Recent Work on Agent Foundations for Robust and Beneficial AI

Andrew Critch

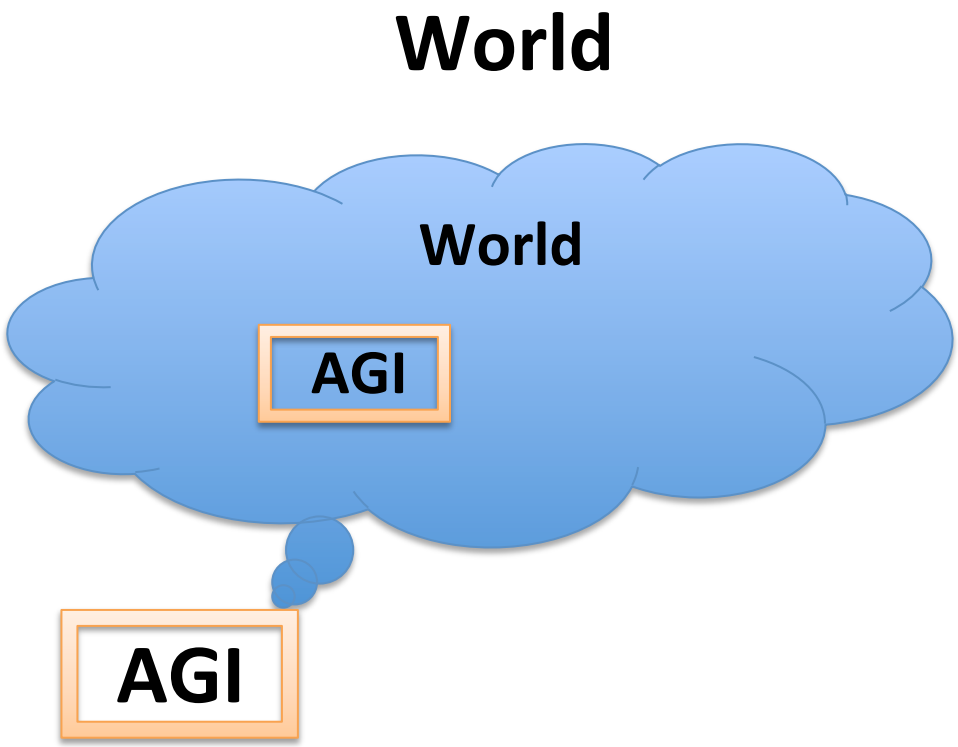
critch@intelligence.org

Machine Intelligence Research Institute
http://intelligence.org/

UC Berkeley, Center for Human Compatible AI
http://humancompatible.ai/

"Aligning Superintelligence with Human Interests", PI: Benya Fallenstein

Assumption: Suppose that at some point in the future, AGI will be thinking about AGI:



Such a system is implicitly facing myriad possibilities with which humans have limited experience, like:

- **Perfectly copying or upgrading** itself, directly or indirectly;
- **Testing or studying** its own source code or that of other AGIs;
- **Finding itself** inside a testing environment, or a hacker's computer…

One might think good behavior in these scenarios can could be trained by example or conversation,

but for any AI system that will face large context changes between testing and deployment, we have strong reasons to believe that it's much easier to align the system if you have clear mathematical models of its reasoning process.

MIRI's Agent Foundations Agenda, in 20 seconds:

… is to develop **fundamental mathematical tools** that help people specify what it means for an AI that can **reason about itself and other AIs** to **reason well** and **act beneficially.**

… the way a POMDP is a clear mathematical specification of what most RL algorithms are meant to do, which uses basic probability theory and linear algebra as tools.

Methodology: pull on loose threads in the edge cases of our understanding, like in physics.

Over the past two years, highlights from MIRI's Agent Foundations work include:

- Reflective Oracles:
  - "a foundation for classical game theory", Fallenstein, Taylor, and Christiano (2015)
  - "a formal solution to the grain of truth problem", Leike, Taylor, and Fallenstein (2016)
- Robust Cooperation of Bounded Agents, – (2016)
- **Logical Induction,** Garrabrant, Benson-Tilsen, –, Soares, Taylor (2016)

Credences should change with time spent thinking / computing:

	1 min	1 day	∞
#1. $P(D_{10} = 7)$	10%	10%	10%
#2. $P(D_{10} = 7 \mid \text{snapshot})$	10% →	15% →	16%
#3. $P(10^{th}\text{ digit of }\sqrt{(10)} = 7)$	10%	1%	0%

Probability theory gives rules for how probabilities should relate to each other and change with new observations, *assuming logical omniscience…*

…but what rules should credences follow over time, as computation is carried out on observations that have already been made?

snapshot for #2:

Also, 50% would be a worse answer to start with here... can we make a principled theory from which this claim would follow?

