Abstract

Autonomous goal-directed systems may behave flexibly with minimal human involvement. Unfortunately, such systems could also be dangerous if pursuing an incorrect or incomplete goal.

Meaningful human control can ensure that each decision ultimately reflects the desires of a human operator, with AI systems merely providing capabilities and advice. Unfortunately, as AI becomes more capable such control becomes increasingly limiting and expensive.

I propose to study an intermediate approach, where a system's behavior is shaped by what a human operator *would have done* if they had been involved, rather than either requiring actual involvement or pursuing a goal without any oversight.

This approach may be able to combine the safety of human control with the efficiency of autonomous operation. But capturing either of these benefits requires confronting new challenges: to be safe, we must ensure that our AI systems do not cause harm by incorrectly predicting the human operator; to be efficient and flexible, we must enable the human operator to provide meaningful oversight in domains that are too complex for them to reason about unaided.

This project will study both of these problems, with the goal of designing concrete mechanisms that can realize the promise of this approach.