Free Energy Principle and Active Inference in Neural Language Models

Maria Raffa[1,*,†], Alessandro Acciai[2,†]

[1]IULM University, via Carlo Bo 1, Milan, Italy
[2]University of Messina, Via Concezione 6, Messina, Italy

Abstract

The aim of this paper is twofold: first, to explore the relationship between Neural Language Models (NLMs) and the Free Energy Principle (FEP), and second, to suggest that NLMs, as undesigned cognitive architectures seen through the lens of FEP and Active Inference (AIF), can be considered potential candidates for the development of AI architectures that promote long-term sustainability. We argue that NLMs can be viewed as "undesigned cognitive architectures", that reflect principles of cognitive efficiency and resource optimisation. While NLMs were not intentionally designed to model cognition, they share significant features with cognitive architectures rooted in the FEP and AIF. These AI systems use generative models that optimise tasks such as language understanding and generation by minimising prediction errors. By aligning NLMs with the FEP and AIF, we show how these models contribute to sustainable AI by balancing performance, transparency and resource use. We also highlight how, despite their passive nature, NLMs share core goals with AIF systems, in particular the minimisation of uncertainty. Specifically, the structure of the paper is as follows: Section 1 introduces the concept of undesigned cognitive architectures, Section 2 explores the relationship between FEP, AIF and NLMs. Following, Section 3 focuses on sustainability considerations, and lastly, Section 4 draws conclusions.

Keywords

Neural Language Models, Free Energy Principle, Active Inference, Sustainability

1. Artificial Systems as Cognitive Architecture

The processes underlying human cognition over the last century have been addressed by artificial intelligence (AI) in an attempt to provide cognitive science with an artificial foundation through which the mind can be studied more thoroughly [1]. This collaboration has led to various attempts [2], and in this regard, the concept of Cognitive Architecture (CA) [3] emerged precisely with the intention of creating a solid foundation for studying the cognitive processes of the human mind with the help of AI. The goal of unifying human cognitive functions within a single architecture has distinguished CA studies from other AI research, as they aim to capture the various forms of human intelligence, which today, with the advent of Neural Language Models (NLMs), is referred to as Artificial General Intelligence [4]. Historically, research in this area has been guided by the use of relatively specific domains, such as vision or language, where the abstraction and simulation of underlying system mechanisms could be emulated and studied through an artificial surrogate. The opportunity now presented by Deep Learning is to leverage artificial artifacts suitable for studying human cognition, with NLMs [5][6] appearing as ideal candidates for this purpose. Even though the nature of the Transformer structure [7] closely tied to language processing and was not specifically designed with the goal of simulating human cognition. Incorporating this new type of AI within an unconventional structure that largely respects the principles of CA [6][7][8] could be seen as the creation of a kind of "undesigned" cognitive architecture. The criteria that the CA of an intelligent system must possess to be considered as such, involve the presence of certain characteristics [9] [10], which, according to our proposal, albeit with some contingencies, are part of the abilities demonstrated by NLMs when performing various tasks typically used to study human cognitive skills [11]. Two fundamental aspects to start with are the presence of recognition ability and decision-making (DM). The criterion of recognition in NLMs, i.e., the ability to relate knowledge

AIxPAC – 2nd Workshop on AI for Perception and Artificial Consciousness – AIxIA '24 – November 25-28, 2024, Bozen, Italy

*Corresponding author.

† These authors contributed equally.

maria.raffa@studenti.iulm.it (M. Raffa); alessandro.acciai@studenti.unime.it (A. Acciai)

0000-0002-4073-7440 (M. Raffa); 0000-0002-0877-706 (A. Acciai)

and be able to infer the correct patterns between agent and object and appropriately categorize situations and events, has been tested in various studies [12][13]. For example, Jin Han and colleagues [14] showed how models from the OpenAI family demonstrate property induction, extending the properties of some categories to others in particular situations when certain elements allow for sharing the same properties and the context is appropriate. Regarding DM ability, that is, making appropriate choices and selecting what is considered the best alternative among those offered in the environment in which one operates, Thilo Hagendorff and colleagues [11] highlight the adoption of "machine intuition," or the emergence of intuitive responses even in hostile contexts, by GPT-3.5 in a battery of tests designed to investigate intuitive DM in humans. The two aspects just mentioned, while important, are not sufficient on their own to achieve a status of ability to satisfy a CA: a broader range of criteria needs to be fulfilled. Two other aspects involve perception and situation assessment, abilities largely observed in the new Multimodal NLMs [15], where even the purely linguistic ones demonstrate a considerable ability to reason and navigate in the surrounding environment [16][17], even based on purely linguistic descriptions. Being able to interact in the environment is not enough. Several studies show that NLMs can also solve complex problems requiring analogical reasoning [18]; tackle tasks that require problem-solving ability using resources external to the system [16]; and exhibit the ability to reason about the behavior of other intelligent agents operating in their environment through the demonstration of higher cognitive functions such as Theory of Mind [19]. Finally, it is worth noting that that in the realization of a CA, especially considering the studies on Free Energy Principle (FEP) and Active Inference (AIF), some criteria more closely related to the issue of embodiment [20], although largely satisfied by current Transformer architectures applied to language processing, could, in any case, be completely exhausted in the broadest sense of the term in a very short time, given the rapid progress in the implementation of NLMs in the field of robotics [21][22][23].

