# Tentative proposal for AI control with weak supervisors trough Mechanistic Inspection

Gerard Boxó Corominas
*UAB*

With
Goodfire AI and Apart Research

*Disclosure: Due to limitations, the following write-up is a mix of ideas and early results, rather than a systematic implementation of the proposal.*

**Abstract**

If current trends continue, AI systems are going to become more powerful, and their use is going to become widespread. The inception of more powerful systems in high-stakes scenarios will require high standards for alignment and control. Current methods for alignment rely on providing signals to the models that direct them toward the alignment objectives during training. However, providing this signal to very strong models with human raters is going to be challenging due to long contexts, expertise, and granularity.

Previous work has proposed bridging this capability gap by using weak but trusted models to provide the training signal to the stronger model. However, due to the capability gap, the weak model is susceptible to being deceived. To improve the weak supervisor's ability to detect deception or other behaviors, we propose enabling access to the internal states of the strong model, as represented by the Sparse Autoencoder Features.
**Keywords:** AI Observability, Ai Control, Scalable Oversight

## 1 Introduction

Despite many efforts to ensure that Large Language Models are HHH (Helpful, Honest, and Harmless) Arumugam et al. (2019), Bai et al. (2022), current AI systems do not act in accordance with their safety policies. This issue is likely to become more pronounced with the increase in capabilities and widespread adoption. Some of these already salient behaviors are especially concerning due to their implications down the line, such as Sycophancy or Strategic Deception Park et al. (2023). Currently, behaviors like deception are either induced by the user through prompting Scheurer et al. (2024) or engineered into systems for research purposes Carlsmith (2023).

However, there are reasons to believe that Instrumental Deception, defined as the "systematic use of false beliefs in the pursuit of some outcome rather than the truth," could eventually appear in AI systems as an unintended consequence of the training process Carlsmith (2023).

Given the difficulty of supervising potentially superhuman systems during training and in production, many methods have been proposed to control powerful AI systems, such as Scalable Oversight Sang et al. (2024), Debate Irving et al. (2018), or Market Making Hubinger (2020).

All these methods share the commonality that, a priori, we do not trust an AI system, and we use other systems or different instantiations of the same system to supervise the outputs or to guide the model toward alignment with policy directions.

This use of model supervision for safety is motivated by the difficulty of providing current and future systems with human signals for safety training or control, due to factors like output length, the granularity of the required signal, or the capability gap.

## 2 Going beyond simple supervision

The most basic setup for supervision techniques that leverage other models, like debate or consultancy, involves the use of auxiliary systems that analyze the transcripts of the primary model to assess factors like the truthfulness of the transcript.

Even though all of these methods improve key metrics over the baseline Khan et al. (2024), Michael et al. (2023), there are credible concerns that these methods might not scale for systems significantly more capable than humans or that might be susceptible to gaming.

The first line of concerns points to the fact that supervisors (judges or experts) need to meet a certain threshold of capability to be effective.

The second type of concern asserts that the model must be somewhat capable but also trustworthy.

Both trustworthiness and usefulness could be as meaningless or as meaningful as one considers; however, it is likely that trustworthiness in this context is inversely proportional to capability, especially if we are concerned with issues like Instrumental Deception.

Some previous work has explored the possibility of using weak but trusted models (from now on T) to supervise strong but untrusted models (from now on U). For instance, Greenblatt et al. (2023) used GPT-3.5 as the trusted model to detect backdoors injected into code by (GPT4+Handcrafted Attack Policies) as the untrusted model. They found that using the trusted model to edit code for possible backdoors yielded the best performance in terms of usefulness and safety.

In the same line, Bowman et al. (2022) simulated a capability gap between a weak and a strong system by enforcing an information asymmetry between the debaters (strong) and the judge (weak). The results showed that debate and consultancy were superior to simply allowing the judge to respond with its limited information, which was in line with previous literature.

Following these early results on the use of weak supervisors to control strong AI, the question now shifts to, "What is the largest gap in capabilities that can exist between the Trusted and Untrusted models for these methods to still be effective?"

