Active Inference for AI

Maria Raffa[1]

[1] IULM University, via Carlo Bo 1, Milan, Italy

Abstract
The aim of this short paper is to present the connection between the cognitive framework of predictive processing and active inference and existing implementation in AI tools, in order to investigate whether these models of cognition can be used to design explainable and sustainable AI architectures.

Keywords
AI, predictive processing, active inference, Bayesian networks

1. Introduction

The aim of this short paper is to trace the connection between the cognitive framework of predictive processing (PP) and the existing implementation in AI tools. PP is related to some formal models of mind, i.e., Karl Friston's free energy principle (FEP) and active inference (AIF), which share their mathematical foundations with generative AI. Therefore, it is worth exploring this relationship to enrich the debate about whether current AI architectures can benefit from frameworks established in neuroscience and philosophy of mind.

For this reason, this paper is organized as follows: section 2 first introduces PP and explains how it is related to AIF and FEP. Section 3 shows the impact of these frameworks on generative AI, and section 4 presents some directions for future research. Finally, some conclusions are drawn.

2. Predictive Processing, Free Energy Principle, Active Inference

By PP, or predictive coding, is meant the hierarchical model theorized by neuroscientist Karl Friston. PP is considered a unified explanation for perception, action and cognition, and sees the brain as a predictive machine that tries to predict its next states based on information gained through the previous interactions with the environment [1]. This means that the brain always tries to minimize the probability of prediction errors and avoid states with a high degree of surprise. This is also the concept of Bayesian inference, which reappears in the context of FEP. FEP, of which Friston is again the father, has been used in different contexts and with different objectives, and has been regarded as a more or less mysterious unification principle, that can be applied to almost any field of knowledge: from thermodynamics, to biology, to the study of the mind. Friston emphasizes that FEP is not a framework, or a theory, but should be understood as a tool or, in other words, a methodology. In its original formulation, FEP is a mathematical formula that states that agents exist because they can persist, maintaining their equilibrium by minimizing free energy. This minimization of energy is achieved by AIF, which – again – means changes in state and the suppression of prediction errors by information gained from the history of previous interactions with the environment [2][3]. AIF can be intuitively described as «feeling our way in darkness: we anticipate what we might touch next and then try to confirm those expectations» [1]. More generally, AIF provides a comprehensive framework to describe,

[1]Corresponding author.
EMAIL: maria.raffa@studenti.iulm.it.
ORCID: 0000-0002-4073-7440.

explain, simulate and understand the mechanisms underlying decision-making, perception, and action [4][5].

AIF modeling is very attractive for several reasons. For example, it has been used in the context of scientific research on introspection, self-modeling and self-access, which has led to several theories of consciousness [6]: introspection, i.e., the ability to access and evaluate one's own mental states, thoughts and experiences, plays a critical role in self-awareness, learning and decision-making, and is a pillar of human consciousness [7]. Mahault Albarracin and colleagues state that 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» [8]. Moreover, «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 [...] 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» [8]. The starting idea for a model of consciousness based on AIF is that for a system to be able to report and evaluate its own conclusions, it must be able to implement some form of self-access, i.e., some parts of the system can use the output of other parts as their own input, for further processing. This is related to the idea that some cognitive processes are "transparent". Like a clean window, they allow us to access other things without being perceptible themselves. Other cognitive processes are "opaque", meaning that they can be evaluated per se, such as introspective self-awareness (e.g., being aware that you are looking at a tree rather than seeing a tree). The idea, then, is that introspective processes make other cognitive processes accessible to the system as such, thus rendering them opaque. AIF has been used by Albarracin and colleagues to develop a three-level generative model that models self-access and introspective abilities in terms processes that govern transparency and opacity at the phenomenal level of description and attentional selection at the psychological level, respectively [8][9].

AIF, FEP and PP are not only used for modelling consciousness in the science of mind, but are also of interest in many other fields such as thermodynamics and biology, as they provide a very general and comprehensive framework for optimizing resources: FEP and AIF can be applied to almost all types of organisms and organizations that are able to sustain themselves with the least waste of resources and through the strategy of minimizing prediction errors. As mentioned earlier, this is the same concept of Bayesian inference, and Bayesian structures are also used in the field of generative AI.

On these grounds, the next section addresses the connection between AIF and AI models based on Bayesian networks.