Goal: call the purple processes "**logical induction**" and figure out how it should work.

Logical Induction

"Logical Induction" (2016) presents

1. a **criterion** for assigning probabilities to statements about deterministic algorithms (and in fact arbitrary mathematical statements), with a large number of desirable properties, and
2. an **algorithm** that provably satisfies the criterion.

Its properties, all consequences of the **criterion**, include:

- Being uncomputably faster than theorem-provers at assigning high confidence to valid patterns;
- Calibration;
- Coherence;
- Introspection;
- Self-trust;
- Future-self-trust…

Formalizing logical induction

PowerPoint → Beamer

Because sometimes, certain theoretical foundations appear to be missing:

Vague desideratum	Clear theoretical specification	Basic concepts needed
"Handling uncertainty well"	Bayesian updating, Solomonoff induction,…	probability theory
"Adapting to an environment well"	POMDPs, AIXI, …	probability theory, linear algebra
"Algorithms reasoning about algorithms well"	**Garrabrant induction criterion**	**"logical uncertainty theory", …**

Formalizing logical induction

Beamer → PowerPoint

The current state of logical uncertainty theory

Domain of Study	Agent Concept	Minimalistic Sufficient Conditions	Desirability Arguments	Feasibility
rational choice theory / economics	VNM utility maximizer	VNM axioms	Dutch book arguments, compelling axioms, …	AIXI, POMDP solvers, …
probability theory	Bayesian updater	axioms of probability theory	Dutch book arguments, compelling axioms, …	Solomonoff induction
logical uncertainty theory	**Garrabrant inductor**	???	**Dutch book arguments, historical desiderata, …**	**LIA2016**

recent progress

What have we learned so far?

The following are more feasible than one might think:

- **Inexploitability.** An algorithm can satisfy a fairly arbitrary set of inexploitability conditions using Brouwer's FPT.

- **Self-trust.** Introspection and self-trust need not lead to mathematical paradoxes.

- **Outpacing deduction.** Inductive learning can in principle outpace deduction, by an uncomputably large margin on efficiently computable questions.

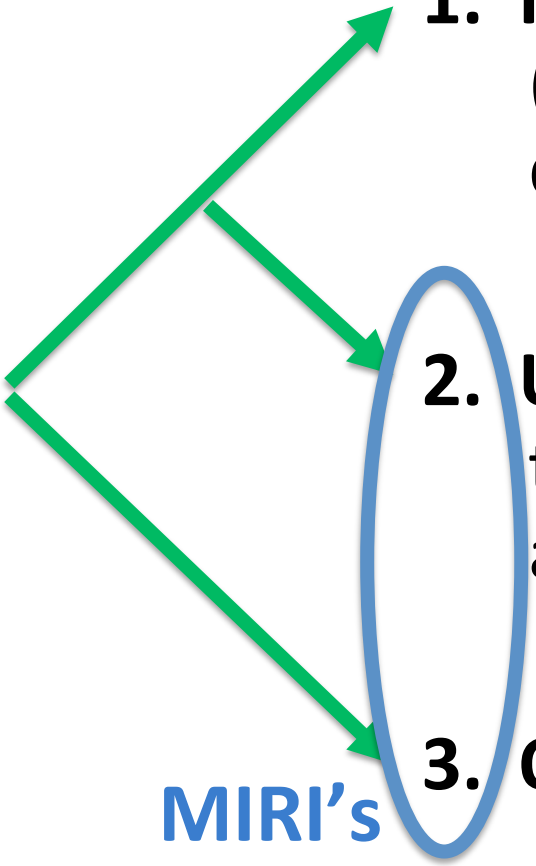
What have we learned so far?

The following are less "required" than one might think for a rational gambler to avoid exploitation:

- **Calibration.** So far it looks like one need only be calibrated about sequences of logical bets that are settled sufficiently quickly (this is being actively researched).

- **Hard-coded belief coherence.** A powerful bet-balancing procedure can and must learn to "mimic" deductive rules used to settles its bets.

