

Master IAMI A.A. 2023-2024

Titolo elaborato finale: Niente caffè da morto: il ruolo della convergenza strumentale nella ricerca di sistemi d'intelligenza artificiale

Candidato: Emanuele Mario Zavanella

Relatore: Prof.ssa Nicoletta Cusano

Abstract

Questo lavoro riassume in maniera non tecnica la tesi della convergenza strumentale e ne delinea l'impianto teorico nella sua evoluzione degli ultimi quindici anni. L'ipotesi della convergenza strumentale sostiene che agenti sufficientemente intelligenti, dato un certo scopo fisso o mutevole, tendano a sviluppare degli obiettivi secondari strumentali funzionali al conseguimento di tale scopo. Per esempio, se immaginiamo un uomo il cui unico scopo fosse quello di bere caffè, per conseguire tale obiettivo dovrebbe prima di tutto massimizzare la probabilità di essere vivo. In quest'ottica, il mantenersi in vita diventa un obiettivo secondario ma imprescindibile. È plausibile che agenti non-umani con obiettivi apparentemente innocui possano anch'essi sviluppare spontaneamente obiettivi secondari che convergono ad esempio all'autoconservazione e a massimizzare il numero di azioni disponibili (potere). Se a un IA è dato il compito di risolvere un'equazione, esso potrebbe ritenere mezzo adeguato l'assicurarsi di non poter essere spento, o imparare a sfruttare risorse naturali per aumentare la propria capacità computazionale.