Interpretability for Multi-Agent Systems Safety

Yoann Poupart *[1] and Nicolas Maudet †[2]

[1]ENS de Lyon
[2]LIP6, Sorbonne University

Abstract

Multi-agent systems (MAS) have been democratised in recent years thanks to the natural language interfacing made possible by large language models (LLM). While their ability to solve complex tasks is undeniable, the dynamics emerging from these systems can be hard to predict, and guarantees are needed. Jailbreak, adversariality or power-seeking are concerning failure modes of MAS, and evaluating these capabilities remains a difficult problem. In this respect, interpretability could be one of the best tools to monitor and control several agents simultaneously and automatically. Indeed the models' internals convey the information used for its prediction and can be used symbolically for gaining understanding or control.

1 Introduction

1.1 Context

Agency, in the meaning of planning and acting to achieve a goal, has always been an objective for AI and systems in general. This skill empowers the models to solve multi-step, complex, and diverse tasks, creating expert models or broader ones that can generalise out-of-distribution. Theoretically, this empowerment can be described as the optimisation of the number of potentially accessible states [1]. Yet, with this simple, intuitive description, it is already easy to see how oversight might be a complex problem not humanly scalable.

In addition to the increasing complexity of the agents and their environment, we need to ensure their safe impact on humans and society in general, and this might involve rethinking AI-human interactions [2]. While leading companies are developing more and more capable models, there is a need to fill in the gap in research on the safety of agentic AI systems [3].

1.2 Agents

In practice, AI can be obtained in various forms with various levels of autonomy, but agents are undoubtedly one the most powerful forms [4]. Their ulterior promise is to be able to achieve a goal with drive, which could make them incredibly useful but also harder to control. In practice, agents can be separated into two kinds.

LLM Agents LLMs can be augmented with extra modalities (vision, sound, video, etc.) and tools like LLaVA-Plus [5] in order to improve their versatility and impact. This is principally possible because LLMs have general capabilities (wide AI) and can orchestrate or delegate to other, more specific systems. These agents can then be configured and composed to create highly complex MAS out-of-the-box [6]. Thus, a MAS can come in different architectures, i.e. different interactions between agents, and be tailored for specific use cases as LLM can easily tackle different modalities and be more or less expert on a task [7].

*yoann.poupart@ens-lyon.org
†nicolas.maudet@lip6.fr

RL Agents RL is a powerful framework, often used to solve tasks that require planning or adaptability, where agents learn to achieve goals, whether direct, i.e. their reward or loss, or indirect, i.e. instrumental objectives that help them achieve their ultimate goals. Furthermore, these models can be accompanied by simulators to increase planning significantly, e.g. tree search for Stockfish [8] or Alpha Zero [9]. The capabilities of agents can be decoupled when interacting with each other in society settings [10]. While multi-agent reinforcement learning (MARL) is only a special case of RL, it contains challenging specificities like training and agents' interactions [11].

1.3 Systems Desiderata

As AI develops, it is essential to keep sound and safe design guidelines in mind. While it is no different for multi-agent systems, just a particular kind of AI, they exhibit specific challenges. I provide below an uncomprehensive[1] list of these challenges.

Transparency The cornerstone of making more trusted AI requires a new framework to audit and evaluate models [12]. It is especially imperative in the context of decision-making; as we give more capability and autonomy to our systems, we need to be able to monitor and oversee them [13].

Cooperation: Teaching cooperation to agents is a fundamental objective towards safer AI [14], and it can be seen as a consensus-reaching problem. Yet consensus, a multi-objective optimisation problem, remains challenging [15], and getting away from the adversarial setting might require thinking over the training and interaction methods [16].

Controllability While MAS controllability can be approached with theoretical graph considerations from complex systems [17, 18], it might be impracticable to apply to LLM or RL agents. Another approach is to consider mechanistic interpretability methods for model oversight and control, more details in section 3.3.

2 Agents Failure Modes

It is important to understand where and how we sometimes fail to exercise control over our systems in order to understand why we need more of it.

