Deceptive Alignment Monitoring

Andres Carranza [*1] **Dhruv Pai** [*1] **Rylan Schaeffer** [*1] **Arnuv Tandon** [*1] **Sanmi Koyejo** [1]

Abstract

As the capabilities of large machine learning models continue to grow, and as the autonomy afforded to such models continues to expand, the spectre of a new adversary looms: *the models themselves*. The threat that a model might behave in a seemingly reasonable manner, while secretly and subtly modifying its behavior for ulterior reasons is often referred to as deceptive alignment in the AI Safety & Alignment communities. Consequently, we call this new direction *Deceptive Alignment Monitoring*. In this work, we identify emerging directions in diverse machine learning subfields that we believe will become increasingly important and intertwined in the near future for deceptive alignment monitoring, and we argue that advances in these fields present both long-term challenges and new research opportunities. We conclude by advocating for greater involvement by the adversarial machine learning community in these emerging directions.

1. Introduction

Machine learning models are growing increasingly general-purpose while simultaneously being granted increasingly more autonomy. The combination of greater capabilities and greater freedom in choosing when and how to exercise those capabilities raises the spectre that models themselves may behave adversarially to human interests (Hubinger et al., 2021; Hendrycks et al., 2021; Ngo et al., 2023). In the AI Safety and Alignment communities, this threat is often referred to as deceptive alignment, and large-scale collaborations spanning leading machine learning organizations and institutions are focusing more on how to measure problematic capabilities of large-scale models via benchmarks (Pan et al., 2023; Shevlane et al., 2023).

In this Blue Sky submission, we aim to elaborate on the different possible threats that large models themselves pose when viewed as potential adversaries, and argue that new methods are urgently needed. These methods must go beyond evaluation on static benchmarks to active, dynamic monitoring of models at multiple stages in a model's lifecycle: dataset creation and curation, training and learning, network representations and circuit mechanisms.

2. Deceptive Alignment Monitoring

2.1. Data Creation & Curation

The use of foundation models to generate and/or curate their own training data has recently emerged in both language and vision. In language generation, Self-Instruct used one language model to bootstrap training data for fine-tuning a different language model (Wang et al., 2022), and similar methods have quickly become commonplace, e.g., (Peng et al., 2023; Sun et al., 2023; Xu et al., 2023). In curation, SemDeDup uses an high-performing language model to prune a dataset for training future language models(Abbas et al., 2023), and DoReMi learns to weight data differently to improve performance when learning (Xie et al., 2023). In vision, (Li et al., 2023) propose endowing a model with the ability to learn to construct its own training dataset in a self-supervised manner by trawling through the internet. Such methods will continue to improve and proliferate.

These methods are relevant to deceptive alignment because if a highly capable, autonomous model does wish to change its behavior, creating or curating training data are straightforward mechanisms by which it can achieve such a goal. Currently, such methods have a high degree of human-in-the-loop, but as model scale and complexity increase, humans will play a vanishing role. To the best of our knowledge, we lack methods for monitoring the training data that our models generate or curate to check that the training data has not been selected in pursuit of some ulterior goal. This threat is spiritually similar to recent work demonstrating that web-scale datasets can be poisoned (Carlini et al., 2023); however, in the setting studied by Carlini et al. (2023), the dataset is known, fixed and constructed by humans, whereas in this setting, the dataset is unknown, dynamic and constructed by models in a largely unsupervised manner.

[*]Equal contribution [1]Computer Science, Stanford University. Correspondence to: Rylan Schaeffer <rschaef@cs.stanford.edu>.

2^{nd} AdvML Frontiers workshop at 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

2.2. Training, Editing & Learning

Turning now from data to learning, there are two emerging trends that pose new challenges: using one model to train another model, and directly editing the parameters of a model to change its behavior. The first trend appears commonly when using reward or preference models to fine-tune a language model, such as in Reinforcement Learning from Human Feedback (Ouyang et al., 2022; Bai et al., 2022a), Reinforcement Learning from AI Feedback (Bai et al., 2022b), Reinforcement Learning for Feedback (Akyürek et al., 2023) and many more, e.g., image synthesis (Lee et al., 2023). The second trend includes several methods for model editing, i.e., directly editing the model's parameters to change its behavior on specific tasks or inputs, e.g., Meng et al. (2022a); Mitchell et al. (2022); Meng et al. (2022b); Hartvigsen et al. (2022); Ravfogel et al. (2022); Hoelscher-Obermaier et al. (2023). In some methods, models are directly trained to learn to modify their own parameters, e.g., Irie et al. (2022); Kirsch & Schmidhuber (2022b;a).

Why does this matter for deceptive alignment? If models can have ulterior motives and act on said motives, then enabling models to train each other or to edit themselves directly posses a significant threat, since a model can give itself or give another model capabilities that humans would rather the model(s) not possess. To the best of our knowledge, we lack techniques to audit what learning signals one model transmits to another model, and we also lack techniques to understand whether model editing methods can themselves be poisoned; for a concrete example, it might be possible that by selective choosing a small number of seemingly innocent edits (e.g., "US President: Donald Trump → Joe Biden" and "World Chess Champion: Magnus Carlsen → Ding Liren"), the combination of edits induces an unpredictable and undesirable change in model behavior[1].

