Project Abstract

AI research has focused on improving the decision-making capabilities of computers, i.e., the ability to select high-quality actions in pursuit of a given objective. When the objective is aligned with the values of the human race, this can lead to tremendous benefits. When the objective is misaligned, improving the AI system's decision-making may lead to worse outcomes for the human race. The objectives of the proposed research are (1) to create a mathematical framework in which fundamental questions of value alignment can be investigated; (2) to develop and experiment with methods for aligning the values of a machine (whether explicitly or implicitly represented) with those of humans; (3) to understand the relationships among the degree of value alignment, the decision-making capability of the machine, and the potential loss to the human; and (4) to understand in particular the implications of the computational limitations of humans and machines for value alignment. The core of our technical approach will be a cooperative, game-theoretic extension of *inverse reinforcement learning*, allowing for the different action spaces of humans and machines and the varying motivations of humans; the concepts of *rational metareasoning* and *bounded optimality* will inform our investigation of the effects of computational limitations.