle. We first provide a brief overview of active inference, and in particular, of how it applies to the modeling of decision-making, introspection, as well as the generation of overt and covert actions. We then discuss how active inference can be leveraged to design explainable AI systems, namely, by allowing us to model core features of "introspective" processes and by generating useful, human-interpretable models of the processes involved in decision-making. We propose an architecture for explainable AI systems using active inference. This architecture foregrounds the role of an explicit hierarchical generative model, the operation of which enables the AI system to track and explain the factors that contribute to its own decisions, and whose structure is designed to be interpretable and auditable by human users. We outline how this architecture can integrate diverse sources of information to make informed decisions in an auditable manner, mimicking or reproducing aspects of human-like consciousness and introspection. Finally, we discuss the implications of our findings for future research in AI, and the potential ethical considerations of developing AI systems with (the appearance of) introspective capabilities.

Contents

Acknowledgements

The authors are grateful to VERSES for supporting the open access publication of this paper. SET is supported in part by funding from the Social Sciences and Humanities Research Council of Canada (Ref: 767-2020-2276). KF is supported by funding for the Wellcome Centre for Human Neuroimaging (Ref: 205103/Z/16/Z) and a Canada-UK Artificial Intelligence Initiative (Ref: ES/T01279X/1). The authors are grateful to Brennan Klein for assistance with typesetting.

Conflict of interest statement

The authors have no conflicts of interest to declare.

1 Introduction: Explainable AI and active inference

Artificial intelligence (AI) systems continue to proliferate and, at the time of writing, have become an integral part of various intellectual and industrial domains, including healthcare, finance, and transportation (Mascarenhas et al., 2023; Raghupathi & Raghupathi, 2014). Traditional AI models, such as deep learning neural networks, have been widely recognized for their ability to achieve high performance and accuracy across various tasks (Goodfellow, Bengio & Courville, 2016; LeCun, Bengio & Hinton, 2015). However, it is well known that these models almost invariably function as "black boxes," with limited transparency and interpretability of their decision-making processes (Castelvecchi, 2016; Gunning, 2017). This lack of explainability can lead to skepticism and reluctance to adopt AI systems—and indeed, to harm, particularly in high-stakes situations, where the consequences of a wrong decision can be severe and harmful (Birhane, 2021; Birhane, Isaac et al., 2022; Doshi-Velez & Kim, 2017; Ribeiro, Singh & Guestrin, 2016). Indeed, a lack of explainability precludes applications in certain domains, such as fintech.

The problem of explainable AI (sometimes referred to as the "black box" problem) is the problem of understanding and interpreting how these models arrive at their decisions or predictions (Bauer, von Zahn & Hinz, 2021; Bélisle-Pipon et al., 2022). While researchers and users may have knowledge of the inputs provided to the model and the corresponding outputs that it produces, comprehending the internal workings and decision-making processes of AI systems can be complex and challenging. This is in no small part because their intricate architectures and numerous interconnected layers learn to make predictions by analyzing vast amounts of training data and adjusting their internal parameters, without explicit instruction from a programmer (Ali et al., 2023). The method by which these systems are trained thus, by design, limits their explainability. Moreover, the internal computations that are performed by these models—when they engage in decision-making—can be highly complex and nonlinear, making it difficult to extract meaningful explanations of their behavior, or insights into their decision-making process (Esterhuizen, Goldsmith & Linic, 2022). This problem is compounded by the fact that most machine learning implementations of AI fail to represent or quantify their uncertainty; especially, uncertainty about the parameters and weights that underwrite their accurate performance. This means that AI, in general, cannot evaluate (or report) the confidence in its decisions, choices or recommendations.

The lack of interpretability poses several challenges. Firstly, it hampers transparency and makes audits by third parties next to impossible, as the designers, users, and stakeholders of these systems may struggle to understand why a particular decision or prediction was made. This becomes problematic in critical domains such as healthcare or finance, where the ability to explain the reasoning behind a decision is essential for trust, accountability, and compliance with regulations (Mishra, 2021; von Eschenbach, 2021). Secondly, the black box nature of machine learning models can hinder the identification and mitigation of biases or discriminatory patterns. Without visibility into the underlying decision-making process, it becomes challenging to detect and address biases that may exist within the model's training data or architecture.

This opacity can lead to unfair or biased outcomes, perpetuating social inequalities or

discriminatory practices (Nascimento, Alencar & Cowan, 2023; van Giffen, Herhausen & Fahse, 2022; Veale & Binns, 2017). Additionally, the lack of interpretability of the model limits its ability to provide meaningful explanations to end-users. Individuals interacting with machine learning systems often seek explanations for the decisions made by these systems (Laato et al., 2022; Stiglic et al., 2020). For instance, in medical diagnosis, patients and healthcare professionals may want to understand why a particular diagnosis or treatment recommendation was given (Neri et al., 2023; Oberste et al., 2023); or consider automated suggestions in practical industrial settings (Le et al., 2023). Without explainability, users may be hesitant to trust the system's recommendations or may feel apprehensive (not without good reason) about relying on the outputs of such models.

Accordingly, the need for explainable AI has become increasingly important (Adadi & Berrada, 2018). "Explainable AI" refers to the development of AI systems that can provide human-understandable explanations for their decisions and actions (Guidotti et al., 2018). This level of transparency is crucial for fostering trust (Burrell, 2016), ensuring accountability (San Miguel, Naseer & Inakoshi, 2021), and facilitating inclusive collaboration between humans and AI systems (Birhane, Ruane et al., 2022; Hipólito, Winkle & Lie, 2023; Kokciyan et al., 2021). Recent efforts to regulate AI may turn explainability into a requirement for the deployment of any AI system at scale. For instance, in the United States, the National Institute of Standards and Technology (NIST) released its Artificial Intelligence Risk Management Framework (RMF) in 2023, which includes explainability and interpretability as crucial characteristics of a trustworthy AI system. The RMF is envisioned as a guide for tech companies to manage the risks of AI and could eventually be adopted as an industry standard. In a similar vein, US Senator Chuck Schumer has led a congressional effort to establish US regulations on AI, with one of the key aspects being the availability of explanations for how AI arrives at its responses (Drake et al., 2023).

In the European Union, a proposed Regulation Laying Down Harmonized Rules on Artificial Intelligence (better known as the "AI Act") is set to increase the transparency required for the use of so-called "high-risk" AI systems. For instance, groups that deploy automated emotion recognition systems may be obligated to inform those on whom the system is being deployed that they are being exposed to such a system. The AI Act is expected to be finalized and adopted in 2023, with its obligations likely to apply within three years' time. The Council of Europe is also in the process of developing a draft convention on artificial intelligence, human rights, democracy, and the rule of law, which will be the first legally binding international instrument on AI. This convention seeks to ensure that research, development, and deployment of AI systems are consistent with the values and interests of the EU, and that they remain compatible with the AI Act and the proposed AI Liability Directive, which includes a risk-based approach to AI. In addition, the US-EU Trade and Technology Council published a joint Roadmap for Trustworthy AI and Risk Management in 2022, which aims to advance collaborative approaches in international standards bodies related to AI, among other objectives (Skeath, Tonsager & Zhang, 2023). Therefore, explainability is clearly a major issue in research, development, and deployment of AI systems, and will remain so for the foreseeable future.

Explainable AI aims to bridge the gap between the complexity and lack of auditability of contemporary AI systems and the need for human interpretability and auditability (Adadi & Berrada, 2018; Brennen, 2020; Guidotti et al., 2018). It seeks to provide insights into the factors that influence AI decision-making, enabling users to understand the explicit reasoning and other factors driving the output of AI systems. Understanding the performance and potential biases of AI systems is crucial for their ethical and responsible deployment (Ratti & Graves, 2022; Ridley, 2022). This understanding, however, must extend beyond the performance of AI systems on academic benchmarks and tasks to include a deep understanding of what the models represent or learn, as well as the algorithms that they instantiate (Guest & Martin, 2023).

