Abstract

This research aims to investigate the Artificial Intelligence (AI) value alignment problem, which refers to the challenge in developing a safe and reliable AI that can achieve our goals and adhere to our values as we intend it to do. A misaligned AI, especially one which transcends all domains of cognitive abilities and has acquired vast computational powers, will be nearly impossible to manage and it will threaten our security. Research addressing this problem is now focused on understanding how to develop AI that can reliably infer our values from our preferences. Thus, preferences are the primary conceptual unit of analysis to the AI value alignment problem. This paper investigates our preferences and seeks to shed light on the issue of obtaining a formal truth that is fundamentally constitutive of our preferences, for the aim of using said formal truth to create a value aligned AI. To do this, this paper gathers data from economics, biological evolution, and neurocognitive studies to bridge the current gaps on the conceptual problem of preferences. The paper concludes with presenting a new kind of security dilemma which stems from the notion of combining a general theoretical framework that fully captures our preferences with the crucial element of uncertainty in AI, effectively showcasing how the two ideas contradict one another. The paper suggests that future research in AI must focus on developing a framework which can reconcile formal truth about our preferences with the key element of uncertainty. This finding serves to show that the problem of AI control is potentially much more difficult to solve than we had initially anticipated, which implies an even greater urgency to support AI value alignment research and encourage more people to work on the problem.