

Stefanie Dach

"Cooperation, Morality, and Artificial Agents"

17th of March (15.50-17.20 in 225V)

Abstract:

Humans cooperate, and humans behave morally. Cooperation and moral behavior are not identical, but both are typically characterized by a focus directed at others. That raises the question of whether these two phenomena stand in a dependence relation. I introduce both directions of such a dependence relation, i.e., cooperation as based on moral commitment and moral commitment as based on cooperation. I develop a more detailed account of the latter in terms of the game-theoretic approach of team reasoning and Wilfrid Sellars's approach to morality and moral concepts. This account chimes well with some empirical approaches to explaining the origin of human morality, particularly with Michael Tomasello's program, in a way that, as I show, addresses a small but significant gap in it. I explore some consequences of this understanding of human morality as "cooperation-plus" for whether artificial agents could be moral agents and for the alignment problem.