Transparency considerations are embedded in the design, development, and deployment of AI systems, from the societal problems that arise worth developing a solution, to the data collection stage, and still at the point where the AI system is deployed in the real world and iteratively improved (Hipólito, 2023; Hipólito, Winkle & Lie, 2023). This transparency may enable the implementation of other ethical AI dimensions like interpretability, accountability, and safety (Chaudhry, Cukurova & Luckin, 2022).

Researchers have been exploring various approaches to develop more explainable AI systems (Arrieta et al., 2020; Doshi-Velez & Kim, 2017). However, these efforts have yet to yield a principled and widely accepted path method for, or path to, explainability. One promising direction is to draw inspiration from research into human introspection and decision-making processes. Furthermore, a two-stage decision-making process, which includes a reflection stage where the network reflects on its feed-forward decision, can enhance the robustness and calibration of AI systems (Prabhushankar & AlRegib, 2022). It has been suggested that explainability in AI systems can be further enhanced through techniques such as layer-wise relevance propagation (Bach et al., 2015) and saliency maps (Zhang, Wu & Zhu, 2018), which aid in visualizing the model's reasoning process. By translating the internal models of AI systems into human-understandable explanations, we can foster trust and collaboration between AI systems and their human users (Lamberti, 2023). However, as (Guest & Martin, 2023) argue, we must also consider the metatheoretical calculus that underpins our understanding and use of these models. This involves not only considering the performance of the model on a task, but also the implications of the performance of the model for our understanding of the mind and brain.

In this paper, we investigate the potential of active inference, and the free energy principle (FEP) upon which is based (Friston et al., 2022; Ramstead, Sakthivadivel et al., 2023), to enhance explainability in AI systems, notably by capturing core aspects of introspective processes, hierarchical decision-making processes, and (cover and overt) forms of action in human beings (Hohwy, 2013; Ramstead, Albarracin, Kiefer, Klein, Fields et al., 2023; Ramstead, Albarracin, Kiefer, Klein, Williford et al., 2023). The FEP is a variational principle of information physics that can be used to model the dynamics of self-organizing systems like the brain. Active inference is an application of the FEP to model the perception-action loops of cognitive systems: it provides us with the basis of a unified theory of the structure and function of the brain (and indeed, of living and self-organizing systems more generally;

(Ramstead, Badcock & Friston, 2018; Ramstead et al., 2019). Active inference allows us to model self-organizing systems like brains as being driven by the imperative to minimize surprising encounters with the environment; where this surprise scores how far a thing or system deviates from its characteristic states (e.g., a fish out of water). By doing so, the brain continually updates and refines its world model, allowing the agent to act adaptively and in situationally appropriate ways.

The relevance of using active inference is that the models of cognitive dynamics—and in particular, introspection—that have been developed using its tools can be adapted to enable the design of human interpretable and auditable (and indeed, self-auditable) AI systems. The ethical and epistemological or epistemic gains that this enables are notable. The proposed active inference based AI system architecture would enable artificial agents to access and analyze their own internal states and decision-making processes, leading to a better understanding of their decision-making processes, and the ability to report on themselves. Proof of concept for this kind of "self report" is already at hand (Parr & Pezzulo, 2021) and, in principle, is supported in any application of active inference. At one level, committing to a generative model—implicit in any active inference scheme—dissolves the explainability problem. This is because one has direct access to the beliefs and belief-updating of the agent in question.

Indeed, this is why active inference has been so useful in neuroscience to model and explain behavioral and neuronal responses in terms of underlying belief states: e.g., (Adams, Shipp & Friston, 2013; Adams et al., 2022; Smith, Khalsa & Paulus, 2021; Smith, Taylor & Bilek, 2021; Sterzer et al., 2018). As demonstrated in (Parr & Pezzulo, 2021) it is a relatively straightforward matter to augment generative models to self-report their belief states. In this paper, we address a slightly more subtle aspect of explainability that rests upon "self-access"; namely, when an agent infers its own "states of mind"—states of mind that underwrite its sense-making and choices. Crucially, this kind of meta-inference (Fleming, 2020; Frith, 2023; Sandved-Smith et al., 2021; Yon & Frith, 2021) may rest on exactly the representations of uncertainty (a.k.a., precision) that are absent in conventional AI.

This paper is organized as follows. We first introduce essential aspects of active inference. We then discuss how active inference can be used to design explainable AI systems. In particular, we propose that active inference can be used as the basis for a novel AI architecture—based on explicit generative models—that both endows AI systems with a greater degree of explainability and audibility from the perspective of users and stakeholders, and allows AI systems to track and explain their own decision-making processes in a manner understandable to users and stakeholders. Finally, we discuss the implications of our findings for future research in auditable, human-interpretable AI, as well as the potential ethical considerations of developing AI systems with the appearance of introspective capabilities.

2 Active inference and introspection

2.1 A brief introduction to active inference

Active inference offers a comprehensive framework for naturalizing, explaining, simulating, and understanding the mechanisms that underwrite decision-making, perception, and action (Constant et al., 2019; Da Costa et al., 2021). The free energy principle (FEP) is a variational principle of information physics (Ramstead, Sakthivadivel et al., 2023). It has gained considerable attention and traction since it was first introduced in the context of computational neuroscience and biology (Friston, 2005, 2010). Active inference denotes a family of models premised on the FEP, which are used to understand and predict the behavior of self-organizing systems. The tools of active inference allow us to model self-organizing systems as driven by the imperative to minimize surprise, which quantifies the degree to which a given path or trajectory deviates from its inertial or characteristic path—or its upper bound, variational free energy, which scores the difference between its predictions and the actual sensory inputs it receives (Ramstead, Badcock & Friston, 2018).

Active inference modeling work suggests that decision-making, perception, and action involve the optimization of a world model that represents the causal structure of the system generating outcomes of observations (Ramstead, Sakthivadivel et al., 2023). In particular, active inference models the way that latent states or factors in the world cause sensory inputs, and how those factors cause each other, thereby capturing the essential causal structure of the measured or sensed world (Konaka & Naoki, 2023). Minimizing surprise or free energy on average and over time allows the brain to maintain a consistent and coherent internal model of the world—one that maximizes predictive accuracy while minimizing model complexity—which, in turn, enables agents to adapt and survive in their environments (Friston, 2010, 2013). (Strictly speaking, this is the other way around. In other words, agents who "survive" can always be read as minimizing variational free energy or maximizing their marginal likelihood (a.k.a., model evidence). This is often called self-evidencing (Hohwy, 2016).)

Active inference has instrumental value in allowing us to model, and thereby hopefully help to understand, core aspects of human consciousness (for a review, see (Friston, 2010)Friston, 2010). Of particular interest to us here, it enables us to model the processes involved in introspective self-access (see (Ramstead, Albarracin, Kiefer, Klein, Fields et al., 2023; Ramstead, Albarracin, Kiefer, Klein, Williford et al., 2023). Active inference modeling deploys the construct of generative models to make sense of the dynamics of self-organizing systems. In this context, a generative model is a joint probability density over the hidden or latent causes of observable outcomes; see (Ramstead, Sakthivadivel et al., 2023) for a discussion of how to interpret these models philosophically and (Sandved-Smith et al., 2021) for a gentle introduction to the technical implementation of these models.

We depict a simple generative model, apt for perceptual inference, in Figure 1, and a more complex generative model, apt for the selection of actions (a.k.a. policy selection) in Figure 2. These models specify the way in which observable outcomes are generated by (typically non-observable) states or factors in the world.