3. Predictive Processing, Active Inference, AI

In the field of AI, the AIF framework, which is essentially Bayesian inference, is used in a wide variety of ways. One of the most popular applications concerns generative models, which are used for neural networks to generate images and text. The idea behind generative models is the same as for predictive perception based on Bayesian inference, namely minimization through errors. Indeed, these algorithms can generate sensory signals that match predicted causes, as they learn from small data sets and generalize to new situations [10]. Bayesian networks are a type of generative models where all processes can be easily tracked and understood: namely the propagation of events evolves through multiple ambiguous points, where the event diverges probabilistically between paths, and at any given point in the network, the probability of that node being visited depends on the joint probability of that node being visited is dependent on the joint probability of the preceding nodes. Therefore, Bayesian networks resemble explicit reasoning, although they are probabilistic. For this reason, they are more reliable in terms of explainability, than other models, such as neural models based on feedforward architectures trained with backpropagation: these models are very powerful in producing fast results, but they often lead to black box situations that are difficult to explain, since the inputs are known, but the path to the outputs is less clear [11].

As mentioned earlier, generative models based on Bayesian networks are used in AI for generating images and text. The most popular AI for images (e.g., Dall-E and Midjourney) are deep generative networks, which are a combination of a generative model and a deep neural network. This means it is a

basic neural network that has been augmented in both size and training data, both of which are required for good performance [12]. The whole is combined with a generative language model, whose task is to process the user's text input to generate the images, i.e., a Generative Pretrained Transformer network (GPT). A transformer network is a deep learning model that applies the mechanism of self-attention, differentially weighting the importance of each part of the input data [13]. These AI algorithms have access to billions of images on the Internet tagged with words. GPT can read this image data and extract some basic features (e.g., the style of a particular artist, or facial shapes), and hence, GPT manages to understand how much the images are coherent with the inputs submitted by the user. It means that during the process of producing new images, the generative networks act gradually adding new features built on the features that they have already seen.

As for a more direct connection between PP, AIF and AI, these cognitive systems are also interesting for robotics. In particular, for robotics applications where the dynamic of the robot or the task are uncertain: for estimation, adaptive control, fault-tolerant control, prospective planning and complex cognition skills (human-robot collaboration, discrimination between self and others) [14]. Pablo Lanillos and Gordon Cheng [15] have developed a computational model that allows a robot to determine its own body configuration by basing it on PP. More in detail, PP is used for a computational perception model that makes any multisensory humanoid robot able to learn, infer and update its body configuration. This model enables generic multisensory integration, by integrating different sources of information (tactile, visual and proprioceptive): the robot estimates its body configuration and adjusts it based on sensory information alone. Prior to Lanillos and Cheng, also Jun Tani [16] and Leo Pio Lopez and colleagues [17] also studied similar AIF-based models for real robots. Similar approaches have also been used for industrial manipulators [18], and for implementing active vision in robots in simulation environments [19]. Toon Van de Maele and his team proved in a simulation environment that a robot agent, moving without knowledge of the workspace, chooses the next visual position by evaluating the expected free energy. This means that it is possible to use the active inference paradigm as a «natural solution for active vision in complex tasks in which the distribution over the environment is not defined upfront» [19]. With the appropriate amount of GPU memory, a more complex generative model could also be trained to compute the expected free energy for all potential viewpoints of an agent, and not just for a limited set of considered ones: this would allow extending the use of the neural network for a real-time control of physical robot manipulators.

All the generative model based on Bayesian networks, which rely on FEP and AIF too, are worth to be investigated also because they could be considered more *sustainable* than more complex models (e.g., neural models based on feedforward architectures) in terms of fairness and explainability [20], since Bayesian networks, as already anticipated, imply processes that can be tracked and understood. They also refer to FEP, which provides a very general framework for resource optimisation. On the other hand, it may be risky to consider this type of algorithms as models for building AI cognition, since they rely on basic decision theory that assumes an optimal decision maker who is able to calculate and choose the move that maximizes the utility function at each stage of problem solving. These theories of maximising expected utility have been criticised as being computationally intractable [21][22], since they are computationally demanding for systems with a large number of random variables, and therefore lead to an exponential increase in computational effort and energy consumption, as the number of states increases.

Hence, the take-away message from this section is that AIF, PP and FEP are not only abstract models for cognition, mind or resource optimization, but they also have very concrete applications in AI. For this reason, research should delve more into this relationship's implications. On these grounds, the next section discusses what has been done on this path until now and what might be investigated in the future.

4. A few steps forward

The previous sections have dealt with the description of the cognitive frameworks of PP and AIF and have shown how they are employed for AI implementation. The relevance of this link has been noted by neuroscientists, who are studying the implications of these tools to draw new kinds of AI, characterized

by top-down systems which prioritize high-level planning and decision-making, rather than bottom-up systems, driven by masses of training data [23].