Building on this foundation, the purposes of this paper are to investigate the relationship among NLMs, FEP and AIF, and to argue that NLMs as undesigned CA under FEP and AIF offer a model for sustainable AI. Indeed, by minimizing prediction errors, NLMs reflect principles of cognitive efficiency that are central to FEP and AIF, making them not only powerful in language tasks but also resource-efficient and adaptable to various contexts. For these aims, in the next two sections we examine the analogies and differences between NLMs, FEP, and AIF and consider the sustainability of NLMs as cognitive architectures, discussing their resource efficiency and adaptability.

2. Neural Language Models in Energy Saving Mode

Having suggested that NLMs can be considered as undesigned CA, we now examine how these systems share key features with the FEP and AIF. Indeed, both FEP and AIF emphasise efficiency and prediction within complex systems, a concept that is reflected in the way NLMs operate. However, there are important differences between NLMs and these frameworks.

The FEP, developed by Karl Friston, is a general principle which states that biological systems – both individuals and more complex systems such as communities and societies – exist because they can maintain the equilibrium between themselves and the environment by minimising free energy [24]. A practical realisation of FEP is the mechanism of predictive processing (PP), i.e., the process by which the brain minimises free energy or surprise. Indeed, the brain minimises prediction errors, namely, signal mismatches between the predicted input and the input actually received from the environment [24]. This minimisation can be achieved in a number of ways: by immediate inference about the hidden states of the world, which may explain perception; by updating a global world model to make better AI predictions, which may explain learning; and finally, by acting to sample sensory data from the world that matches the predictions [25]. PP has been advocated as a unified account for perception, action and cognition and can be described as an approximate Bayesian inference process based on Gaussian inference [25]. This means that in order to reduce surprise or uncertainty about their next states, systems use the information gained from previous interaction with the environment, using generative models to predict sensory inputs and minimise free energy [24]. The free energy minimisation is achieved through AIF and internal autoregulation, which ensures a constant updating of information gained from the environment, leading to accurate predictions of future next states [24][26]. In other words, through AIF, an organism (let's call it an agent) minimises free energy by updating its model of the world through observation and inference about the states of the world itself, as well as through action. This means that an agent actively modifies its own environment or its behaviour – which is defined by its actions – in order to make the environment – and the future – more predictable. In summary: FEP is a general theoretical framework that describes how single organisms or more complex systems minimise uncertainty in order to maintain their states. PP is the specific mechanism by which the brain reduces this uncertainty, as well as an operational implementation of the FEP. AIF is the process by which the organism/system acts to reduce uncertainty by integrating perception, action and learning.