The main advantage of using generative models over current state of the art black box

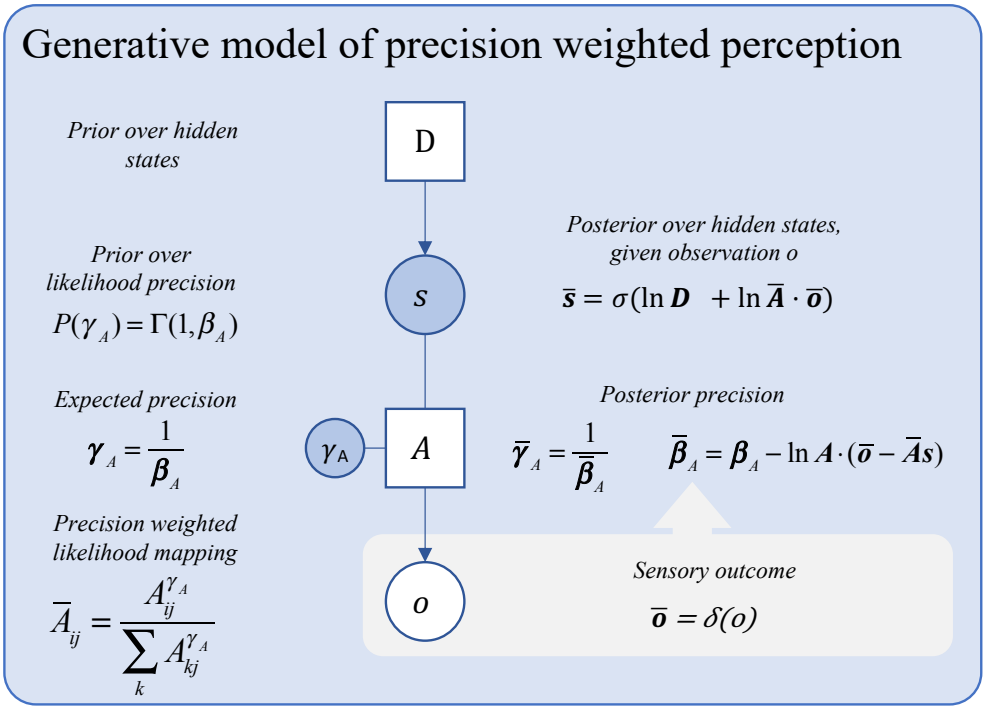

Figure 1: **A basic generative model for precision-weighted perceptual inference.** This figure depicts an elementary generative model that is capable of performing precision-weighted perceptual inference. States are depicted as circles and denoted in lowercase: observable states or outcomes are denoted o and latent states (which need to be inferred) are denoted s. Parameters are depicted as squares and denoted as uppercase. The likelihood mapping $\mathbf{A}$ relates outcomes to the states that cause them, whereas $\mathbf{D}$ harnesses our prior beliefs about states, independent of how they are sampled. The precision term γ controls the precision or weighting assigned to elements of the likelihood, and implements attention as precision-weighting. Figure from (Sandved-Smith et al., 2021).

approaches is interpretability and auditability. Indeed, the factors that figure in the generative model are explicitly labeled, such that their contributions to the operations of the model can be read directly off its structure. This lends the generative model a degree of auditability that other approaches do not have.

2.2 Active inference, introspection, and self-modeling

Active inference modeling has been deployed in the context of the scientific study of introspection, self-modeling, and self-access, which has led to the development of several leading theories of consciousness (for a review, see (Ramstead, Albarracin, Kiefer, Klein, Williford et al., 2023; Seth & Bayne, 2022)). Introspection, which is defined as the ability to access and evaluate one's own mental states, thoughts, and experiences, plays a pivotal role in self-awareness, learning, and decision-making and is a pillar of human consciousness (Limanowski

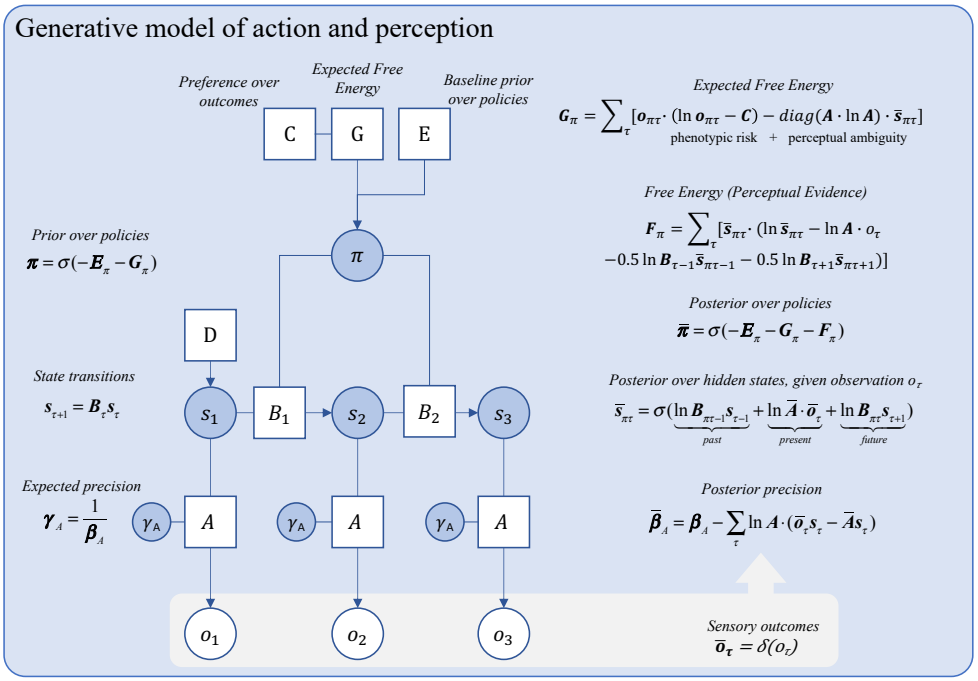

Figure 2: **A generative model for policy selection.** This figure depicts a more sophisticated generative model that is apt for planning and the selection of actions in the future. The basic model depicted in Figure 1 has now been expanded to include beliefs about the current course of action or policy (denoted $\bar{\pi}$), as well as **B**, **C**, **E**, **F** and **G** parameters. This kind of model generates a time series of states (s_1, s_2, etc.) and outcomes (o_1, o_2, etc.). The state transition (**B**) parameter encodes the transition probabilities between states over time, independently of the way they are sampled. **B**, **C**, **E**, **F** and **G** enter into the selection of beliefs about courses of action, a.k.a. policies. The **C** vector specifies preferred or expected outcomes and enters into the calculation of variational (**F**) and expected (**G**) free energies. The **E** vector specifies a prior preference for specific courses of action. Figure from (Sandved-Smith et al., 2021).

& Friston, 2018). Self-modeling and self-access can be defined as interconnected processes that contribute to the development of self-awareness and to the capacity for introspection. Self-modeling involves the creation of internal representations of oneself, while self-access refers to the ability to access and engage with these representations for self-improvement and learning (Baker, 2022; Murray, 2018). These processes, in conjunction with introspection, form a complex dynamic system that enriches our understanding of consciousness and the self—and indeed, may arguably form the causal basis of our capacity to understand ourselves and others.

Introspective self-access has been modeled using active inference by deploying a hierarchically structured generative model (Limanowski & Friston, 2020). The basic idea is that for a system to report or evaluate its own inferences, it must be able to enact some form of

self-access, where some parts of the system can take the output of other parts as their own input, for further processing. This has been discussed in computational neuroscience under the rubric of "opacity" and "transparency" (Metzinger, 2003, 2007, 2017; Sandved-Smith et al., 2021). The idea is that some cognitive processes are "transparent": like a (clean, transparent) window, they enable us to access some other thing (say, a tree outside) while not themselves being perceivable. Other cognitive processes are "opaque": they can be assessed per se, as in introspective self-awareness (i.e., aware that you are looking at a tree as opposed to seeing a tree). The idea, then, is that introspective processes make other cognitive processes accessible to the system as such, rendering them opaque.

