Open ended project

What are the major challenges in the field of AI?

Avoiding Negative Side Effects: How can we ensure that an AI system will not disturb its environment in negative ways while pursuing its goals, e.g. a cleaning robot knocking over a vase because it can clean faster by doing so?

Avoiding Reward Hacking: How can we avoid gaming of the reward function? For example, we don't want this cleaning robot simply covering over messes with materials it can't see through.

Scalable Oversight: How can we efficiently ensure that a given AI system respects aspects of the objective that are too expensive to be frequently evaluated during training? For example, if an AI system gets human feedback as it performs a task, it needs to use that feedback efficiently because asking too often would be annoying.

Safe Exploration: How do we ensure that an AI system doesn't make exploratory moves with very negative repercussions? For example, maybe a cleaning robot should experiment with mopping strategies, but clearly it shouldn't try putting a wet mop in an electrical outlet.

Robustness to Distributional Shift: How do we ensure that an AI system recognizes, and behaves robustly, when it's in an environment very different from its training environment? For example, heuristics learned for a factory work floor may not be safe enough for an office.