Some things in the world behave like “agents”: their actions can be understood as directed toward achieving particular goals.

Agent foundations is a research program which tries to understand the nature of agents and their properties, often in a mathematically precise way. It asks questions such as “What are the necessary components of an agent?”, “Can we predict which systems will be agents, and how they will behave?”, and “Which agent designs are tolerant of human error?” There are many open problems in agent foundations research.

Some frameworks which various research groups are developing in order to study agent foundations are embedded agency, Infra-Bayesianism, and shard theory, among others.

Researchers disagree about which features of a system are essential for “agency". Some researchers have a fairly narrow view of what qualifies as an agent, requiring a long list of abilities. Other researchers (including the originators of the agent foundations research program) consider the term “agent” to be ambiguous, and include any system that optimizes towards a "goal” (in a broad sense) in their research on agent foundations.

Related

- What is an agent?
- What are "true names" in the context of AI alignment?
- Why is agent foundations research important for aligning superintelligent syst…
- What is the relationship between goals, intelligence, agency and optimization?
- Why is optimization the thing people worry about?
- What is "coherent extrapolated volition (CEV)"?
- What are the different versions of decision theory?

Scratchpad

Steven rebullet attempt 05/24

- Agent foundations is an AI alignment research agenda
- It’s roughly about figuring out theoretical foundations for intelligent agents
- It’s what MIRI was focused on until recently
- It includes areas such as embedded cognition, decision theory, logical uncertainty, and Vingean uncertainty
- These areas are also sometimes grouped under “highly reliable agent design”
- Agent foundations is distinct from other approaches: prosaic alignment, …
- The logic behind studying agent foundations is (... deconfusion, being able to be very confident in the reliability of systems through understanding their theoretical foundations, https://arxiv.org/pdf/2201.02950 , example of chess in comments, thing about finding a coherent target …)