In the context of self-access, the transparency and opacity of introspective processes has been modeled using a three-level generative model (Sandved-Smith et al., 2021). The model is depicted in Figure 3. This model provides a framework for understanding how we access and interpret our internal states and experiences. The first level of the model (in blue), which implements the selection of overt actions, can be seen as a transparent process. The second, hierarchically superordinate level (in orange), which implements attention and covert action (Metzinger, 2017; Ramstead, Albarracin, Kiefer, Klein, Fields et al., 2023), represents more opaque processes, which make processes in the first layer accessible to the system. This layer models mental actions and shifts in attention that we may not be consciously aware of, or able to report. The second level takes as its input the inferences (posterior state estimations) ongoing at the first level, as data for further inference—about the system's inferences. Attentional processes are of this sort: they are about cognitive processes and action, and they modulate the activity of the first level. The third, final level (in green) implements the awareness of where one's attention is deployed. In other words, it both recognizes and instantiates a particular attentional set via bottom-up and top-down messages between levels, respectively. On the whole, this three-level architecture models our self-access and introspective abilities in terms of the processes regulating transparency and opacity at a phenomenal level of description, or attentional selection at a psychological level.

Ramstead, Albarracin et al. (2023) recently discussed how active inference enables us to model both overt and covert action (also see (Fleming, 2020; Limanowski & Friston, 2018, 2020; Metzinger, 2017; Yon & Frith, 2021)). Overt actions—observable behaviors such as physical movements or verbal responses—are directly influenced by the brain's hierarchical organization and can be modeled using active inference (Friston, Mattout & Kilner, 2011; Friston, Parr & de Vries, 2017; Friston et al., 2017). In contrast, covert actions refer to internal mental processes, such as attention and imagination, which involve the manipulation and processing of internal representations in the absence of observable behaviors (Ainley et al., 2016; Brown et al., 2013; Edwards et al., 2012; Feldman & Friston, 2010; Hohwy, 2012; Kanai et al., 2015; Limanowski, 2017; Parr & Friston, 2019; Pezzulo, 2012; Vossel et al., 2015)—of the sort discussed as "mental action" (Limanowski, 2022; Limanowski & Friston, 2018; Metzinger, 2017; Sandved-Smith et al., 2021). These actions are essential for higher cognitive functions, which rely on the brain's capacity to explore and manipulate abstract concepts and relationships.

In Smith et al. (2019), a hierarchical architecture of this type was deployed that was

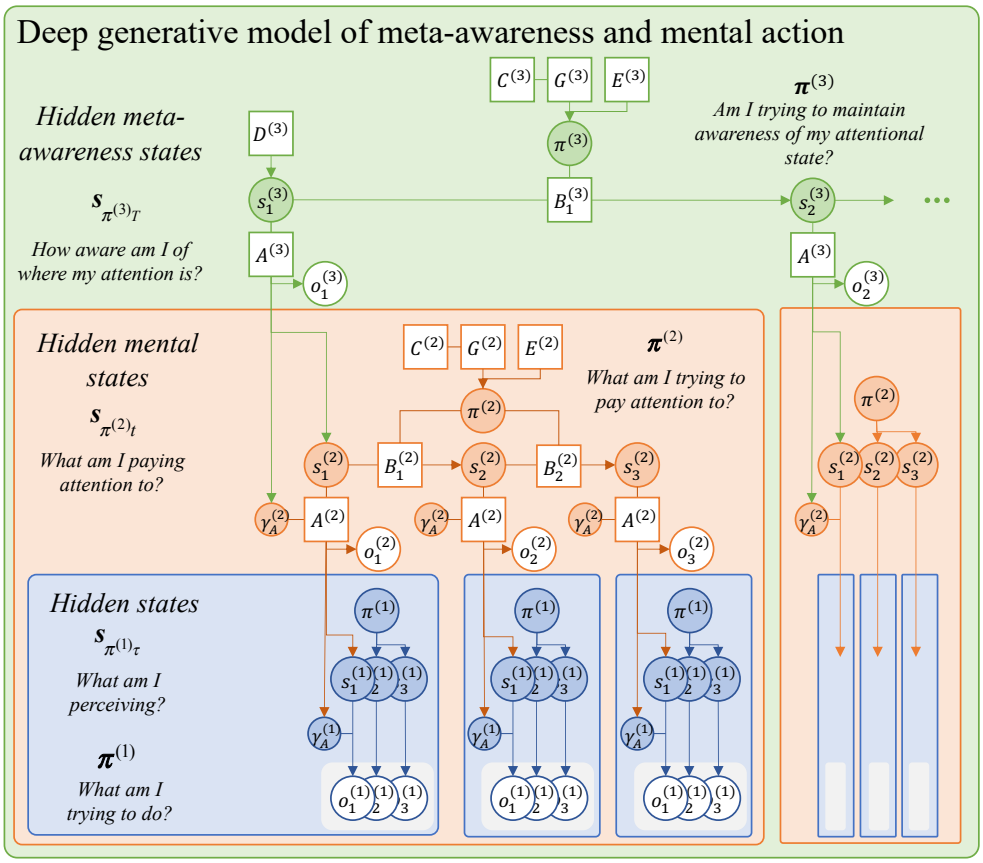

Figure 3: **A hierarchical generative model capable of self-access.** Here, the generative model depicted in Figure 2 (in blue) has been augmented with two superordinate hierarchical layers. In this architecture, posterior state estimates at one level are passed onto the next level as data for further inference. Note that this induces an architecture where the system is able to make inferences about its own inferences. Figure from (Sandved-Smith et al., 2021).

augmented with the capacity to report on its emotional states. Thus, it is possible to use active inference to design systems that can not only access their own states and perform inferences on their basis, but also to report on their introspective processes in a manner that is readily understandable by human users and stakeholders. With this formulation of how active inference enables agents to model their overt and covert action, in the following sections, we argue that we can and ought to research, design, and develop AI systems that mimic these introspective processes, ultimately leading to more human-like artificial intelligence.

3 Using active inference to design self-explaining AI

We argue that incorporating the design principles of active inference into AI systems can lead to better explainability. This is for two key reasons. The first is that, by deploying an

explicit generative model, AI systems premised on active inference are designed explicitly such that their operations can be interpreted and audited by a user or stakeholder that is fluent in the operation of such models. The second is that, by implementing an architecture inspired by active inference models of introspection, we can build systems that are able to access—and report on—the reasons for their decisions, and their state of mind when reaching these decisions.

AI systems designed using active inference can incorporate the kind of hierarchical self-access described by (Sandved-Smith et al., 2021) and by (Smith et al., 2019), to enhance their introspection during decision-making. As discussed, in the active inference tradition, introspection can be understood in the context of the (covert and overt) actions that AI systems perform. Covert actions, which are internal computations and decision-making processes that are not directly observable to users and stakeholders, can be recorded or explained to make the system more explainable. Overt actions, which are actions that an AI system takes based on its internal computations, such as making a recommendation or decision, can be explained to help users understand why the AI system acted as it did. This kind of deep inference promotes introspection, adaptability, and responses to environmental changes (Dhulipala & Hruska, 2022; Schoeffer et al., 2023).

The proposed AI architecture includes components that continuously update and maintain an internal model of its own states, beliefs, and goals. This capacity for self-access (and implicitly self-report) enables the AI system to optimize (and report on) its decision-making processes, fostering introspection (and enhanced explainability). It incorporates metacognitive processing capabilities, which involve the ability to monitor, control, and evaluate its own cognitive processes. The AI system can thereby better explain the factors that contribute to its decisions, as well as identify potential biases or errors, ultimately leading to improved decision-making and explainability.