2.1 Jailbreak

Jailbreak is a shortcut to refer to adversarial attacks that remove the safety training, the jail, in LLMs. Adversarial attacks, a classical in the vision domain [19], derived on encoders like BERT [20], are widely spreading with language models and their new modalities [21, 22]. While it has been shown that language models can learn dangerous capabilities [23], they often can somewhat be mitigated by using methods like RLHF [24]. Yet, as LLMs have become more capable of simulating humans' thoughts, it is now possible to psychologically analyse their thinking process and thus attack them [25].

Obviously, MAS is not exempt from this flaw, as adding more agents to the system only makes the defence more complicated. First, LLMs can be hacked using intermediate agents as hackers [26]. In addition, LLM agents equipped with tools are also vulnerable to external data they might retrieve from these tools, leading to indirect prompt injection [27].

2.2 Deception

The behaviour of an agent is said to be deceptive when it acts to maximise A while apparently aiming for B. This behaviour is especially concerning as it can persist even after quantitative alignment

[1] I left aside issues like fairness or misuse as they might fall into ethics and governance, not explored in this proposal.

training, producing the so-called "Sleeper Agents" [28] and pointing to the imperfection of alignment training like RLHF [24].

Adversariality While deception emergence might not always be the default, it remains a concern in terms of generalisation and data poisoning. For example, this phenomenon can emerge as a particular form of overfitting, where the model might exploit flaws in our measurements, e.g. goal misgeneralisation [29].

Power-seeking While agents can be trained to seek power, it also is a natural property emerging as an instrumental goal [30]. In decision-making, it is common as it maximises the system possibilities of action and thus its utility [31].

3 Interpretability

Interpretability is about trying to understand AI knowledge and thus can be a tool to learn from AI, like in Alpha Zero [32]. But it can also be used to monitor and control AI and make ourselves understood by AI, like with the representation engineering framework [33].

3.1 LLM Interpretability

While certain capabilities of LLMs can be evaluated intuitively, the exact models' internal processes remain poorly understood, even for the most basic capabilities, like addition [34]. Some methods are trying to automate the interpretability process using causal edition to find circuits [35], but still need a human to step in the loop. In this respect, scalability can be increased using agents to automatically evaluate the relevance of interpretability [36] or by replacing the human interpreter with a language model [37], but explaining complex reasoning remains a challenge [38].

3.2 RL Interpretability

Interpreting planning, which is mostly emerging from the search component of these agents, remains an open question. For example, in chess with Stockfish [8] using the efficient alpha-beta pruning [39] or with Alpha Zero [9] using MCTS [40]. Existing works trying to interpret these agents limit their analysis to the heuristic network without the tree search [41, 42], even on simpler environments like mini-chess or Hex [43, 44]. Even if dynamic concepts were introduced in [32], they only cover a few search steps. While specific methods for interpreting these agents might be required [45], probing for concepts or activation vectors remain well-established methods [46, 47].

3.3 A Tool for Control

Concepts or features can commonly be found in the embedding or latent space of the model by analysing its internals [48]. These findings can illustrate how the model represents well-known concepts, like truth, [49, 50], and they can even describe information processing by finding functional components [51]. These concepts can then be used to modify the model to gain more control over it [52–54]. This domain, known as Representation Engineering [33], seeks to manipulate the model's internals in order to make them more transparent and controllable, e.g. removing biases, restraining the outputs, aligning the beliefs, .etc. It is a new form of symbolic control which could enable stronger safety guarantees from logical computations as classically used in RL settings [55, 56].

4 Problem Fit

I now briefly describe why I think I have the capacity to explore the outlined problems and make a significant contribution to the field. While my background is in physics, I turned towards computer

science early, building side projects and graduating with a degree in complex systems. This training gave me the **theoretical** background and my projects the **practical** one, which is needed for the challenges ahead. While employed as an **ML engineer**, I harvested the skillset to create modular and clean software as well as train, evaluate and deploy **AI models**. I have already worked on **RL interpretability** projects, some of which are under continuation, and I am deeply motivated to make **AI safer**.

References

[1] A. Klyubin, D. Polani, and C. Nehaniv, "Empowerment: a universal agent-centric measure of control," in *2005 IEEE Congress on Evolutionary Computation*, vol. 1, pp. 128–135 Vol.1, 2005.