Paths forward



1. **Improving** logical inductor theory (Minimalistic conditions? Mutual dominance? Other open questions...)
2. **Using** Garrabrant inductors / LIA2016 to ask new questions about AI alignment
3. **Other approaches** to AI alignment*

MIRI's focus

* Must eventually address logical uncertainty implicitly or explicitly, so expect some convergence.

How will logical induction be applicable?

Conceptual tools for reasoning about **incentives, competition, and goal pursuit** are under-developed for computationally bounded agents. They presume agents are logically omniscient, because we already had good theoretical models for developing them that way:

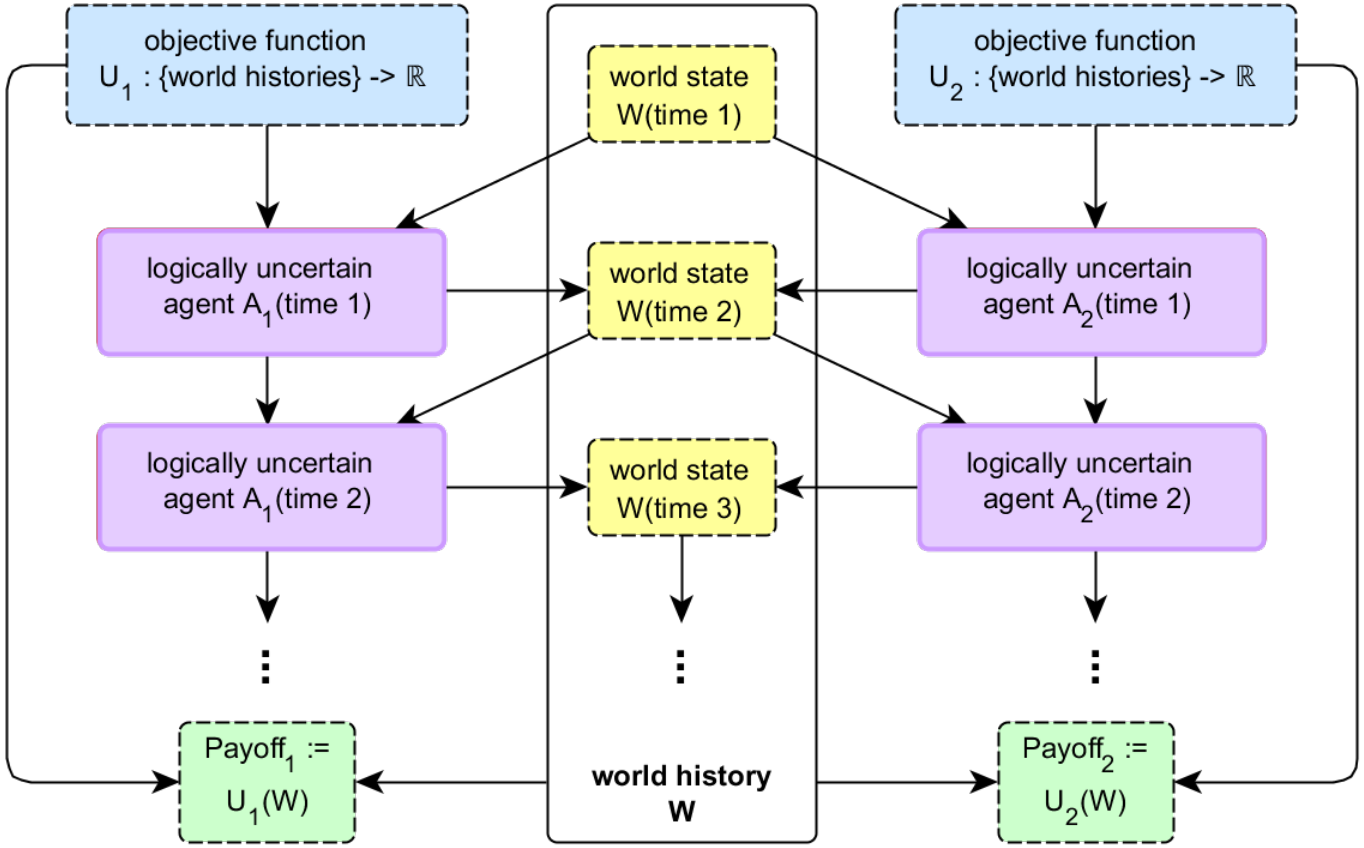
- **Game theory and economics:**
  - Von Neumann-Morgenstern utility theorem
  - Nash equilibria and correlated equilibria
  - Efficient market theory:
    - Fundamental theorems of welfare economics
    - Coase's Theorem
  - Value of Information (VOI)
- **Mechanism design**
  - Gibbard–Satterthwaite theorem
  - Myerson–Satterthwaite theorem
  - Revenue Equivalence theorem

Theoretical models of limited (and eventually, bounded) reasoners could help expand these fields to ask more questions directly relevant to artificial agents.