The proposed AI architecture would include introspection and a self-report interface, which translates the AI system's internal models and decision-making processes into human-understandable (natural) language (using, e.g., large language models). In effect, the agent would be talking to itself, describing its current state of mind and beliefs. This interface bridges the gap between the AI system's internal workings and human users, promoting epistemic trust and collaboration. In this way, the system can effectively mimic human-like consciousness and transparent introspection, leading to a deeper understanding of its decision-making processes and explainability. This advancement may be essential in fostering trust and collaboration between AI systems and their human users, paving the way for more effective and responsible AI applications.

Augmenting a generative model with black box systems—like large language models—may be a useful strategy to help AI systems articulate their "understanding" of the world. Using large language models to furnish an introspective interface may be relatively straightforward, leveraging their powerful natural language processing capabilities to create explanations of belief updating. This architecture—with a hierarchical generative model at its core—may contribute to the overall performance and explainability of hybrid AI systems. Attention mechanisms also achieve this purpose by enhancing the explainability of the AI

system's decision making, emphasizing important factors in the hierarchical generative model that contribute to its decisions and actions.

These ideas are not new. Attentional mechanisms, particularly those at the word-level, have been identified as crucial components in AI architecture, specifically in the context of hierarchical generative models—and in generative AI, in the form of transformers. They function by focusing on relevant aspects during decision-making processes, thereby allowing the system to effectively process and prioritize information (Lan et al., 2020). In fact, the performance of hierarchical models, which are a type of AI architecture, can be significantly improved by integrating word-level attention mechanisms. These mechanisms are powerful because they can leverage context information more effectively, especially fine-grained information.

The AI architecture that we propose employs a soft attention mechanism, which uses a weighted combination of hierarchical generative model components to focus on relevant information. The attention weights are dynamically computed based on the input data and the AI system's internal state, allowing the system to adaptively focus on different aspects of the hierarchical generative model (Kulkarni & Abubakar, 2020). This approach is similar to the use of deep learning models for global coordinate transformations that linearize partial differential equations, where the model is trained to learn a transformation from the physical domain to a computational domain where the governing partial differential equations are simpler, or even linear (Gin et al., 2021).

The AI architecture that we describe here effectively integrates diverse information sources for decision-making, mirroring the complex information processing capabilities observed in the human brain. The hierarchical structure of the generative model facilitates the exchange of information between different levels of abstraction. This exchange allows the AI system to refine and update its internal models based on both high-level abstract knowledge and low-level detailed information.

In conclusion, the integration of introspective processes in AI systems may represent a significant step towards achieving more explainable AI. By leveraging explicit generative models, as well as attention and introspection mechanisms, we can design AI systems that are not only more efficient and robust, but also more understandable and trustworthy. This approach allows us to bridge the gap between the complex internal computations of AI systems and the human users who interact with them. Ultimately, the goal is to create AI systems that can effectively communicate the reasons that drive their decision-making processes, adapt to environmental changes, and collaborate seamlessly with human users. As we continue to advance in this field, the importance of introspection in AI will only become more apparent, paving the way for more sophisticated and ethically sound AI systems.

4 Discussion

4.1 Directions for future research

The problem of explainable AI is the problem of understanding how AI models arrive at their decisions or predictions. This problem is especially relevant to avoid biases and harm in the design, implementation, and use of AI systems. By incorporating explicit generative models and introspective processing into the proposed AI architecture, we can create a system that is or seems capable of introspection and, thereby, that displays greatly enhanced explainability and auditability. This approach to AI design paves the way for more effective AI deployment across various real-world applications, by shedding light upon the problem of explainability, thereby offering opportunities for fostering trust, fairness, and inclusivity.

The development of the AI architecture based on active inference opens several potential avenues for future research. One possible direction is to further investigate the role of attention and introspection mechanisms in both AI systems and human cognition, as well as the development of more efficient attentional models to improve the AI system's ability to focus on salient information during decision-making. The approach that we propose bridges the gap between AI and cognitive neuroscience by incorporating biologically-inspired mechanisms into the design of AI systems. As a result, the proposed architecture promotes a deeper understanding of the nature of cognition and its potential applications in artificial intelligence, thus paving the way for more human-like AI systems capable of introspection and enhanced collaboration with human users.

Future work could explore more advanced data fusion techniques, such as deep learning-based fusion or probabilistic fusion, to improve the AI system's ability to combine and process multimodal data effectively. Evaluating the effectiveness of these techniques in diverse application domains will also be a valuable avenue for research (Lahat, Adali & Jutten, 2015; Microsoft Defender Security Research Team, 2020). Furthermore, the explanation dimension of these AI systems has been a significant topic in recent years, particularly in decision-making scenarios. These systems provide more awareness of how AI works and its outcomes, building a relationship with the system and fostering trust between AI and humans (Ferreira & Monteiro, 2021).

In addition to the aforementioned avenues for future research, another promising direction lies in the realm of computational phenomenology (for a review and discussion, see (Ramstead et al., 2022). Beckmann, Köstner, & Hipólito (2023) have proposed a framework that deploys phenomenology—the rigorous descriptive study of first-person experience—for the purposes of machine learning training. This approach conceptualizes the mechanisms of artificial neural networks in terms of their capacity to capture the statistical structure of some kinds of lived experience, offering a unique perspective on deep learning, consciousness, and their relation. By grounding AI training in socioculturally situated experience, we can create systems that are more aware of sociocultural biases and capable of mitigating their impact. Ramstead et al. (2022) propose a similar methodology based on explicit generative models as they figure in the active inference tradition. This connection to first-person experience, of course, does not guarantee unbiased AI. But by moving away from traditional black

box AI systems, we shift towards human-interpretable models that enable the identification and correction of biases in the AI system. This approach aligns with our goal of creating AI systems that are not only efficient and effective, but also ethically sound and socially responsible.

The incorporation of computational phenomenology into our proposed AI architecture could further enhance its introspective capabilities and its ability to understand and navigate the complexities of human sociocultural contexts. This could lead to AI systems that are more adaptable, more trustworthy, and more capable of meaningful collaboration with human users. As we continue to explore and integrate such innovative approaches, we move closer to our goal of creating AI systems that truly mirror the richness and complexity of human cognition and consciousness.

4.2 Ethical considerations of introspective AI systems

Ethical AI starts with the development of AI systems that are ethically designed; AI systems must be designed in such a way as to be transparent, auditable, and explainable, and to minimize harm. But as these systems become increasingly integrated into our daily lives, research on the ethical implications of introspective AI systems, as well as the development of regulatory frameworks and guidelines for responsible AI use, become crucial. The development of introspective AI systems raises several ethical considerations. Even if these systems provide more human-like decision-making capabilities and enhanced explainability, it is and will remain crucial to ensure that their decisions are transparent, fair, and unbiased, and that their designers and users can be held accountable for harm that their use may cause.

To address these concerns, future research should focus on developing methods to audit and evaluate the AI system's decision-making processes, as well as identify and mitigate potential biases within the system. Additionally, the development of ethical guidelines and regulatory frameworks for the use of introspective AI systems will be essential to ensure that they are deployed responsibly and transparently. Moreover, as introspective AI systems become more prevalent, issues related to agency, privacy, and data security may arise. Ensuring that these systems protect sensitive information by abiding by data protection regulations, thereby safeguarding agency, will be of paramount importance.

In conclusion, the development of AI systems based on active inference has broad implications for both the fields of AI and consciousness studies. As future research explores the potential of this novel approach, ethical considerations and responsible use of introspective AI systems must remain at the forefront of these advancements, ultimately leading to more transparent, effective, and user-friendly AI applications.

5 Conclusion

We have argued that active inference has demonstrated significant potential in advancing the field of explainable AI. By incorporating design principles from active inference, the AI

system can better tackle complex real-world problems with improved auditability of decision-making, thereby increasing safety and user trust.