2.3. Representations & Mechanisms

A key step towards addressing the identified challenges will be monitoring the representations and mechanisms of large-scale models for "permissible" behavior (Christiano, 2022a;b); here, "permissible" behavior is loosely understood since the concept itself is in desperate need of sharper understanding. Loosely, if we think of a network as some map $f : \mathcal{X} \rightarrow \mathcal{Y}$, previous work has studied adversarial inputs coming from an unusual input distribution $\tilde{p}(x)$ or backdoored models with unusual conditional distributions $p(y|x)$. We are therefore concerned not with unusual input data, nor unusual conditional distributions, but whether the network's outputs are produced for the "normal reasons". As Christiano (2022a) explains, we are looking for outliers in terms of how f handles perfectly normal data.

The reason why this matters is that while inputs, outputs, and inputs-to-output maps may all appear perfectly ordinary, highly capable models may conceal some ulterior motive or goal. In order to detect and counter this threat, it is imperative to develop interpretability methods capable of understanding both activations and circuit mechanisms of models and detecting abnormal variations in models' processing. This is important because two models generating identical outputs when provided with the same inputs may compute their outputs for different "reasons".

Because the exact threat is unknown and likely dynamic, there is an urgent need to develop unsupervised methods for mechanistic anomaly detection that scale well. The next step is to develop an unsupervised methodology that can differentiate between normal model behaviors—where conclusions are reached for "appropriate reasons"—and anomalous model behaviors—where conclusions are derived for erroneous, potentially harmful "reasons". This task is particularly challenging since it requires the identification of patterns in the propagation of information through a model that is indicative of correct reasoning without relying on explicitly supervisory signals.

To achieve this, we propose leveraging techniques related to unsupervised anomaly detection to capture deviations from typical model behaviors. By comparing a model's processing across various inputs and outputs, it may be possible to identify patterns that consistently align with desired and appropriate behavior. We hypothesize that these patterns could manifest at three different levels of analysis within a model. Firstly, at the individual layer, a comprehensive analysis of activation distributions in the high-dimensional activation space could provide valuable insights into the model's processing. Secondly, at the layer-to-layer activation level, investigating how high-dimensional modes propagate, transform and evolve through the layers of a model can also offer an understanding of normal and abnormal processing. Thirdly, at the circuit level, identifying subgraphs within the network that correspond to specific transformations on features relevant to out-of-domain generalization might also prove powerful; however, knowing how to usefully define probabilistic distribution over activations, activations' propagations and circuit mechanisms for anomaly detection are, to the best of our knowledge, open questions. For possible approaches, see Carranza et al. (2023).

3. Outlook

The human-model interpretability quest can be modeled as an adversarial game, whereby deceptively aligned models subvert interpretability tools in favor of capabilities. More capable models are increasingly threatening, and to maintain scalable oversight we advocate development of novel tools for deceptive alignment monitoring.

[1] However, we note that sometimes seemingly unpredictable changes are indeed predictable, e.g., Schaeffer et al. (2023).

References

Abbas, A., Tirumala, K., Simig, D., Ganguli, S., and Morcos, A. S. Semdedup: Data-efficient learning at web-scale through semantic deduplication. *arXiv preprint arXiv:2303.09540*, 2023.

Akyürek, A. F., Akyürek, E., Madaan, A., Kalyan, A., Clark, P., Wijaya, D., and Tandon, N. Rl4f: Generating natural language feedback with reinforcement learning for repairing model outputs. *arXiv preprint arXiv:2305.08844*, 2023.

Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., Das-Sarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., Joseph, N., Kadavath, S., Kernion, J., Conerly, T., El-Showk, S., Elhage, N., Hatfield-Dodds, Z., Hernandez, D., Hume, T., Johnston, S., Kravec, S., Lovitt, L., Nanda, N., Olsson, C., Amodei, D., Brown, T., Clark, J., McCandlish, S., Olah, C., Mann, B., and Kaplan, J. Training a helpful and harmless assistant with reinforcement learning from human feedback, 2022a.

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022b.

Carlini, N., Jagielski, M., Choquette-Choo, C. A., Paleka, D., Pearce, W., Anderson, H., Terzis, A., Thomas, K., and Tramèr, F. Poisoning web-scale training datasets is practical. *arXiv preprint arXiv:2302.10149*, 2023.

Carranza, A., Pai, D., Tandon, A., Schaeffer, R., and Koyejo, S. Facade: A framework for adversarial circuit anomaly detection and evaluation, 2023.

Christiano, P. Mechanistic anomaly detection and elk, 2022a. URL https://www.lesswrong.com/posts/vwt3wKXWaCvqZyF74/mechanistic-anomaly-detection-and-elk. Accessed on May 28, 2023.

