Mechanism design for AI alignment

Abhimanyu Pallavi Sudhir – abhimanyu.io

market fundamentalism as a philosophical position

tired: logical positivism answers all philosophical questions (but uncertainty!)

wired: Bayes and utility functions answer all philosophical questions (but logical uncertainty!)

inspired: markets and capitalism answer all philosophical questions

—basically all philosophical questions boil down to bounded rationality, e.g.

- verification vs falsification
- if a tree falls down
- ethics
- decision theory
- why trust logic?

bounded agents are like markets

- beliefs, decisions
- incomplete beliefs
- inconsistency = arbitrage
- bounded rationality = algorithmic EMH
- computational cost = transaction costs
- learning = budgets/capitalism

elementary case: prediction markets

"Market-maker" generalizes "Bayesian prior". Payment per bit of information on question.

Beyond finite events – three ways:

- perpetual options – very nice, inf-sup
- dividends – probably the right way

Better than logical induction, because you don't have to trust math, just rich people.

related work

hints at the idea

- Garrabrant+ (2016), Logical Induction (but I don't like to trust theories)
- Oesterheld+ (2021), A theory of bounded inductive rationality (but … IDK, something seems missing)
- John Wentworth's sub-agents (but not program markets!)
- Gwern (2018), Evolution as a backstop for RL

mechanism design reference

- Hanson (2003), LMSR
- Conitzer (2012) […] and co-operative game theory

decisions and latent space

- Chen+ (2011), Decision markets with good incentives
- tailcalled on LessWrong (2023), Latent variables for prediction markets

research agenda

"multi-layer" program markets

- derived demand
- credit assignment
- automated mechanism design
- "optimizing over transaction costs"
- latent space discovery

alignment

- just set an objective function for the market, right?
- no: mis-alignment is like institutional failure*;* spying as generalized interpretability

affiliation: University of Warwick, supervisor: Long Tran-Thanh