As emphasised above, PP operates as an approximate Bayesian inference process, as the brain uses predictions based on prior experience to minimise the error between predicted inputs and actual sensory data. And this ongoing process of prediction error correction is central to maintaining cognitive efficiency and reducing uncertainty about the environment. Similarly, NLMs employ pre-trained generative architectures that perform tasks such as speech generation, comprehension, and context prediction by minimising errors in predicting next word sequences. Although NLMs are not explicitly designed to model uncertainty in the same way as biological systems, they exhibit behaviour consistent with PP. Indeed, just as the brain adjusts its predictions based on incoming sensory data to minimise prediction errors, NLMs adjust their word predictions based on large amounts of prior data to produce contextually appropriate output. Although their mechanism is based on statistical learning rather than explicit Bayesian inference, the overarching principle of reducing prediction error is similar to the goal of PP. Thus, although NLMs are not explicitly designed to minimise uncertainty in the same way as systems based on FEP, they exhibit behaviours consistent with the principles of error minimisation and efficient prediction inherent in PP. In contrast to the passive nature of NLMs, AIF systems engage in active exploration of the environment, constantly updating their predictions based on interactions with the world. Giovanni Pezzulo and colleagues [27] have argued that AIF generative models are characterised by being active, i.e., they incorporate action as a core mechanism for reducing uncertainty. In contrast, NLMs are generative models that operate passively – they generate predictions based on pre-existing data rather than through interaction with a dynamic environment. This difference is crucial: while NLMs are powerful in terms of language processing, they lack the adaptive, environment-driven characteristics inherent in the AIF model. Thus, the primary difference between NLMs and FEP and AIF models lies in their interaction with uncertainty. NLMs are trained on static data and passively generate responses based on previous inputs, whereas AIF models actively seek to minimise uncertainty through dynamic interaction. Despite these differences, both systems share the overarching goal of minimising error, which makes NLMs conceptually related to FEP and AIF in their prediction mechanisms. In terms of practical implementations of AIF, they are particularly valuable in uncertain environments, such as robotics, where estimation, adaptive control and human-robot collaboration rely on constant updates to predict and adapt based on sensory input. For example, models using PP have been applied to enable robots to learn and infer their body configurations from multisensory data [28]. In robotic applications, AIF-based systems have been shown to use active vision, selecting the most informative viewpoints to reduce uncertainty in dynamic environments. This adaptability is particularly valuable in tasks where the distribution over the environment is not predefined, as seen in recent simulations where robotic agents choose actions based on expected free energy to optimise task performance [29].

AIF models also promote transparency and traceability, making DM processes more understandable and ethically accountable. Unlike more complex models, such as deep neural networks based on feedforward architectures, AIF's reliance on Bayesian networks provides clearer, more interpretable processes, improving accountability and fairness. This transparency ensures that stakeholders can trust the DM process and that the system's actions can be easily traced, which is also a core principle of ethical AI [30]. In addition, AIF-based systems are highly adaptive, continuously updating and refining their models to respond to changing environments and contexts. This dynamic approach makes DM processes more robust and context-aware, allowing systems to balance short-term and long-term objectives. This adaptability, combined with the transparency and continuous improvement of AIF, provides a strong foundation for the development of sustainable and accountable AI systems [31].

In summary, although NLMs and AIF models have several differences, both share the goal of minimising prediction error, which links them to the principles of FEP and PP. In particular, AIF-based AI offers advantages in terms of accountability, transparency and sustainability, providing a robust framework for building systems that actively reduce uncertainty and dynamically adapt to their environment.

3. Undesigned but Sustainable Cognitive Architectures

So far, we explored how NLMs can be considered as undesigned CA and examined the links between NLMs, the FEP and AIF. All that considered, now we analyse why NLMs can also be considered sustainable CA through the lens of AIF-based sustainable models.

Sustainability in AI is a multifaceted concept. It includes sustainability in terms of the goals of the technology – such as creating tools that address sustainability challenges – and sustainability in terms of resource efficiency, both computationally and energetically [32]. There is also the dimension of social sustainability to consider: socially sustainable AI is also ethical AI, ensuring accountability and transparency [30]. AIF models promote transparency by making DM processes traceable from start to finish, allowing stakeholders to understand and trust these systems. In addition, AIF models are highly adaptable, dynamically adjusting to changing environments and requirements, which is critical in real-world scenarios where conditions can change rapidly. This adaptability results in DM processes that are more resilient and context-aware, rather than driven solely by immediate benefits. The continuous learning and updating capabilities of AIF further enhance their predictive capabilities, enabling these systems to refine strategies and optimise performance over time. AIF models are also equipped to operate across multiple time scales, balancing short-term and long-term objectives to improve overall system efficiency [31].

NLMs align with these sustainability principles through their inherent versatility and efficiency. Unlike traditional AI systems designed for specific tasks, NLMs are general-purpose models that can handle a wide range of applications – from text generation to translation – without requiring extensive retraining for each new task. This flexibility reduces the need for specialised models, saving both computational and human resources. The emergent capabilities of NLMs allow them to scale across domains, making them valuable tools for general problem solving, while maintaining a level of resource efficiency in line with sustainability goals. Once trained, NLMs operate efficiently across multiple tasks with minimal additional energy requirements, in contrast to the high cost of continuously retraining task-specific models.

Concerning practical examples of the relationship between FEP and AIF and NLMs, we can refer to medical applications: AIF has been employed to enhance the precision and contextual relevance of LLM responses, particularly in guiding the development of models that generate more accurate and contextually relevant results. For example, researchers have integrated AIF principles to enhance the efficacy of NLM-guided medical interventions, wherein models, informed by AIF, act as human therapists. The aforementioned systems comprise a "therapist agent" who responds to patients' queries and a "supervisor agent" who assesses the veracity and dependability of these responses. This method employs AIF to iteratively minimise prediction errors and enhance the quality of NLM-generated advice in intricate medical scenarios, particularly in the context of conditions such as insomnia therapy [33].