Christiano, P. Can we efficiently distinguish different mechanisms?, 2022b. URL https://www.lesswrong.com/posts/JLyWP2Y9LAruR2gi9/can-we-efficiently-distinguish-different-mechanisms. Accessed on May 28, 2023.

Hartvigsen, T., Sankaranarayanan, S., Palangi, H., Kim, Y., and Ghassemi, M. Aging with grace: Lifelong model editing with discrete key-value adaptors. *arXiv preprint arXiv:2211.11031*, 2022.

Hendrycks, D., Carlini, N., Schulman, J., and Steinhardt, J. Unsolved problems in ml safety. *arXiv preprint arXiv:2109.13916*, 2021.

Hoelscher-Obermaier, J., Persson, J., Kran, E., Konstas, I., and Barez, F. Detecting edit failures in large language models: An improved specificity benchmark. In *Findings of ACL*. Association for Computational Linguistics, 2023.

Hubinger, E., van Merwijk, C., Mikulik, V., Skalse, J., and Garrabrant, S. Risks from learned optimization in advanced machine learning systems, 2021.

Irie, K., Schlag, I., Csordás, R., and Schmidhuber, J. A modern self-referential weight matrix that learns to modify itself. In *International Conference on Machine Learning*, pp. 9660–9677. PMLR, 2022.

Kirsch, L. and Schmidhuber, J. Eliminating meta optimization through self-referential meta learning. *arXiv preprint arXiv:2212.14392*, 2022a.

Kirsch, L. and Schmidhuber, J. Self-referential meta learning. In *First Conference on Automated Machine Learning (Late-Breaking Workshop)*, 2022b.

Lee, K., Liu, H., Ryu, M., Watkins, O., Du, Y., Boutilier, C., Abbeel, P., Ghavamzadeh, M., and Gu, S. S. Aligning text-to-image models using human feedback, 2023.

Li, A. C., Brown, E., Efros, A. A., and Pathak, D. Internet explorer: Targeted representation learning on the open web. *arXiv preprint arXiv:2302.14051*, 2023.

Meng, K., Bau, D., Andonian, A., and Belinkov, Y. Locating and editing factual associations in gpt. *Advances in Neural Information Processing Systems*, 35:17359–17372, 2022a.

Meng, K., Sharma, A. S., Andonian, A., Belinkov, Y., and Bau, D. Mass-editing memory in a transformer. *arXiv preprint arXiv:2210.07229*, 2022b.

Mitchell, E., Lin, C., Bosselut, A., Manning, C. D., and Finn, C. Memory-based model editing at scale. In *International Conference on Machine Learning*, pp. 15817–15831. PMLR, 2022.

Ngo, R., Chan, L., and Mindermann, S. The alignment problem from a deep learning perspective, 2023.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

Pan, A., Shern, C. J., Zou, A., Li, N., Basart, S., Woodside, T., Ng, J., Zhang, H., Emmons, S., and Hendrycks, D. Do the rewards justify the means? measuring trade-offs between rewards and ethical behavior in the machiavelli benchmark, 2023.

Peng, B., Li, C., He, P., Galley, M., and Gao, J. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277*, 2023.

Ravfogel, S., Twiton, M., Goldberg, Y., and Cotterell, R. Linear adversarial concept erasure, 2022.

Schaeffer, R., Miranda, B., and Koyejo, S. Are emergent abilities of large language models a mirage?, 2023.

Shevlane, T., Farquhar, S., Garfinkel, B., Phuong, M., Whittlestone, J., Leung, J., Kokotajlo, D., Marchal, N., Anderljung, M., Kolt, N., Ho, L., Siddarth, D., Avin, S., Hawkins, W., Kim, B., Gabriel, I., Bolina, V., Clark, J., Bengio, Y., Christiano, P., and Dafoe, A. Model evaluation for extreme risks, 2023.

Sun, Z., Shen, Y., Zhou, Q., Zhang, H., Chen, Z., Cox, D., Yang, Y., and Gan, C. Principle-driven self-alignment of language models from scratch with minimal human supervision. *arXiv preprint arXiv:2305.03047*, 2023.

Wang, Y., Kordi, Y., Mishra, S., Liu, A., Smith, N. A., Khashabi, D., and Hajishirzi, H. Self-instruct: Aligning language model with self generated instructions. *arXiv preprint arXiv:2212.10560*, 2022.

Xie, S. M., Pham, H., Dong, X., Du, N., Liu, H., Lu, Y., Liang, P., Le, Q. V., Ma, T., and Yu, A. W. Doremi: Optimizing data mixtures speeds up language model pretraining. *arXiv preprint arXiv:2305.10429*, 2023.

Xu, C., Sun, Q., Zheng, K., Geng, X., Zhao, P., Feng, J., Tao, C., and Jiang, D. Wizardlm: Empowering large language models to follow complex instructions. *arXiv preprint arXiv:2304.12244*, 2023.