Social Choice for AI Alignment

Vincent Conitzer
Carnegie Mellon University

Foundation models such as GPT-4 are fine-tuned to avoid unsafe or otherwise problematic behavior, so that they refuse to comply with requests for help with committing crimes, refuse to produce racist text, etc. One approach to this fine-tuning is to let humans express which of multiple outputs they prefer and to learn from that, an approach commonly referred to as Reinforcement Learning from Human Feedback (RLHF). Another approach is Constitutional AI, in which the only input from humans is a list of high-level principles. But which humans get to provide the feedback or the principles, and how are they weighed against each other when they conflict? This is a natural question for the field of social choice that I will discuss in this talk, drawing on a workshop last month on Social Choice for AI Ethics and Safety (organized together with Jobst Heitzig and Wesley Holliday; please let us know if you are interested in the report).