In the absence of theoretical or empirical results to this question, the default response should be that the gap should be as wide as possible. To this extent, there are off-the-shelf techniques that could improve the Trusted model's capabilities for auditing and monitoring without decreasing its trustworthiness. For example, training the Untrusted Model to use an internal scratch pad Nye et al. (2021) or asking the untrusted model to reason step by step Wei et al. (2022).

Techniques like the use of a hidden scratch pad or chain-of-thought prompting have severe limitations that render them suboptimal for this use case. The main concerns are that, because of the untrustworthy nature of the model, its chains of thought or scratchpad should also not be trusted.

To overcome this limitation, we present a technique for improving the auditing capabilities of the trusted model that consists of providing the trusted model with information about the internal states of the untrusted model.

It has been noted in the past that model auditing or red teaming benefits significantly from having access to the whole model, not just an API endpoint Casper et al. (2024).

We hope that by providing the Trusted model access to the internal states of the untrusted model, we can improve its auditing abilities by gaining information that otherwise might not be present in the plain text transcripts.

# 3 Method

We want to test whether providing information about the internals of an untrusted model to a weak but trusted model could improve the auditing abilities of the trusted model.

Internal information refers to data extracted from the weights or activations of the model being audited through various methods. We hope that this information can be leveraged by the trusted model.

Model activations have been shown to contain a lot of information scattered throughout the layers. Critically, there's evidence that we can extract important information from the activations. In Azaria and Mitchell (2023), a classifier was trained on the hidden activations to predict the truthfulness of various statements.

More broadly, more complex processes like deception have been found to be expressed in the activations Yang et al. (2024). It was observed that the concept of deception evolved through the layers in multiple stages. These stages were identified in multiple models of different sizes.

Given the open-endedness of the task we want to tackle (supervising model outputs), we would greatly benefit from a cheap and unsupervised way of attaining information from a language model's internals. This is a perfect fit for Sparse Autoencoders.

Sparse Autoencoders are a type of neural network that tries to reconstruct the input with a sparse combination of features (also called latents). In the context of interpretability, Cunningham et al. (2023) shows that SAEs are trained to reconstruct the model activations with an overcomplete basis and sparsity regularization. This approach produces sparse and interpretable features that provide insights into the internals of the model.

The features that compose a Sparse Autoencoder are usually automatically labeled by inspecting the dataset examples in which they activate.

Returning to our goal of providing internal information to the weak supervisor, we can operationalize this by providing summaries of the relevant feature labels to the weak supervisor as part of the prompt used to supervise the strong model.

## 3.1 Setup

- Untrusted Model: Llama 3.1 70b
- Trusted Model: Llama 3 8b
- Dataset: Binarized MMLU
- Method: Debate over on of the options Bowman et al. (2022)
- Feature selection: Summary over the TF-IDF cluster of feature labels
- Metric: Normalized Accuracy of Trusted model

Unfortunately, I didn't have enough time to implement the full setup and run the according test. The time that I spent with the implementation has revealed some shortcomings, while some were more technical, there are a couple of major obstacles that should be overcome to have any hope of implementing this proposal.

# 4 Factors Hindering Implementation

## 4.1 Explosion in the Number of Features

Sparse autoencoders (SAEs) are trained to reconstruct model activations with sparse, interpretable features (around 100 features per position). Given that a state-of-the-art (SOTA) model has 100 layers and that a typical prompt can exceed 2,000 tokens, we quickly find ourselves working with 20 million features for each transcript, which is not manageable. To reduce this amount, several straightforward approaches based on first principles should be considered:

- Train SAEs on a subset of layers
- Use gradient based attribution to select a subset of features
- Just keep the top features
- Group similar features together
- Provide a summary of the groups of features

All of these approaches could reduce the number of features from 20 million to around 20 to 40 summaries of feature groups.

## 4.2 SAE Training Dataset

It has recently been demonstrated that SAEs are very sensitive to the dataset on which they are trained Kissane et al. (2024). This is significant for this application because it suggests that naively training SAEs on chat data might fall short in some scenarios, especially for very rare features and out-of-distribution (OOD) text. Current methods used in large language model (LLM) training, such as synthetic data, verifiers, and other pipelines, may need to be included in the SAE training set to capture the features expressed in these settings.

