عنوان مقاله:

Kantian Fallibilist Ethics for AI alignment

محل انتشار:

فصلنامه پژوهش های فلسفی, دوره 18, شماره 47 (سال: 1403)

تعداد صفحات اصل مقاله: 16

نویسنده:

Vadim Chaly - Lomonosov Moscow State University, Immanuel Kant Baltic Federal University, Russia

خلاصه مقاله:

The problem of AI alignment has parallels in Kantian ethics and can benefit from its concepts and arguments. The Kantian framework allows us to better answer the question of what exactly AI is being aligned to, what are the problems of alignment of rational agents in general, and what are the prospects for achieving a state of alignment. Having described the state of discussions about alignment in AI, I will reformulate them in Kantian terms. Thus, the process of alignment is captured by the concept of enlightenment, and for the final state of alignment in Kant's lexicon there is the concept of the "kingdom of ends." I will argue that the discourse of alignment and the Kantian ethical program ۱) are devoted to the same general end of harmonizing the thinking and acting of rational agents, ۲) encounter similar difficulties, well known in the Kantian discussions with its comparatively longer history, and ۳) for a number of reasons lying on the side of humanity, do not have and, despite the hopes and attitudes of some participants in the AI discussions, will not have a theoretically rigorous, harmonious and practically implementable, conflict-free solution - alignment will remain a regulative idea in the Kantian sense, but will not become a reality.

کلمات کلیدی:

AI alignment, moral deliberation, moral fallibilism specification gaming, kingdom of ends, categorical imperative, misgeneralization