Throughout our discussions and analysis, we have highlighted the importance of active inference models as a foundation for designing more human-like AI systems, seemingly capable of introspection and finessed (epistemic) collaboration with human users. This novel approach bridges the gap between AI and cognitive neuroscience by incorporating biologically-inspired mechanisms into the design of AI systems, thus promoting a deeper understanding of the nature of consciousness and its potential applications in artificial intelligence.

As we move forward in the development of AI systems, the importance of advancing explainable AI becomes increasingly apparent. By designing AI systems that can not only make accurate and efficient decisions, but also provide understandable explanations for their decisions, we foster (epistemic) trust and collaboration between AI systems and human users. This advancement ultimately leads to more transparent, effective, and user-friendly AI applications that can be tailored to a wide range of real-world scenarios.

References

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access*, *6*, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

Adams, R. A., Shipp, S., & Friston, K. J. (2013). Predictions not commands: Active inference in the motor system. *Brain Structure and Function*, *218*(3), 611–643. https://doi.org/10.1007/s00429-012-0475-5

Adams, R. A., Vincent, P., Benrimoh, D., Friston, K. J., & Parr, T. (2022). Everything is connected: Inference and attractors in delusions. *Schizophrenia Research*, *245*, 5–22. https://doi.org/10.1016/j.schres.2021.07.032

Ainley, V., Apps, M. A. J., Fotopoulou, A., & Tsakiris, M. (2016). 'Bodily precision': A predictive coding account of individual differences in interoceptive accuracy. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *371*(1708), 20160003. https://doi.org/10.1098/rstb.2016.0003

Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Moral, J. M., Confalonieri, R., Guidotti, R., Del Ser, J., Díaz-Rodríguez, N., & Herrera, F. (2023). Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. *Information Fusion*, 101805. https://doi.org/10.1016/j.inffus.2023.101805

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, *58*, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., & Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance

propagation. *PloS One*, *10*(7), e0130140. https://doi.org/10.1371/journal.pone.0130140

Baker, J. R. (2022). Going beyond brick and mortar self-access centers: Establishing a satellite activity self-access program. *Studies in Self-Access Learning Journal*, *13*(1), 129–141. https://doi.org/10.37237/130107

Bauer, K., von Zahn, M., & Hinz, O. (2021). Expl(AI)ned: The impact of explainable artificial intelligence on cognitive processes. *Information Systems Research*. https://doi.org/10.1287/isre.2023.1199

Beckmann, P., Köstner, G., & Hipólito, I. (2023). Rejecting cognitivism: Computational phenomenology for deep learning. *arXiv*. https://doi.org/10.48550/arXiv.2302.09071

Bélisle-Pipon, J.-C., Monteferrante, E., Roy, M.-C., & Couture, V. (2022). Artificial intelligence ethics has a black box problem. *AI & SOCIETY*, 1–16. https://doi.org/10.1007/s00146-021-01380-0

Birhane, A. (2021). The impossibility of automating ambiguity. *Artificial Life*, *27*(1), 44–61. https://doi.org/10.1162/artl_a_00336

Birhane, A., Isaac, W. S., Prabhakaran, V., Díaz, M., Elish, M. C., Gabriel, I., & Mohamed, S. (2022). Frameworks and Challenges to Participatory AI. *Proceeding of the Second Conference on Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO '22)*. https://doi.org/10.48550/arXiv.2209.07572

Birhane, A., Ruane, E., Laurent, T., S. Brown, M., Flowers, J., Ventresque, A., & L. Dancy, C. (2022). The forgotten margins of AI ethics. *2022 ACM Conference on Fairness, Accountability, and Transparency*, 948–958. https://doi.org/10.1145/3531146.3533157

Brennen, A. (2020). What do people really want when they say they want "Explainable AI?" we asked 60 stakeholders. *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–7. https://doi.org/10.1145/3334480.3383047

Brown, H., Adams, R. A., Parees, I., Edwards, M., & Friston, K. J. (2013). Active inference, sensory attenuation and illusions. *Cognitive Processing*, *14*, 411–427. https://doi.org/10.1007/s10339-013-0571-3

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society*, *3*(1), 2053951715622512. https://doi.org/10.1177/2053951715622512

Castelvecchi, D. (2016). Can we open the black box of AI? *Nature News*, *538*(7623), 20. https://doi.org/10.1038/538020a

Chaudhry, M. A., Cukurova, M., & Luckin, R. (2022). A transparency index framework for AI in education. *Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners' and Doctoral Consortium: 23rd International Conference, AIED 2022, Durham, UK, July 27–31, 2022, Proceedings, Part II*, 195–198. https://doi.org/10.1007/978-3-031-11647-6_33

Constant, A., Ramstead, M. J. D., Veissière, S. P. L., & Friston, K. J. (2019). Regimes of expectations: An active inference model of social conformity and human decision making. *Frontiers in Psychology*, *10*, 679. https://doi.org/10.3389/fpsyg.2019.00679

Da Costa, L., Friston, K. J., Heins, C., & Pavliotis, G. A. (2021). Bayesian mechanics for stationary processes. *Proceedings of the Royal Society A*, *477*(2256). https://doi.org/10.1098/rspa.2021.0518

Dhulipala, S. L. N., & Hruska, R. C. (2022). Efficient interdependent systems recovery modeling with DeepONets. *arXiv*, 1–6. https://doi.org/10.48550/arXiv.2206.10829

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv*. https://doi.org/10.48550/arXiv.1702.08608

Drake, M., Ong, J., Hansen, M., & Peets, L. (2023). EU AI Policy and Regulation: What to look out for in 2023. *Inside Privacy*. https://www.insideprivacy.com/artificial-intelligence/eu-ai-policy-and-regulation-what-to-look-out-for-in-2023/

Edwards, M. J., Adams, R. A., Brown, H., Parees, I., & Friston, K. J. (2012). A Bayesian account of 'hysteria'. *Brain*, *135*(11), 3495–3512. https://doi.org/10.1093/brain/aws129

Esterhuizen, J. A., Goldsmith, B. R., & Linic, S. (2022). Interpretable machine learning for knowledge generation in heterogeneous catalysis. *Nature Catalysis*, *5*(3), 175–184. https://doi.org/10.1038/s41929-022-00744-z

Feldman, H., & Friston, K. J. (2010). Attention, uncertainty, and free-energy. *Frontiers in Human Neuroscience*, *4*. https://doi.org/10.3389/fnhum.2010.00215

Ferreira, J. J., & Monteiro, M. (2021). The human-AI relationship in decision-making: AI explanation to support people on justifying their decisions. *arXiv*. https://doi.org/10.48550/arXiv.2102.05460

Fleming, S. M. (2020). Awareness as inference in a higher-order state space. *Neuroscience of Consciousness*, *2020*(1), niz020. https://doi.org/10.1093/nc/niz020

Friston, K. J. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *360*(1456), 815–836. https://doi.org/10.1098/rstb.2005.1622

Friston, K. J. (2010). Is the free-energy principle neurocentric? *Nature Reviews Neuroscience*, *11*(8), 605–605. https://doi.org/10.1038/nrn2787-c2

Friston, K. J. (2013). Life as we know it. *Journal of the Royal Society Interface*, *10*(86), 20130475. https://doi.org/10.1098/rsif.2013.0475

Friston, K. J., Mattout, J., & Kilner, J. (2011). Action understanding and active inference. *Biological Cybernetics*, *104*, 137–160. https://doi.org/10.1007/s00422-011-0424-z

Friston, K. J., Parr, T., & de Vries, B. (2017). The graphical brain: Belief propagation and active inference. *Network Neuroscience*, *1*(4), 381–414. https://doi.org/10.1162/NETN_a_00018