Moreover, Albarracin and colleagues [8] have argued that incorporating the design principles of AIF into AI systems can lead to better explainability for two reasons: the first is that «by deploying an explicit generative model, AI systems premised on AIF are designed to be interpreted by a user or a stakeholder, that can be fluent in the operation of such models». The second is that an architecture inspired by AIF models of introspection is useful to «build systems that are able to access and report on the reasons for their decisions, and their state of mind when reaching these decisions». Specifically, they propose an architecture including components that update and maintain an internal model of its own state, beliefs and goals: this capacity of self-access, enables the AI system to optimize and report in its decision-making processes, fostering introspection and enhanced explainability. This is done by a soft attention mechanism, which uses a weighted combination of hierarchical generative model components to focus on relevant information: the attention weights are 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 [8].

All that considered, Albarracin and colleagues indicate different research paths for future development. It is worth exploring, among other things, 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. The explanation and transparency of these kind of algorithms has been a significant topic in recent years, especially in decision-making scenarios.

In this sense, an interesting path might be found in the research about Bayesian networks for decision-making, specifically exploring a specific subset of a Bayesian network, i.e., a Markov blanket (MB): «the Markov blanket of a variable X is the set consisting of the parents of X, the children of X, and the variables sharing a child with X» [24]. This is, in a Bayesian network, the nodes forming the MB of X simplify a lot the computation of the value of a variable, since «if X is embedded in a graph with a hundred variables, but its MB consists of five variables, one can safely ignore ninety-five variables in the computation» [25]. Hence, in the framework of FEP and AIF, Friston introduced the statistical concept of MB, by advocating it as a tool to describe a specific form of conditional independence between a dynamical system and its environment, and by addressing it as real boundaries of living organisms [2][3]. Indeed, they are associated with the physical boundaries surrounding cells (i.e., their membrane), through which all influences between the intracellular and extracellular spaces are mediated. More broadly, MBs represent a separation of the world into different sets of states, which interact only via MB. Hence, although living systems need to interact with the environment, since they are open systems, they also need to distinguish themselves from the external environment: in this context, conditional independence occurs: when the state on the organism/environment boundary, i.e., the MB, is fixed, what happens on one side of the boundary does not influence what happens on the other side. This means that all the necessary information for explaining the behaviour of the internal system is given by the state of the blanket [26][27].

Besides the theoretical framework, MBs are employed in AI implementation, joint with AIF, for instance, in music generation [28] and in robotics. If we consider again one of the examples seen in the previous section, i.e., Van de Maele's simulation of the robotic agent moving in a space AIF based, also MBs structures occur. Indeed, in that context the robotic agent was designed as separated from the world state through a MB, meaning that the agent can only update its knowledge about the surrounding space by interacting with the world through its chosen actions and its observed sensory information [19].

Hence, when establishing research on MBs, one should be very careful about the basic assumptions to be made: this is, on the one hand, MB can be used instrumentally in their original statistical form, i.e., as a formal mathematical construct for inference on a generative model, e.g., a Bayesian network. On the other hand, this usage cannot come only from the theoretical framework established by Friston and cannot be justified with traditional statistics. However, grasping the potential of this tool and delving into its possible implications might be worth for further research on using AIF for AI architecture.

5. Conclusion

So far, this short paper has described the theoretical frameworks of PP and AIF, showing their linkages to FEP, and how all of them are employed in the field of AI. The paper has reported the research conducted by Mahault Albarracin and colleagues on the possibility of using AIF for implementing explainable and transparent AI architecture. Specifically, this work's final proposal is to follow this path by focusing on the potential of the statistical structures of MBs, which might be helpful for the development of more sustainable and transparent decision-making processes. This follows from the observation that the investigation of the relationship between AI and the cognitive frameworks provided by the philosophy of mind and neuroscience still has not yet received sufficient attention by literature and should be taken into account for future AI development, and, vice versa, delving into AI implementation might be helpful to shed some light on these cognitive frameworks.

References

[1] K. Friston, The free energy principle: A unified brain theory?, Nature Reviews Neuroscience 11 2 (2010) 127-138.

[2] K. Friston, J. Mattout, J. Kilner, Action understanding and active inference, Biological cybernetics, 104 1 (2011) 137-160.

[3] M. Kirchhoff, T. Parr, E. Palacios, The Markov Blanket of life: Autonomy, active inference and the free energy principle, Journal of the Royal Society Interface, 15 138 (2018). https://doi.org/10.1098/rsif.2017.0792.