[2] C. Mitelut, B. Smith, and P. Vamplew, "Intent-aligned ai systems deplete human agency: the need for agency foundations research in ai safety," 2023.

[3] Y. Shavit, S. Agarwal, M. Brundage, C. authors, S. Adler, C. O'Keefe, R. Campbell, T. Lee, P. Mishkin, T. Eloundou, A. Hickey, K. Slama, L. Ahmad, P. McMillan, A. Beutel, A. Passos, and D. G. Robinson, "Practices for governing agentic ai systems," 2023.

[4] M. R. Morris, J. Sohl-dickstein, N. Fiedel, T. Warkentin, A. Dafoe, A. Faust, C. Farabet, and S. Legg, "Levels of agi: Operationalizing progress on the path to agi," 2024.

[5] S. Liu, H. Cheng, H. Liu, H. Zhang, F. Li, T. Ren, X. Zou, J. Yang, H. Su, J. Zhu, L. Zhang, J. Gao, and C. Li, "Llava-plus: Learning to use tools for creating multimodal agents," 2023.

[6] Q. Wu, G. Bansal, J. Zhang, Y. Wu, B. Li, E. Zhu, L. Jiang, X. Zhang, S. Zhang, J. Liu, A. H. Awadallah, R. W. White, D. Burger, and C. Wang, "Autogen: Enabling next-gen llm applications via multi-agent conversation," 2023.

[7] T. Guo, X. Chen, Y. Wang, R. Chang, S. Pei, N. V. Chawla, O. Wiest, and X. Zhang, "Large language model based multi-agents: A survey of progress and challenges," 2024.

[8] Y. Nasu, "Nnue efficiently updatable neural-network based evaluation functions for computer shogi," *Ziosoft Computer Shogi Club*, 2018.

[9] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, Y. Chen, T. Lillicrap, F. Hui, L. Sifre, G. van den Driessche, T. Graepel, and D. Hassabis, "Mastering the game of go without human knowledge," *Nature*, vol. 550, pp. 354–, Oct. 2017.

[10] M. Zhuge, H. Liu, F. Faccio, D. R. Ashley, R. Csordás, A. Gopalakrishnan, A. Hamdi, H. A. A. K. Hammoud, V. Herrmann, K. Irie, L. Kirsch, B. Li, G. Li, S. Liu, J. Mai, P. Piekos, A. Ramesh, I. Schlag, W. Shi, A. Stanić, W. Wang, Y. Wang, M. Xu, D.-P. Fan, B. Ghanem, and J. Schmidhuber, "Mindstorms in natural language-based societies of mind," 2023.

[11] S. V. Albrecht, F. Christianos, and L. Schäfer, *Multi-Agent Reinforcement Learning: Foundations and Modern Approaches*. MIT Press, 2024.

[12] I. D. Raji, A. Smart, R. N. White, M. Mitchell, T. Gebru, B. Hutchinson, J. Smith-Loud, D. Theron, and P. Barnes, "Closing the ai accountability gap: defining an end-to-end framework for internal algorithmic auditing," *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 2020.

[13] B. Lepri, N. Oliver, E. Letouzé, A. S. Pentland, and P. Vinck, "Fair, transparent, and accountable algorithmic decision-making processes," *Philosophy & Technology*, vol. 31, pp. 611–627, 2018.

[14] T. Franzmeyer, M. Malinowski, and J. F. Henriques, "Learning altruistic behaviours in reinforcement learning without external rewards," 2022.

[15] A. Amirkhani and A. H. Barshooi, "Consensus in multi-agent systems: A review," *Artif. Intell. Rev.*, vol. 55, p. 3897–3935, jun 2022.

[16] L. Yuan, Z. Zhang, L. Li, C. Guan, and Y. Yu, "A survey of progress on cooperative multi-agent reinforcement learning in open environment," 2023.

[17] A. Rahmani, M. Ji, M. Mesbahi, and M. Egerstedt, "Controllability of multi-agent systems from a graph-theoretic perspective," *SIAM Journal on Control and Optimization*, vol. 48, no. 1, pp. 162–186, 2009.