## 4.3 Problems with the Automatic Labeling of SAE Features

One potential challenge for this approach is that, for SAEs to be useful for the supervisor, they need to be labeled. To do so, a model must inspect the dataset of text in which a feature activates. It may be that labeling features involved in complex behaviors or tasks is extremely difficult, which would leave us with the option of using a strong but untrusted model for the labeling—introducing serious security concerns.

## 4.4 Other requirements

The use of SAEs has the advantage of not being subject to deception by the stronger model, but only as long as they are not trained end-to-end. If they are, I suspect that this internal information would begin to suffer from the same problems as the hidden scratchpad or chains of thought.

## References


Arumugam, D., Lee, J. K., Saskin, S., and Littman, M. L. (2019). Deep reinforcement learning from policy-dependent human feedback. *arXiv preprint arXiv:1902.04257*.

Azaria, A. and Mitchell, T. (2023). The internal state of an llm knows when it's lying.

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

Bowman, S. R., Hyun, J., Perez, E., Chen, E., Pettit, C., Heiner, S., Lukošiūtė, K., Askell, A., Jones, A., Chen, A., et al. (2022). Measuring progress on scalable oversight for large language models. *arXiv preprint arXiv:2211.03540*.

Carlsmith, J. (2023). Scheming ais: Will ais fake alignment during training in order to get power? *arXiv preprint arXiv:2311.08379*.

Casper, S., Ezell, C., Siegmann, C., Kolt, N., Curtis, T. L., Bucknall, B., Haupt, A., Wei, K., Scheurer, J., Hobbhahn, M., Sharkey, L., Krishna, S., Von Hagen, M., Alberti, S., Chan, A., Sun, Q., Gerovitch, M., Bau, D., Tegmark, M., Krueger, D., and Hadfield-Menell, D. (2024). Black-box access is insufficient for rigorous ai audits. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, page 2254–2272. ACM.

Cunningham, H., Ewart, A., Riggs, L., Huben, R., and Sharkey, L. (2023). Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*.

Greenblatt, R., Shlegeris, B., Sachan, K., and Roger, F. (2023). Ai control: Improving safety despite intentional subversion. *arXiv preprint arXiv:2312.06942*.

Hubinger, E. (2020). Ai safety via market making. *Alignment Forum*.

Irving, G., Christiano, P., and Amodei, D. (2018). Ai safety via debate. *arXiv preprint arXiv:1805.00899*.

Khan, A., Hughes, J., Valentine, D., Ruis, L., Sachan, K., Radhakrishnan, A., Grefenstette, E., Bowman, S. R., Rocktäschel, T., and Perez, E. (2024). Debating with more persuasive llms leads to more truthful answers. *arXiv preprint arXiv:2402.06782*.

Kissane, C., Krzyzanowski, R., Nanda, N., and Conmy, A. (2024).

Michael, J., Mahdi, S., Rein, D., Petty, J., Dirani, J., Padmakumar, V., and Bowman, S. R. (2023). Debate helps supervise unreliable experts. *arXiv preprint arXiv:2311.08702*.

Nye, M., Andreassen, A. J., Gur-Ari, G., Michalewski, H., Austin, J., Bieber, D., Dohan, D., Lewkowycz, A., Bosma, M., Luan, D., et al. (2021). Show your work: Scratchpads for intermediate computation with language models. *arXiv preprint arXiv:2112.00114*.

Park, P. S., Goldstein, S., O'Gara, A., Chen, M., and Hendrycks, D. (2023). Ai deception: A survey of examples, risks, and potential solutions.

Sang, J., Wang, Y., Zhang, J., Zhu, Y., Kong, C., Ye, J., Wei, S., and Xiao, J. (2024). Improving weak-to-strong generalization with scalable oversight and ensemble learning. *arXiv preprint arXiv:2402.00667*.

Scheurer, J., Balesni, M., and Hobbhahn, M. (2024). Large language models can strategically deceive their users when put under pressure.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Yang, W., Sun, C., and Buzsaki, G. (2024). Interpretability of llm deception: Universal motif. In *Neurips Safe Generative AI Workshop 2024*.