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

[5] L. Da Costa, K. J. Friston, C. Heins, and G. A. Pavliotis, Bayesian mechanics for stationary processes, Proceedings of the Royal Society, A 477.2256 (2021). doi: 10.1098/rspa.2021.0518.

[6] M. Ramstead, M. Albarracin, A. Kiefer, B. Klein, C. Fields, K. Friston, A. Safron, The inner screen model of consciousness: applying the free energy principle directly to the study of conscious experience, (2023).

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

[8] M. Albarracin, I. Hipolito, S.E. Tremblay, Designing explainable artificial intelligence with active inference: A framework for transparent introspection and decision-making, ArXiv (2023). doi: arXiv:2306.04025.

[9] L. Sandved-Smith, C. Hesp, J. Mattout, K. J. Friston, A. Lutz, and M. J. D. Ramstead, Towards a computational phenomenology of mental action: Modelling meta-awareness and attentional control with deep parametric active inference, Neuroscience of Consciousness, 2021.1 (2021). doi: 10.1093/nc/ niab018.

[10] A. Seth, The brain as a prediction machine, in: D. Mendonça, M. Curado, S. S. Gouveia (Eds.), The Philosophy and Science of Predictive Processing, Bloomsbury, London, 2020, pp. XIV-XVII.

[11] I. P. Derks, A. deWaal, A taxonomy of explainable Bayesian networks, in: A. Gerber (Ed.), Artificial Intelligence Research, CCIS, Springer, Cham, 2020. https://doi. org/10.48550/arXiv.2101.11844.

[12] J. Kaplan, S. McCandlish, T. Henighan, Scaling Laws for Neural Language Models, arXiv (2020). arXiv:2001.08361v1.

[13] A. Vaswani, N. Shazeer, N. Parmar, Attention Is All You Need, ArXiv (2017). arXiv: 1706.03762.

[14] P. Lanillos, C. Meo, C. Pezzato, Active inference in robotics and artificial agents: survey and challenges, ArXiv (2021). arXiv:2112.01871.

[15] P. Lanillos, G. Cheng, 2018, Adaptive robot body learning and estimation through predictive coding, IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE (2018) 4083–4090.

[16] J. Tani, Learning to generate articulated behavior through the bottom-up and the top-down interaction processes, Neural networks 16(1) (2003) 11–23.
[17] L. Pio-Lopez, A. Nizard, K. Friston, G. Pezzulo, Active inference and robot control: a case study, Journal of The Royal Society Interface 13 (122) (2016).
[18] C. Pezzato, M. Baioumy, C. H. Corbato, N. Hawes, M. Wisse, R. Ferrari, Active inference for fault tolerant control of robot manipulators with sensory faults, International Workshop on Active Inference, Springer, (2020) 20–27.
[19] T. Van De Maele, T. Verbelen, O. Çatal, Active Vision for Robot Manipulators Using the Free Energy Principle, Frontiers in Neurorobotics 15 (2021). doi: 10.3389/fnbot.2021.642780.
[20] K. Imai, Z. Jiang, Principal Fairness for Human and Algorithmic Decision-Making, ArXiv (2022) arXiv:2005.10400.
[21] J. Kwisthout, I. van Rooij, Computational resource demands of a predictive Bayesian brain, Comput Brain Behav 3 (2020) 174-188. doi: https://doi.org/10.1007/s42113-019-00032-3.
[22] A. Lieto, Cognitive design for artificial minds, Routledge, New York, 2021.
[23] M. Cullen, B. Davey, K.J. Friston, R.J. Moran, Active Inference in OpenAI Gym: A Paradigm for Computational Investigations Into Psychiatric Illness, Biol Psychiatry Cogn Neurosci Neuroimagining 3 (9) (2018). doi: 10.1016/j.bpsc.2018.06.010.
[24] T. Koski, J. M. Noble, Bayesian Networks: An introduction, Wiley and Sons, Chichester, 2009.
[25] M. Facchin, Extended predictive minds: Do Markov blankets matter?, Review of Philosophy and Psychology (2021) 1-30. doi: https://doi.org/10.1007/S13164-021-00607-9.
[26] J. Hohwy, Quick'n'Lean or Slow and Rich? Andy Clark on predictive processing and embodied cognition, in: M. Colombo, E. Irvine, M. Stapleton (Eds.), Andy Clark and His Critics, Oxford University Press, New York, 2019, pp. 191-205.
[27] E. R. Palacios, A. Razi, T. Parr, On Markov Blanket and hierarchical self-organization, Journal of Theoretical Biology 486 (110089) (2020).
[28] J. Cruz, Deep learning vs Markov model in music generation, Honors College Theses, graduate thesis (2019).