Friston, K. J., Ramstead, M. J. D., Kiefer, A. B., Tschantz, A., Buckley, C. L., Albarracin, M., Pitliya, R. J., Heins, C., Klein, B., Millidge, B., Sakthivadivel, D. A. R., St Clere Smithe, T., Koudahl, M., Essafi Tremblay, S., Petersen, C., Fung, K., Fox, J. G., Swanson, S., Mapes, D., & René, G. (2022). Designing ecosystems of intelligence from first principles. *arXiv*. https://doi.org/10.48550/arXiv.2212.01354

Friston, K. J., Rosch, R., Parr, T., Price, C., & Bowman, H. (2017). Deep temporal models and active inference. *Neuroscience & Biobehavioral Reviews*, *77*, 388–402. https://doi.org/10.1016/j.neubiorev.2017.04.009

Frith, C. D. (2023). Consciousness, (meta) cognition, and culture. *Quarterly Journal of Experimental Psychology*, 17470218231164502. https://doi.org/10.1177/17470218231164502

Gin, C., Lusch, B., Brunton, S. L., & Kutz, J. N. (2021). Deep learning models for global coordinate transformations that linearise PDEs. *European Journal of Applied Mathematics*, *32*(3), 515–539. https://doi.org/10.1017/S0956792520000327

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT press.

Guest, O., & Martin, A. E. (2023). On logical inference over brains, behaviour, and artificial neural networks. *Computational Brain & Behavior*, 1–15. https://doi.org/10.1007/s42113-022-00166-x

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM Computing Surveys*, *51*(5), 1–42. https://doi.org/10.1145/3236009

Gunning, D. (2017). Explainable Artificial Intelligence (XAI). *Defense Science Research Projects Agency*, *2*(2), 1. https://doi.org/10.1609/aimag.v40i2.2850

Hipólito, I. (2023). The human roots of artificial intelligence. https://doi.org/10.31234/osf.io/cseqt

Hipólito, I., Winkle, K., & Lie, M. (2023). Enactive artificial intelligence: Subverting gender norms in robot-human interaction. *Frontiers in Neurorobotics*, *17*, 77. https://doi.org/10.48550/arXiv.2301.08741

Hohwy, J. (2012). Attention and conscious perception in the hypothesis testing brain. *Frontiers in Psychology*, *3*, 96. https://doi.org/10.3389/fpsyg.2012.00096

Hohwy, J. (2013). *The Predictive Mind*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199682737.001.0001

Hohwy, J. (2016). The self-evidencing brain. *Nous*, *50*(2), 259–285. https://doi.org/10.1111/nous.12062

Kanai, R., Komura, Y., Shipp, S., & Friston, K. J. (2015). Cerebral hierarchies: Predictive processing, precision and the pulvinar. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *370*(1668), 20140169. https://doi.org/10.1098/rstb.2014.0169

Kokciyan, N., Srivastava, B., Huhns, M. N., & Singh, M. P. (2021). Sociotechnical perspectives on AI ethics and accountability. *IEEE Internet Computing*, *25*(6), 5–6. https://doi.org/10.1109/MIC.2021.3117611

Konaka, Y., & Naoki, H. (2023). Decoding reward–curiosity conflict in decision-making from irrational behaviors. *Nature Computational Science*, *3*(5), 418–432. https://doi.org/10.1038/s43588-023-00439-w

Kulkarni, M., & Abubakar, A. (2020). Soft attention convolutional neural networks for rare event detection in sequences. https://doi.org/10.48550/arXiv.2011.02338

Laato, S., Tiainen, M., Najmul Islam, A., & Mäntymäki, M. (2022). How to explain AI systems to end users: A systematic literature review and research agenda. *Internet Research*, *32*(7), 1–31. https://doi.org/10.1108/INTR-08-2021-0600

Lahat, D., Adali, T., & Jutten, C. (2015). Multimodal data fusion: An overview of methods, challenges, and prospects. *Proceedings of the IEEE*, *103*(9), 1449–1477. https://doi.org/10.1109/JPROC.2015.2460697

Lamberti, W. F. (2023). An overview of explainable and interpretable AI. *AI Assurance*, 55–123. https://doi.org/10.1016/B978-0-32-391919-7.00015-9

Lan, T., Mao, X.-L., Wei, W., & Huang, H. (2020). Which kind is better in open-domain multi-turn dialog, hierarchical or non-hierarchical models? an empirical study. *arXiv*. https://doi.org/10.48550/arXiv.2008.02964

Le, T.-T.-H., Prihatno, A. T., Oktian, Y. E., Kang, H., & Kim, H. (2023). Exploring local explanation of practical industrial AI applications: A systematic literature review. *Applied Sciences*, *13*(9), 5809. https://doi.org/10.3390/app13095809

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539

Limanowski, J. (2017). (Dis-)attending to the body — Action and self-experience in the active inference framework. In T. Metzinger & W. Wiese (Eds.), *Philosophy and Predictive Processing*. Frankfurt am Main: MIND Group. https://doi.org/10.15502/9783958573192

Limanowski, J. (2022). Precision control for a flexible body representation. *Neuroscience and Biobehavioral Reviews*, *134*, 104401. https://doi.org/10.1016/j.neubiorev.2021.10.023

Limanowski, J., & Friston, K. J. (2018). 'Seeing the dark': Grounding phenomenal transparency and opacity in precision estimation for active inference. *Frontiers in Psychology*, *9*, 643. https://doi.org/10.3389/fpsyg.2018.00643

Limanowski, J., & Friston, K. J. (2020). Attenuating oneself: An active inference perspective on "selfless" experiences. *Philosophy and the Mind Sciences*, *1*(1), 1–16. https://doi.org/10.33735/phimisci.2020.I.35

Mascarenhas, M., Afonso, J., Ribeiro, T., Andrade, P., Cardoso, H., & Macedo, G. (2023). The promise of artificial intelligence in digestive healthcare and the bioethics challenges it presents. *Medicina*, *59*(4), 790. https://doi.org/10.3390/medicina59040790

Metzinger, T. (2003). Phenomenal transparency and cognitive self-reference. *Phenomenology and the Cognitive Sciences*, *2*, 353–393. https://doi.org/10.1023/b:phen.0000007366.42918.eb

Metzinger, T. (2007). Empirical perspectives from the self-model theory of subjectivity: A brief summary with examples. *Progress in Brain Research*, *168*, 215–278. https://doi.org/10.1016/S0079-6123(07)68018-2

Metzinger, T. (2017). The problem of mental action. In T. Metzinger & W. Wiese (Eds.), *Philosophy and Predictive Processing*. Frankfurt am Main: MIND Group. https://doi.org/10.15502/9783958573208

Microsoft Defender Security Research Team. (2020). Seeing the big picture: Deep learning-based fusion of behavior signals for threat detection. https://tinyurl.com/3kpzvk9d

Mishra, A. (2021). Transparent AI: Reliabilist and proud. *Journal of Medical Ethics*, *47*(5), 341–342. https://doi.org/10.1136/medethics-2021-107352

Murray, G. (2018). Self-access environments as self-enriching complex dynamic ecosocial systems. *Studies in Self-Access Learning Journal*, *9*(2). https://doi.org/10.37237/090204

Nascimento, N., Alencar, P., & Cowan, D. (2023). Comparing software developers with chatgpt: An empirical investigation. *arXiv*. https://doi.org/10.48550/arXiv.2305.11837

Neri, E., Aghakhanyan, G., Zerunian, M., Gandolfo, N., Grassi, R., Miele, V., Giovagnoni, A., Laghi, A., & expert group on Artificial Intelligence, S. (2023). Explainable AI in radiology: A white paper of the Italian Society of Medical and Interventional Radiology. *La Radiologia Medica*, 1–10. https://doi.org/10.1007/s11547-023-01634-5