Another interesting example lies in the field of education, where the combination of AIF and NLMs facilitates the simulation of more active and embodied learning experiences. NLMs can be incorporated into educational settings, such as Montessori classrooms, where the tenets of AIF inform active learning. In this instance, LLMs are employed to facilitate interactions, assist students in formulating hypotheses, test them and reduce prediction errors. This hybrid approach emphasises exploration and engagement with material environments, in accordance with the predictive processing frameworks that drive human learning [34]. These examples illustrate how AIF enhances the real-world application of NLM by introducing an active, feedback-driven process that aligns with human cognitive and interactive dynamics.

However, we cannot neglect the sustainability challenges posed by NLMs, in particular the significant energy consumption during the initial training phase. As Joan Kwisthout and Iris van Rooij [35] note, systems based on Bayesian inference – including those aligned with FEP – become exponentially more computationally demanding as the number of variables increases. This complexity also affects NLMs, where large-scale models require significant resources. Mitigating this energy demand remains a critical challenge for the future development of sustainable AI. However, advances in hardware optimisation and more energy-efficient architectures can further reduce the environmental impact of NLMs training and contribute to the overall sustainability of these models.

Despite these challenges, NLMs offer a unique opportunity for advancing sustainable AI through their flexibility and explainability in virtue of the opportunity to compare their abilities based on tests used to study human cognition. Generative models such as NLMs can be traced, making their decision processes more interpretable than those of other AI systems. This traceability fosters ethical accountability, which is a critical component of sustainability. In this way, NLMs represent a compelling intersection of efficiency, adaptability, and functionality, key elements of sustainable AI. All the above considered, although NLMs and AIF differ in their approach to handling uncertainty and interaction with the environment, NLMs still exhibit features that make them viable candidates for sustainable AI architectures. Their generality, resource efficiency, and potential for transparency position them as critical models for future AI development, balancing performance with sustainability goals.

4. Conclusions

In this paper, we have explored the conceptual parallels between NLMs, FEP, and AIF, considering NLMs as undesigned CA. Through an examination of their shared characteristics – such as prediction error minimization and resource efficiency – alongside their differences in interaction with the environment, we have argued that NLMs represent a unique form of undesigned CA that offers new perspectives on both artificial cognition and sustainability in AI. By their very nature, NLMs do not follow explicit cognitive designs, yet they exhibit emergent behaviours consistent with the principles of the FEP and AIF. These models show a remarkable ability to generalise across tasks, minimising the need for highly specialised architectures. As a result, NLMs inherently promote sustainability goals within AI by optimising the use of data, computation and energy. Their versatility, coupled with resource-efficient operation, reflects the adaptive and resilient characteristics required for long-term sustainability. This analysis suggests that the future of AI development should increasingly consider undesigned CA as viable pathways for creating systems that balance high performance with sustainable resource use.

References

[1] J. Haugeland, Artificial Intelligence: The Very Idea, MIT, 1985.

[2] D. E. Rumelhart, J. L. McCelland, On learning the past tenses of English verbs, 2, 1986, pp. 216-271.

[3] A. Newell, Physical symbol systems, Cognitive Science, 4, (1980): 135-183.

[4] B. Goertzel, Artificial general intelligence: Concept, state of the art, and future prospects, Journal of Artificial General Intelligence, 5, (2014): 1-46.

[5] J. Devlin, M. W. Chang, K. Lee, K. Toutanova (Eds.), BERT: Pre-training of deep bidirectional transformers for language understanding, Proceedings North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, 2019: 4171-4186.

[6] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, Language models are few-shot learners, ArXiv, (2020), https://doi.org/10.48550/arXiv.2005.14165.

[7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in Neural Information Processing Systems, (2017): 6000-6010.

[8] R. Sun, Desiderata for cognitive architectures, Philosophical Psychology 17, (2004): 341-373.

[9] P. Langley, J. E. Laird, S. Rogers. Cognitive architectures: Research issues and challenges, Cognitive Systems Research 10, (2009): 141-160.

[10] A. Lieto, Cognitive design for artificial minds, Routledge, New York, NY, 2021.

[11] T. Hagendorff, S. Fabi, M. Kosinski, Machine intuition: Uncovering human-like intuitive decision-making in GPT-3.5, ArXiv, (2022), abs/2212.05206.