[18] Y.-Y. Liu, J.-J. E. Slotine, and A.-L. Barabási, "Controllability of complex networks," *Nature*, vol. 473, pp. 167–173, 2011.

[19] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," 2015.

[20] L. Li, R. Ma, Q. Guo, X. Xue, and X. Qiu, "Bert-attack: Adversarial attack against bert using bert," 2020.

[21] E. Shayegani, Y. Dong, and N. Abu-Ghazaleh, "Jailbreak in pieces: Compositional adversarial attacks on multi-modal language models," 2023.

[22] X. Qi, K. Huang, A. Panda, P. Henderson, M. Wang, and P. Mittal, "Visual adversarial examples jailbreak aligned large language models," 2023.

[23] J. Yang, A. Prabhakar, S. Yao, K. Pei, and K. R. Narasimhan, "Language agents as hackers: Evaluating cybersecurity skills with capture the flag," in *Multi-Agent Security Workshop @ NeurIPS'23*, 2023.

[24] P. Christiano, J. Leike, T. B. Brown, M. Martic, S. Legg, and D. Amodei, "Deep reinforcement learning from human preferences," 2023.

[25] R. Shah, Q. Feuillade-Montixi, S. Pour, A. Tagade, S. Casper, and J. Rando, "Scalable and transferable black-box jailbreaks for language models via persona modulation," 2023.

[26] M. Terekhov, R. Graux, E. Neville, D. Rosset, and G. Kolly, "Second-order jailbreaks: Generative agents successfully manipulate through an intermediary," in *Multi-Agent Security Workshop @ NeurIPS'23*, 2023.

[27] K. Greshake, S. Abdelnabi, S. Mishra, C. Endres, T. Holz, and M. Fritz, "Not what you've signed up for: Compromising real-world llm-integrated applications with indirect prompt injection," 2023.

[28] E. Hubinger, C. Denison, J. Mu, M. Lambert, M. Tong, M. MacDiarmid, T. Lanham, D. M. Ziegler, T. Maxwell, N. Cheng, A. Jermyn, A. Askell, A. Radhakrishnan, C. Anil, D. Duvenaud, D. Ganguli, F. Barez, J. Clark, K. Ndousse, K. Sachan, M. Sellitto, M. Sharma, N. DasSarma, R. Grosse, S. Kravec, Y. Bai, Z. Witten, M. Favaro, J. Brauner, H. Karnofsky, P. Christiano, S. R. Bowman, L. Graham, J. Kaplan, S. Mindermann, R. Greenblatt, B. Shlegeris, N. Schiefer, and E. Perez, "Sleeper agents: Training deceptive llms that persist through safety training," 2024.

[29] L. Langosco, J. Koch, L. Sharkey, J. Pfau, L. Orseau, and D. Krueger, "Goal misgeneralization in deep reinforcement learning," 2023.

[30] A. M. Turner, L. R. Smith, R. Shah, A. Critch, and P. Tadepalli, "Optimal policies tend to seek power," in *Advances in Neural Information Processing Systems* (A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, eds.), 2021.

[31] A. M. Turner and P. Tadepalli, "Parametrically retargetable decision-makers tend to seek power," 2022.

[32] L. Schut, N. Tomasev, T. McGrath, D. Hassabis, U. Paquet, and B. Kim, "Bridging the human-ai knowledge gap: Concept discovery and transfer in alphazero," 2023.

[33] A. Zou, L. Phan, S. Chen, J. Campbell, P. Guo, R. Ren, A. Pan, X. Yin, M. Mazeika, A.-K. Dombrowski, S. Goel, N. Li, M. J. Byun, Z. Wang, A. Mallen, S. Basart, S. Koyejo, D. Song, M. Fredrikson, J. Z. Kolter, and D. Hendrycks, "Representation engineering: A top-down approach to ai transparency," 2023.

[34] P. Quirke and F. Barez, "Understanding addition in transformers," 2024.

[35] A. Conmy, A. N. Mavor-Parker, A. Lynch, S. Heimersheim, and A. Garriga-Alonso, "Towards automated circuit discovery for mechanistic interpretability," 2023.

