Philipp Alexander Kreer

Research Statement: Fundamental Physics & Science of Deep Learning

Motivation

My career has centered around the fundamental question: *What is mass?*, explored through various lenses: *in the Standard Model of particle physics, in general relativity, and in the learning processes of artificial intelligences*. Though seemingly distinct, my interdisciplinary research in theoretical physics and the science of deep learning has revealed the deeper connections between these perspectives.

My Contribution to Fundamental Physics

The Standard Model of particle physics describes the interactions of elementary particles at the smallest scales. In this framework, the Higgs boson gives mass to fundamental particles through spontaneous symmetry breaking. While the 2012 discovery of the Higgs was a milestone, key questions about the Higgs boson's properties remain. For instance, its interaction strength with the top quark — the heaviest Standard Model particle is only inaccurately measured. Deviations from the predicted coupling could reveal new physics and the mystery of dark matter. Even though dark matter constitutes approximately 27% of the universe, essentially nothing is known about it.

The measurement of the coupling of the Higgs particle to top quarks relies on precise Standard Model predictions. In my paper *One loop QCD corrections to gg $\to t\bar{t}H$ at $\mathcal{O}\left(\epsilon^2\right)$*, I computed precision corrections to so-called scattering amplitudes which are essential building blocks for high-precision predictions at the Large Hadron Collider. A key challenge was optimizing the computational implementation. While the original method overwhelmed large clusters, my approach reduced the runtime to minutes on a personal laptop.

In my paper *H graph with unequal masses in quantum field theory*, I adapted these techniques to study the motion of black holes. This cross-disciplinary link between general relativity and the Standard Model highlighted the nuanced role of mass and demonstrated the power of quantum field theory, the mathematical foundation of the Standard Model.

The rapid progress of AI has opened new paths to understanding mass. For example, neural networks have uncovered the constituents of protons, which, along with neutrons, form atomic nuclei. This analysis confirmed the existence of so-called intrinsic charm quarks in protons, resolving a half-century debate. However, the behavior of cutting-edge AI models remains poorly understood, making it crucial to grasp their underlying dynamics to drive future advancements.

My Contribution to the Science of Deep Learning

Misaligned AI behaviors during deployment limit their real-world applicability. Misalignment can originate from subtleties, such as biased or insufficient training data, distributional shifts, insufficient capability evaluations, etc. During training, phenomena like phase transitions and grokking cause drastic shifts in model performance. Understanding these error sources is key to designing reliable and transparent AI systems for real-world use.

The science of deep learning aims to address these challenges with a comprehensive framework. The interplay between theoretical physics and the science of deep learning has led to innovations like physics-informed neural networks, which incorporate physical laws to enhance robustness. Equivariant networks leverage geometric properties to improve training, generalization, and robustness. These models solve real-world problems, from predicting WIFI strengths to classifying quantum particles. Beyond novel architectures, theoretical physics also provides unique insights into the inner workings of AI systems.

Successful AI systems are often wider than they are deep. So-called effective field theory descriptions leverage this property to predict large-scale behaviors of advanced AI models. Critical hyperparameter settings have been derived using tools from theoretical particle physics and general relativity. These critical parameters facilitate learning in neural networks and transformer models, significantly reducing the need for extensive hyperparameter searches.

The constituents of AI models, such as neurons or attention heads, behave similarly to interacting particles in complex physical systems. Singular learning theory formalizes this analogy by translating the mathematical foundation of theoretical physics to AI models. This theory successfully explained phase transitions in Anthropic's toy models. Loosely speaking, phase transitions in AI models equal phase transitions of materials, like water freezing to ice. Despite its success, singular learning theory remains limited to toy models, as computations are currently intractable for frontier AI models.

Building on singular learning theory, I developed a novel technique to detect sandbagging — strategic underperformance — in AI systems. Sandbagging has been observed in modern large language models, posing a significant security risk as it could allow evasion of governmental regulations. In my paper, *Sandbagging Detection through Model Impairment*, presented at NeurIPS, I demonstrated that injecting noise into a model's parameters can unexpectedly improve its performance, a counterintuitive indicator of sandbagging. This research earned me the Apart Research Fellowship and highlights the advantages of physics-informed approaches to AI safety.

Future Work

Going forward, I will leverage my expertise in theoretical particle physics and deep learning to design aligned AI systems. My focus will be on addressing distributional shifts — situations where training data fails to match deployment scenarios — which frequently lead to uncontrolled behaviors. Drawing on my work detecting sandbagging in large language models, I will analyze these shifts using singular learning theory and tackle computational bottlenecks with effective field theory descriptions of AI models. I plan to test these insights on particle physics distributions, which are ideal due to their analytic nature, extensive experimental validation, and large datasets. Therefore, they uniquely combine theoretical advantages with real-world setups. In addition, my previous research on scattering amplitudes allows me to carefully study the critical effects of, for example, singularities.

My unique background positions me to contribute to the design of aligned AI systems and the progress of modern particle physics by bridging abstract theoretical frameworks with practical, empirical machine-learning experiments.