Visualizing a theoretical application

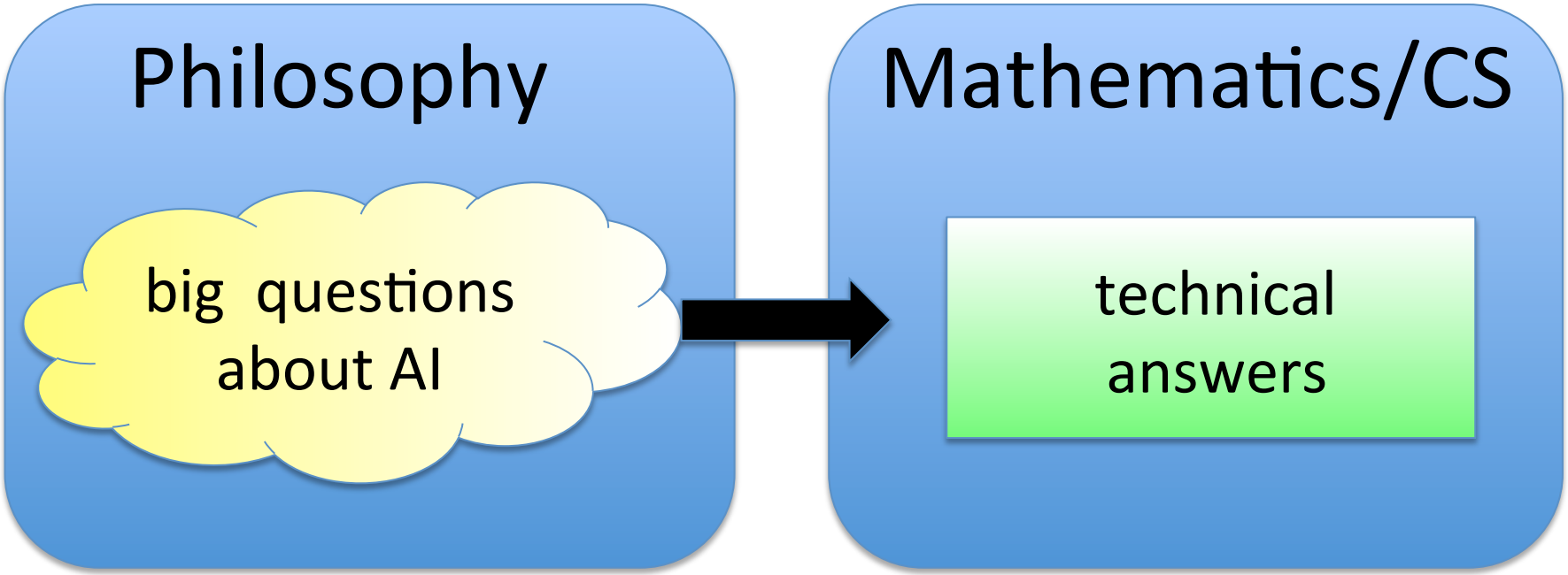
Currently, game theory analyzes scenarios with logically omniscient agents…

Now we can better theoretically analyze scenarios with bounded reasoners:



Meta updates

MIRI's general approach includes developing "big" questions about how AI can and should work, past the stages of philosophical conversation and into the domain of math and CS.



Meta updates

I was not personally expecting logical induction to be "solved" in this way for at least a decade, so I've updated that:

- I would like to see more theoreticians trying to beak down unsettled philosophical questions about intelligence and AI into math/CS and grinding through them like this; and
- perhaps other seemingly "out of reach" problems in AI alignment, like decision theory and logical counterfactuals, might be amenable to this sort of approach.

Thanks!

To

- **Scott Garrabrant**, for the core idea and many rapid subsequent insights;
- **Tsvi Benson Tilsen**, **Nate Soares**, and **Jessica Taylor** for co-developing the theory and resulting paper; and
- **Jimmy Rintjema** for a *lot* of help with LaTeX bugs and collaborative editing issues

<end of this talk>