Researchers use SAE latents to steer model behaviors, yet human-designed algorithms such as "analyze relative frequencies of features given certain types of data and outputs, then make a custom hook function to modify them" are unlikely to reach any sort of optimum for steering tasks such as unlearning or helpfulness steering. It should be possible to add components into SAEs that act on the latents and train them on a task using gradient descent. These trained components could learn optimal values and algorithms, and if we chose their structure carefully, they can retain the interpretable properties of the SAE latent itself. I call these fine-tuning methods Interpretable Sparse Autoencoder Representation Fine Tuning or "ISaeRFT".

This research direction would contribute towards answering "How can intentional or accidental misuse of advanced AI systems be prevented?" by answering the narrower subquestion "Can the fine-tuning process of model development be made interpretable?" This is important for reducing global catastrophic risk because many important aspects of LLM behavior are introduced during fine-tuning: its personality, preferences, ideology, how it responds to certain stimuli, refusal behavior, reasoning behavior, and more are all mostly undetermined after pretraining but become important aspects of the model after fine tuning. AI developers could use interpretable fine-tuning as part of their safety standards for auditing their data and fine tuning process. Additionally, stakeholders who care about the interpretability of their model and are willing to sacrifice performance may opt for a model which has interpretable fine tuning components.

This builds off of work such as Representation Engineering: A Top-Down Approach to AI Transparency and since it involves small modifications to semantic meaning, it builds off of Locating Semantic Understanding: A Study of Semantically Similar Sentences in Large Language Models.

This is my current side project, and I have made significant progress, but there is still a lot of work to be done.
Github: https://github.com/AMindToThink/interpretable-fine-tuning
Preliminary results:
https://docs.google.com/document/d/1uttDTD16hWF8UriLnMjVOkA-8Lu8-QaAey7560gEVYI/edit?tab=t.0

Main doubts:
It is likely that changing one part of the model modifies the semantic meaning of other parts of the model. The more changes made, the more significant this change would be, and then the preexisting SAE latent labels would lose their validity. This might be a significant challenge, but if necessary a regularization technique could keep the learned components from changing the semantic meanings of the latent by too much.

It remains to be seen whether a small number of interpretable components can be expressive enough to perform meaningfully well on fine-tuning tasks.

Weekly breakdown:
Week 1: Train a SAE steering vector on Gemma-2-2b and Pythia70m using SFT using a helpful-harmless dataset (AMindToThink/moss-002-sft-data-instruction-output). Apply LoRA finetuning to the models as a baseline for comparing performance.

Week 2: Apply Attribution Patching to the learned steering vectors to find which indices were most impacted, and use their Neuronpedia labels to make a human-readable description of what changed in the model during fine tuning.

Week 3: Make ISaeRFT compatible with TRL, especially DPOTrainer. Do training runs and interpret the changes.

Week 4: Try other interpretable components instead of steering vectors, especially 1 hidden layer residual neural networks.

Week 5: Try applying ISaeRFT to other finetuning tasks — summarization, translation, reasoning.

Week 6: Apply interpretable fine tuning components to the attention SAEs.

Week 7: Do a thorough hyperparameter search to find which layers, components, learning rates, etc work best for ISaeRFT finetuning.

Week 8: Start the paper and do final experiments.

Week 9: Finish the paper.

Week 10: Create a blog post, write documentation, make an example Google Colab notebook, and make my GitHub repository usable by other researchers.