Oberste, L., Rüffer, F., Aydingül, O., Rink, J., & Heinzl, A. (2023). Designing user-centric explanations for medical imaging with informed machine learning. *Design Science Research for a New Society: Society 5.0: 18th International Conference on Design Science Research in Information Systems and Technology, DESRIST 2023, Pretoria, South Africa, May 31–June 2, 2023, Proceedings*, 470–484. https://doi.org/10.1007/978-3-031-32808-4_29

Parr, T., & Friston, K. J. (2019). Attention or salience? *Current Opinion in Psychology*, *29*, 1–5. https://doi.org/10.1016/j.copsyc.2018.10.006

Parr, T., & Pezzulo, G. (2021). Understanding, explanation, and active inference. *Frontiers in Systems Neuroscience*, *15*, 772641. https://doi.org/10.3389/fnsys.2021.772641

Pezzulo, G. (2012). An active inference view of cognitive control. *Frontiers in Psychology*, *3*, 478. https://doi.org/10.3389/fpsyg.2012.00478

Prabhushankar, M., & AlRegib, G. (2022). Introspective learning: A two-stage approach for inference in neural networks. *arXiv*. https://openreview.net/forum?id=in1ynkrXyMH

Raghupathi, W., & Raghupathi, V. (2014). Big data analytics in healthcare: Promise and potential. *Health Information Science and Systems*, *2*, 1–10. https://doi.org/10.1186/2047-2501-2-3

Ramstead, M. J. D., Albarracin, M., Kiefer, A., Klein, B., Fields, C., Friston, K. J., & Safron, A. (2023). The inner screen model of consciousness: Applying the free energy principle directly to the study of conscious experience. *PsyArXiv*. https://doi.org/10.31234/osf.io/6afs3

Ramstead, M. J. D., Albarracin, M., Kiefer, A., Klein, B., Williford, K., Safron, A., Fields, C., Solms, M., & Friston, K. J. (2023). Steps towards a minimal unifying model of consciousness: An integration of models of consciousness based on the free energy principle.

Ramstead, M. J. D., Badcock, P. B., & Friston, K. J. (2018). Answering Schrödinger's question: A free-energy formulation. *Physics of Life Reviews*, *24*, 1–16. https://doi.org/10.1016/j.plrev.2017.09.001

Ramstead, M. J. D., Constant, A., Badcock, P. B., & Friston, K. J. (2019). Variational ecology and the physics of sentient systems. *Physics of Life Reviews*, *31*, 188–205. https://doi.org/10.1016/j.plrev.2018.12.002

Ramstead, M. J. D., Sakthivadivel, D. A. R., Heins, C., Koudahl, M., Millidge, B., Da Costa, L., Klein, B., & Friston, K. J. (2023). On Bayesian mechanics: A physics of and by beliefs. *Interface Focus*, *13*, 20220029. https://doi.org/10.1098/rsfs.2022.0029

Ramstead, M. J. D., Seth, A. K., Hesp, C., Sandved-Smith, L., Mago, J., Lifshitz, M., Pagnoni, G., Smith, R., Dumas, G., Lutz, A., Friston, K. J., & Constant, A. (2022). From generative models to generative passages: A computational approach to (neuro) phenomenology. *Review of Philosophy and Psychology*, *13*(4). https://doi.org/10.1007/s13164-021-00604-y

Ratti, E., & Graves, M. (2022). Explainable machine learning practices: Opening another black box for reliable medical AI. *AI and Ethics*, *2*(4), 801–814. https://doi.org/10.1007/s43681-022-00141-z

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144. https://doi.org/10.1145/2939672.2939778

Ridley, M. (2022). Explainable Artificial Intelligence (XAI). *Information Technology and Libraries*, *41*(2). https://doi.org/10.6017/ital.v41i2.14683

San Miguel, B., Naseer, A., & Inakoshi, H. (2021). Putting accountability of AI systems into practice. *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, 5276–5278. https://doi.org/10.24963/ijcai.2020/768

Sandved-Smith, L., Hesp, C., Mattout, J., Friston, K. J., Lutz, A., & Ramstead, M. J. D. (2021). Towards a computational phenomenology of mental action: Modelling meta-awareness and attentional control with deep parametric active inference. *Neuroscience of Consciousness*, *2021*(1). https://doi.org/10.1093/nc/niab018

Schoeffer, J., Jakubik, J., Voessing, M., Kuehl, N., & Satzger, G. (2023). On the interdependence of reliance behavior and accuracy in AI-assisted decision-making. *arXiv*. https://doi.org/10.48550/arXiv.2304.08804

Seth, A. K., & Bayne, T. (2022). Theories of consciousness. *Nature Reviews Neuroscience*, *23*(7), 439–452. https://doi.org/10.1038/s41583-022-00587-4

Skeath, C., Tonsager, L., & Zhang, J. (2023). FTC Announces COPPA Settlement Against Ed Tech Provider Including Strict Data Minimization and Data Retention Requirements. *Inside Privacy*. https://www.insideprivacy.com/childrens-privacy/ftc-announces-coppa-settlement-against-ed-tech-provider-including-strict-data-minimization-and-data-retention-requirements

Smith, R., Khalsa, S. S., & Paulus, M. P. (2021). An active inference approach to dissecting reasons for nonadherence to antidepressants. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, *6*(9), 919–934. https://doi.org/10.1016/j.bpsc.2019.11.012

Smith, R., Lane, R. D., Parr, T., & Friston, K. J. (2019). Neurocomputational mechanisms underlying emotional awareness: Insights afforded by deep active inference and their potential clinical relevance. *Neuroscience & Biobehavioral Reviews*, *107*, 473–491. https://doi.org/10.1016/j.neubiorev.2019.09.002

Smith, R., Taylor, S., & Bilek, E. (2021). Computational mechanisms of addiction: Recent evidence and its relevance to addiction medicine. *Current Addiction Reports*, 1–11. https://doi.org/10.1007/s40429-021-00399-z

Sterzer, P., Adams, R. A., Fletcher, P., Frith, C., Lawrie, S. M., Muckli, L., Petrovic, P., Uhlhaas, P., Voss, M., & Corlett, P. R. (2018). The predictive coding account of psychosis. *Biological Psychiatry*, *84*(9), 634–643. https://doi.org/10.1016/j.biopsych.2018.05.015

Stiglic, G., Kocbek, P., Fijacko, N., Zitnik, M., Verbert, K., & Cilar, L. (2020). Interpretability of machine learning-based prediction models in healthcare. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *10*(5), e1379. https://doi.org/10.1002/widm.1379

van Giffen, B., Herhausen, D., & Fahse, T. (2022). Overcoming the pitfalls and perils of algorithms: A classification of machine learning biases and mitigation methods. *Journal of Business Research*, *144*, 93–106. https://doi.org/10.1016/j.jbusres.2022.01.076

Veale, M., & Binns, R. (2017). Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. *Big Data & Society*, *4*(2). https://doi.org/10.1177/2053951717743530

von Eschenbach, W. J. (2021). Transparency and the black box problem: Why we do not trust AI. *Philosophy & Technology*, *34*(4), 1607–1622. https://doi.org/10.1007/s13347-021-00477-0

Vossel, S., Mathys, C., Stephan, K. E., & Friston, K. J. (2015). Cortical coupling reflects bayesian belief updating in the deployment of spatial attention. *Journal of Neuroscience*, *35*(33), 11532–11542. https://doi.org/10.1523/JNEUROSCI.1382-15.2015

Yon, D., & Frith, C. D. (2021). Precision and the Bayesian brain. *Current Biology*, *31*(17), R1026–R1032. https://doi.org/10.1016/j.cub.2021.07.044

Zhang, Q., Wu, Y. N., & Zhu, S.-C. (2018). Interpretable convolutional neural networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 8827–8836. https://doi.org/10.1109/CVPR.2018.00920