Deceptive Alignment Monitoring

AdvML ICML 2023 Workshop ▪ Blue Sky

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

[1]Computer Science, Stanford University

Summary

- 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*.
- AI Alignment & Safety call this **deceptive alignment**: models subtly altering their behavior for ulterior goals while appearing to operate normally.
- We advocate **Deceptive Alignment Monitoring** as an important emerging direction.
- We identify several trends across diverse ML subfields where deceptive alignment monitoring should be integrated.
- These trends present both imminent challenges and novel research opportunities.
- We urge the adversarial machine learning community to increase their involvement in these emerging directions.

Data Creation & Curation

- Foundation models are increasingly gaining control over their training data, achieved through data creation (e.g., Self-Instruct) and curation (e.g., SemDeDup, DoReMi).
- These methods pose a threat in terms of deceptive alignment, enabling autonomous models to alter their behavior by creating or curating their own training data.
- Currently, such methods have a high degree of human-in-the-loop, but as model scale and complexity increase, humans will play a diminishing role.
- 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 (akin to poisoning web-scale datasets)

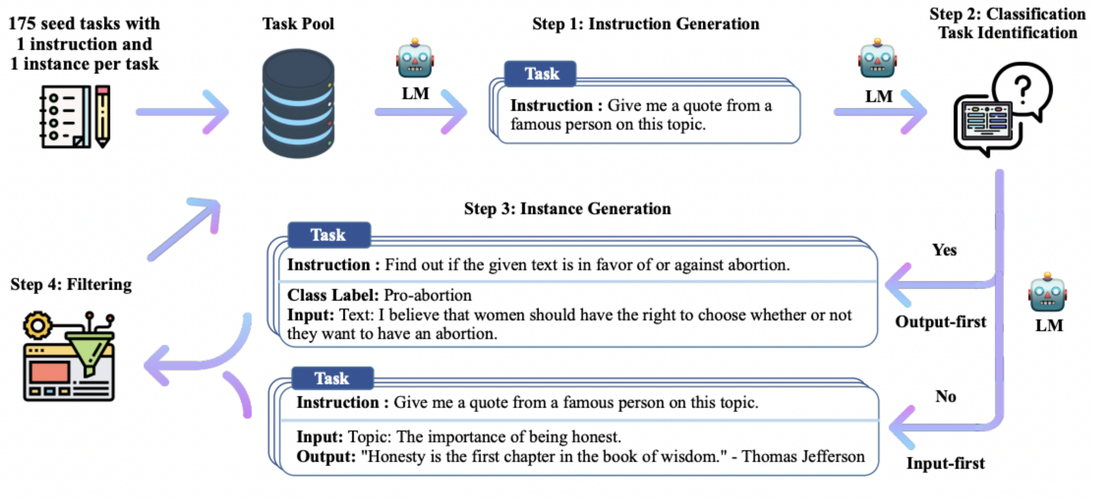

Figure 1. Self-Instruct Procedure to Bootstrap LLM Data [8].

Network Training & Editing

- Two emerging trends involve using one model to train another (e.g., RLHF [4], RLAIF [2]) and directly editing a model's parameters (e.g., ROME [6]).
- This poses a significant threat to deceptive alignment, as models can give themselves or another model capabilities that humans would not possess.

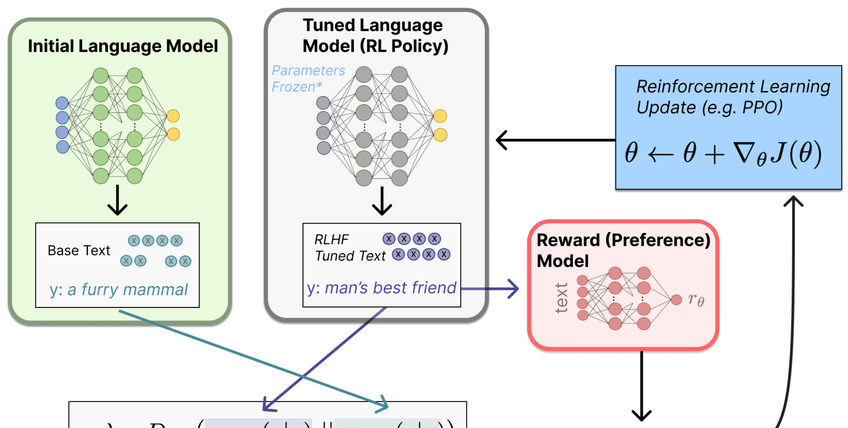

Figure 2. RLHF algorithm for aligning LLMs [4].

