# Interpretability in Artificial Intelligence

*SOUL course proposal, Milton Lin*

## Course Description

As artificial intelligence models like chatGPT become increasingly capable and ubiquitous, the need to understand their inner workings intensifies. Imagine an autonomous vehicle taking an unexpected turn or a medical AI diagnosing a life-altering condition; the importance cannot be overstated. Despite their widespread applications, our grasp of these models remains alarmingly limited. This course is designed to survey this gap. A special emphasis is placed on *mechanistic interpretability*, a subfield that rigorously investigates AI networks at the neuronal level.

The course will engage students through rigorous reading assignments, interactive discussions, and a hands-on project, providing multiple avenues for grasping the complexities and nuances of interpretability in AI. Papers from researchers/groups in this area, such as AnthropicAI, OpenAI, and MIT's Tegmark Group, form the cornerstone of our course.

## Course Topics

Various research papers and articles to be distributed during the course.

1. Introduction to Interpretability in AI [ 1 week] and basics of transformed language models. [ 1 Week]

2. Mechanistic Interpretability [ 4 weeks], example papers include, [2], and a good collection of articles is collected here.

3. Concept-based Interpretability [ 1 week ] example papers include, [1].

## Required Background

Students are expected to have a basic understanding of calculus, linear algebra, probability theory, and coding. A reading list will be provided to be completed before the commencement of the course.

## Assessment

Student assessment will be based on both weekly reading assignments and a course-long project. Students can opt for either a coding project or a written project, aimed at deeply exploring a sub-field of interpretability. The grade distribution will be as follows: Weekly Readings: 60%, Course-long Project: 40%.

## References

[1] Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt, *Discovering latent knowledge in language models without supervision*, 2022.

[2] Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter, *Zoom in: An introduction to circuits*, Distill (2020). https://distill.pub/2020/circuits/zoom-in.