

CSUREMM 2026 Statement of Interest

Peyton Jackson

March 2026

My name is Peyton Jackson and I study Applied Mathematics at SEAS. Outside of my classes, I am also interested in AI safety – I am involved with CUAIAAC (Columbia University Artificial Intelligence Alignment Club), where I pursue technical work in alignment (as well as humanities and policy angles). I have also participated in math modeling opportunities at Columbia (the math modeling workshop and ICM/MCM contest), and greatly enjoyed applying my mathematical knowledge to real problems in a creative way.

I find myself most attracted to mathematical modeling by the idea of connecting abstraction and reality. There is a constant push and pull between building a clear and usable model and the often unbounded complexity of real world behavior. I have so far greatly enjoyed the process of considering the relative impact of environmental parameters, making relevant modeling choices, and formalizing the description with mathematics. Mathematical modeling is like a combination of art (which I identify with as a musician) and rigorous logic (which speaks to me as a mathematics student).

My academic background is strongly based around pure mathematics, particularly in high school where I took several post-calculus math classes, covering topics such as abstract algebra, algebraic combinatorics, number theory, and complex analysis, and more. I chose to study applied mathematics at SEAS rather than continue my focus on pure mathematics so that I could be closer to real world problems and use cases – I want to use mathematical thinking to address tangible issues, and math modeling projects are perhaps the best way of doing this.

During my coursework, both here at Columbia and before attending, as well as in my own projects involving stochastic modeling and simulation, I have gained experience in translating concrete conditions into mathematics. I particularly enjoy deriving conclusions from structural mathematical constraints, such as those reflected in bounds or approximations, in areas like computer science and physics.

My interest in AI safety is another avenue I've found for appreciating the power of mathematical descriptions. Alignment, particularly on the empirical side (as LLMs' black box nature makes effective first-principles work difficult), often becomes modeling problems that require formalization of various concepts like incentives and feedback. In the case of alignment particularly, mathematical modeling provides a means to formalize what would otherwise remain philosophical debates.

If given the opportunity to pursue a project at CSUREMM, I would love to get to work intensively on a math modeling project that is deeply interesting to me for a longer period of time than I have before. While I enjoyed the ICM/MCM experience, for example, I struggled with the short turnaround time, and would have preferred to work on a problem of my own design – both downsides that I would not have at CSUREMM.

Working at CSUREMM would allow me to develop a more mature approach to mathematical modeling than short term experiences have so far allowed. As my academic interests increasingly center on using structural mathematics to understand complex systems, I see CSUREMM as a natural extension of that trajectory. I would contribute my creativity, analytical ability, and openness to alternative perspectives as I build my idea into a rigorous and coherent model.

CSUREMM 2026 Proposal

Peyton Jackson

March 2026

The problem I would like to explore is reward hacking in preference optimization in the field of AI safety. I would use an approach that compares “proxy” versus “true” mismatch across training to model how optimization strength can cause a reward function to diverge from the objective it was intended to represent. In particular, I am interested in modeling reward hacking as a measurable breakdown in correlation between a proxy reward and a “true” evaluation metric as a policy is optimized more aggressively, motivated by existing research detailed below. This problem connects some of my long-standing interests in AI safety with math modeling in the context of unexpected behavior arising from mathematical structural constraints.

Current alignment work on deployed models is done through optimization of proxy objectives, such as human preference in RLHF (Reinforcement Learning through Human Feedback) style learning. A key issue that can arise is that the proxy can fail to completely and precisely indicate the objective, leading to reward hacking – when the system scores well under the reward model while behaving undesirably under the actual objective (for example, sycophancy in RLHF).

This problem is particularly relevant as AI systems are deployed in increasingly more use cases; models failing to perform outside of their test environment can lead to an abundance of issues, from poor efficiency outcomes to economic loss. Modeling how proxy-true mismatch arises could allow for evaluation of prevention measures that clarify and optimize actual performance.

In particular, two questions that I would like to explore are:

1. Can we characterize measurable signals for reward hacking – for example, a decay in proxy–true alignment as optimization strength increases – under different levels of training data diversity and annotation noise?
2. Can we utilize the model to compare intervention mechanisms that preserve proxy-true alignment under limited training data?

Recent work has used the proxy-true idea to formalize reward hacking as a breakdown in alignment under “optimization-induced distribution shift” (heavy training unfavorably changes outcome likelihoods). In particular, the paper Correlated Proxies (Laidlaw et al., 2025)[1] defines reward hacking in terms of the correlation between a learned proxy reward and a true objective. This motivates its use as a measurable signal to be analyzed under increasing optimization strength in the model.

Additionally, the paper Sail into the Headwind (Rashidinejad & Tian, 2025) [2] analyzes how limited dataset coverage and statistical noise contribute to proxy failure in offline preference optimization, supporting the idea of modeling data diversity and label noise as explicit variables leading to reward hacking. Rather than aiming to propose new strategies for mitigating reward hacking, my project seeks to build a mathematical model that treats optimization strength, training data diversity, and label noise as parameters, allowing reward hacking to be analyzed as a robustness problem.

References

- [1] Cassidy Laidlaw, Sumeet Singhal, and Anca Dragan. Correlated proxies: A new definition and improved mitigation for reward hacking. In *International Conference on Learning Representations (ICLR)*, 2025.
- [2] Paria Rashidinejad and Yifan Tian. Sail into the headwind: Alignment via robust rewards and dynamic labels against reward hacking. In *International Conference on Learning Representations (ICLR)*, 2025.