[36] S. Schwettmann, T. R. Shaham, J. Materzynska, N. Chowdhury, S. Li, J. Andreas, D. Bau, and A. Torralba, "Find: A function description benchmark for evaluating interpretability methods," 2023.

[37] S. Bills, N. Cammarata, D. Mossing, H. Tillman, L. Gao, G. Goh, I. Sutskever, J. Leike, J. Wu, and W. Saunders, "Language models can explain neurons in language models." `https://openaipublic.blob.core.windows.net/neuron-explainer/paper/index.html`, 2023.

[38] Y. Hou, J. Li, Y. Fei, A. Stolfo, W. Zhou, G. Zeng, A. Bosselut, and M. Sachan, "Towards a mechanistic interpretation of multi-step reasoning capabilities of language models," 2023.

[39] D. E. Knuth and R. W. Moore, "An analysis of alpha-beta pruning," *Artificial intelligence*, vol. 6, no. 4, pp. 293–326, 1975.

[40] L. Kocsis and C. Szepesvári, "Bandit based monte-carlo planning," in *Machine Learning: ECML 2006* (J. Fürnkranz, T. Scheffer, and M. Spiliopoulou, eds.), (Berlin, Heidelberg), pp. 282–293, Springer Berlin Heidelberg, 2006.

[41] T. McGrath, A. Kapishnikov, N. Tomaš ev, A. Pearce, M. Wattenberg, D. Hassabis, B. Kim, U. Paquet, and V. Kramnik, "Acquisition of chess knowledge in AlphaZero," *Proceedings of the National Academy of Sciences*, vol. 119, nov 2022.

[42] A. Pálsson and Y. Björnsson, "Unveiling concepts learned by a world-class chess-playing agent,"

[43] P. Hammersborg and I. Strümke, "Reinforcement learning in an adaptable chess environment for detecting human-understandable concepts," *IFAC-PapersOnLine*, vol. 56, no. 2, pp. 9050–9055, 2023.

[44] C. Lovering, J. Forde, G. Konidaris, E. Pavlick, and M. Littman, "Evaluation beyond task performance: Analyzing concepts in alphazero in hex," *Advances in Neural Information Processing Systems*, vol. 35, pp. 25992–26006, 2022.

[45] P. Hammersborg and I. Strümke, "Information based explanation methods for deep learning agents–with applications on large open-source chess models," *arXiv preprint arXiv:2309.09702*, 2023.

[46] G. Alain and Y. Bengio, "Understanding intermediate layers using linear classifier probes," 2018.

[47] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas, and R. Sayres, "Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav)," 2018.

[48] G. Dar, M. Geva, A. Gupta, and J. Berant, "Analyzing transformers in embedding space," 2023.

[49] S. Marks and M. Tegmark, "The geometry of truth: Emergent linear structure in large language model representations of true/false datasets," 2023.

[50] C. Tigges, O. J. Hollinsworth, A. Geiger, and N. Nanda, "Linear representations of sentiment in large language models," 2023.

[51] J. Dunefsky and A. Cohan, "Observable propagation: A data-efficient approach to uncover feature vectors in transformers," 2023.

[52] C. J. Anders, L. Weber, D. Neumann, W. Samek, K.-R. Müller, and S. Lapuschkin, "Finding and removing clever hans: Using explanation methods to debug and improve deep models," 2020.

[53] N. Belrose, D. Schneider-Joseph, S. Ravfogel, R. Cotterell, E. Raff, and S. Biderman, "Leace: Perfect linear concept erasure in closed form," 2023.

[54] N. Rimsky, N. Gabrieli, J. Schulz, M. Tong, E. Hubinger, and A. M. Turner, "Steering llama 2 via contrastive activation addition," 2023.

[55] N. Jansen, B. Könighofer, S. Junges, A. C. Serban, and R. Bloem, "Safe reinforcement learning via probabilistic shields," 2019.

[56] W.-C. Yang, G. Marra, G. Rens, and L. D. Raedt, "Safe reinforcement learning via probabilistic logic shields," 2023.