[12] B. Z. Li, M. Nye, J. Andreas (Eds.), Implicit representations of meaning in Neural Language Models, Proceedings of the 59th Meeting of the Association for Computational Linguistics, ACL, (2021): 1813-1827.

[13] G. H. Patel, D. M. Kaplan, L. H. Synder, Topographic organization in the brain: searching for general principles, The Science of Consciousness 18, (2014): 351-363.

[14] J. Han, M. Kamber, J. Pei, Data mining: concepts and techniques, 3rd edition, MK, 2012.

[15] H. Yang, S. Yang, A. M. Isen, Positive affect improves working memory: Implications for controlled cognitive processing, Cognition and Emotion 27, 3, Taylor & Francis, 2013, pp. 474-482.

[16] S. Bubeck, V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y. T. Lee, Y. Li, S. Lundberg, H. Nori, H. Palangi, M. T. Ribeiro, Y. Zhang, Sparks of Artificial General Intelligence: Early experiments with GPT-4, ArXiv, (2023), abs/2303.12712.

[17] J. S. Park, L. Popowski, C. J. Cai, M. R. Morris, P. Liang, M. S. Bernstein (Eds.), Social simulacra: Creating populated prototypes for social computing systems, Proceeding of ACM Symposium on User Interface Software and Technology 2022, ACM, pp. 1-18.

[18] T. Webb, K. J. Holyak, H. Lu, Emergent analogical reasoning in large language models, NHB 7, (2023): 1526-1541.

[19] M. Kosinski, Theory of mind may have spontaneously emerged in Large Language Models, ArXiv, (2023), abs/2302.02083.

[20] D. Vernon, Artificial cognitive systems: A primer, MIT, 2014.

[21] M. Ahn, A. Brohan, N. Brown, Y. Chebotar, O. Cortes, B. David, C. Finn, C. Fu, K. Gopalakrishnan, K. Hausman, A. Herzog, Do as I can, not as I say: Grounding language in robotic affordances, ArXiv, (2022), abs/2204.01691.

[22] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, J. Dabis, C. Finn, K. Gopalakrishnan, K. Hausman, A. Herzog, J. Hsu, J. Ibarz, B. Ichter, A. Irpan T. Jackson, S. Jesmonth, N. J. Joshi, R. Julian, D. Kalashnikov, Y. Kuang, I. Leal, K. Lee, S. Levine, Y. Lu, U. Malla, D. Manjunath, RT-1 Robotics Transformer for Real-World Control at Scale, ArXiv, (2022), abs/2212.06817.

[23] R. Firoozi, J. Tucker, S. Tian, A. Majumdar, J. Sun, W. Liu, Y. Zhu, S. Song, A. Kapoor, K. Hausman, B. Ichter, D. Driess, J. Wu, C. Lu, M. Schwager, Foundation Models in Robotics: Applications, Challenges, and the Future, ArXiv, (2023), abs/2312.07843.

[24] K. C. Friston, J. Mattout, Action understanding and active inference, Biological Cybernetics 104, (2011): 137–160.

[25] B. Millidge, A. Seth, K. C. Friston, Predictive coding: A theoretical and experimental review, ArXiV, (2022), doi: https://doi.org/10.48550/arXiv.2107.12979.

[26] M. D. Kirchhoff, Predictive processing, perceiving and imagining: Is to perceive to imagine, or something close to it?, Philosophical Studies 175, (2018): 751–767.

[27] G. Pezzulo, T. Parr, Generative meaning: Active inference and the scope and limits of passive AI, Trends in Cognitive Sciences 28, (2024).

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

[29] 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.

[30] F. Mazzi, L. Floridi (Eds.), The Ethics of Artificial Intelligence for the Sustainable Development Goals, Springer, Cham, Switzerland, 2023.

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

[32] S. N. Jan-Christoph Heilinger, Hendrik Kempt, Beware of sustainable AI! Uses and abuses of a worthy goal, AI Ethics, (2023): 1–12.

[33] R. Shusterman, A. C., Waters, S. O'Neill, P. Luu, D. M. Tucker, 2023, An Active Inference Strategy for Prompting Reliable Responses from Large Language Models in Medical Practice, arXiv: 2407.21051.

[34] L. D. Di Paolo, B. White, A. Guénin-Carlut, A. Constant, A. Clark, 2024, Active inference goes to school: the importance of active learning in the age of large language models. Phil. Trans. R. Soc. B 379: 20230148. https://doi.org/10.1098/rstb.2023.0148.

[35] J. Kwisthout, I. van Rooij, Computational resource demands of a predictive Bayesian brain, Computational Brain Behavior 3, (2024): 174–188.