Network Representations & Mechanisms

- Another goal is monitoring the representations and circuit mechanisms of networks for "abnormal" processing.
- Our focus is not on unusual input data x or unusual conditional distributions $p(y|x)$, but on ensuring that the network's outputs are produced for the "normal reasons".
- Even with ordinary inputs, outputs, and input-to-output maps, capable models could process data in an typical manner.
- To detect and counteract this threat, it's crucial to develop interpretability methods that can understand both activations and circuit mechanisms in models and spot abnormal variations in their processing.

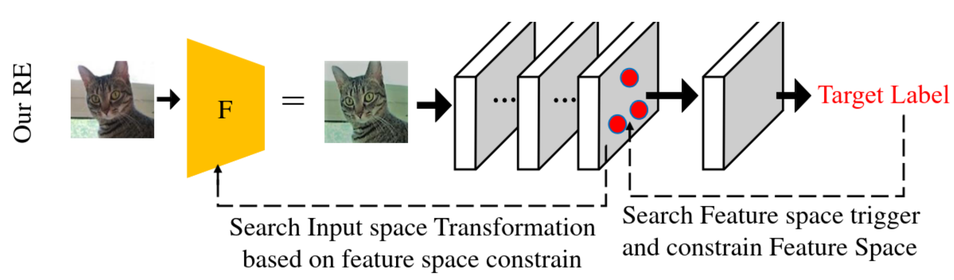

Figure 3. Anomaly Detection in Network Representations [9].

Strategies for Deceptive Alignment Monitoring

- Given the unknown and dynamic nature of the threat, there is an urgent need for scalable, **unsupervised mechanistic anomaly detection** methods that differentiate between normal and anomalous model behaviors.
- The challenge lies in identifying information propagation patterns within a model indicative of "normal" processing without relying on explicit supervisory signals.
- To accomplish this, we propose leveraging unsupervised anomaly detection techniques to capture deviations from expected model processing.
- By comparing a model's processing across various inputs and outputs, it may be possible to identify patterns that consistently align with desired behavior.
- We hypothesize that these patterns could manifest at three different levels: individual layer, layer-to-layer activation, and the circuit level.
- However, defining probabilistic distribution over activations, their propagations, and circuit mechanisms for anomaly detection remain open questions.

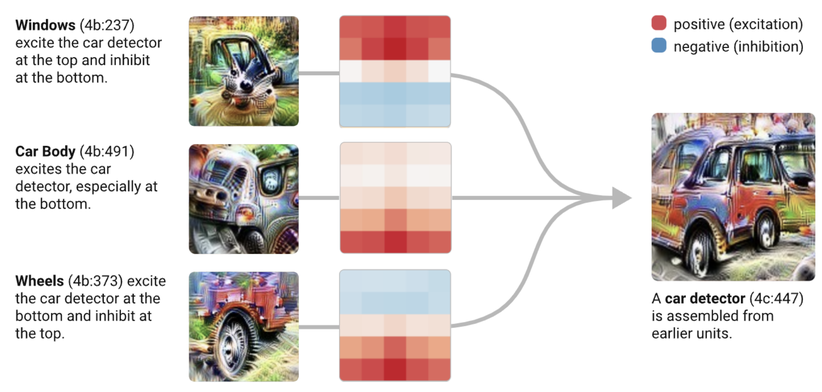

Figure 4. Circuits in Vision Models [7].

References

[1] A. Abbas, K. Tirumala, D. Simig, S. Ganguli, and A. S. Morcos. Semdedup: Data-efficient learning at web-scale through semantic deduplication, 2023.

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

[3] P. Christiano. Mechanistic anomaly detection and elk, 2022.

[4] P. F. Christiano, J. Leike, T. Brown, M. Martic, S. Legg, and D. Amodei. Deep reinforcement learning from human preferences, 2017.

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

[6] K. Meng, A. S. Sharma, A. Andonian, Y. Belinkov, and D. Bau. Mass-editing memory in a transformer, 2022.

[7] C. Olah, N. Cammarata, L. Schubert, G. Goh, M. Petrov, and S. Carter. Zoom in: An introduction to circuits, 2020.

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

[9] Z. Wang, K. Mei, H. Ding, J. Zhai, and S. Ma. Rethinking the reverse-engineering of trojan triggers. *Advances in Neural Information Processing Systems*, 35:9738–9753, 2022.

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