Why ain’t you rich?

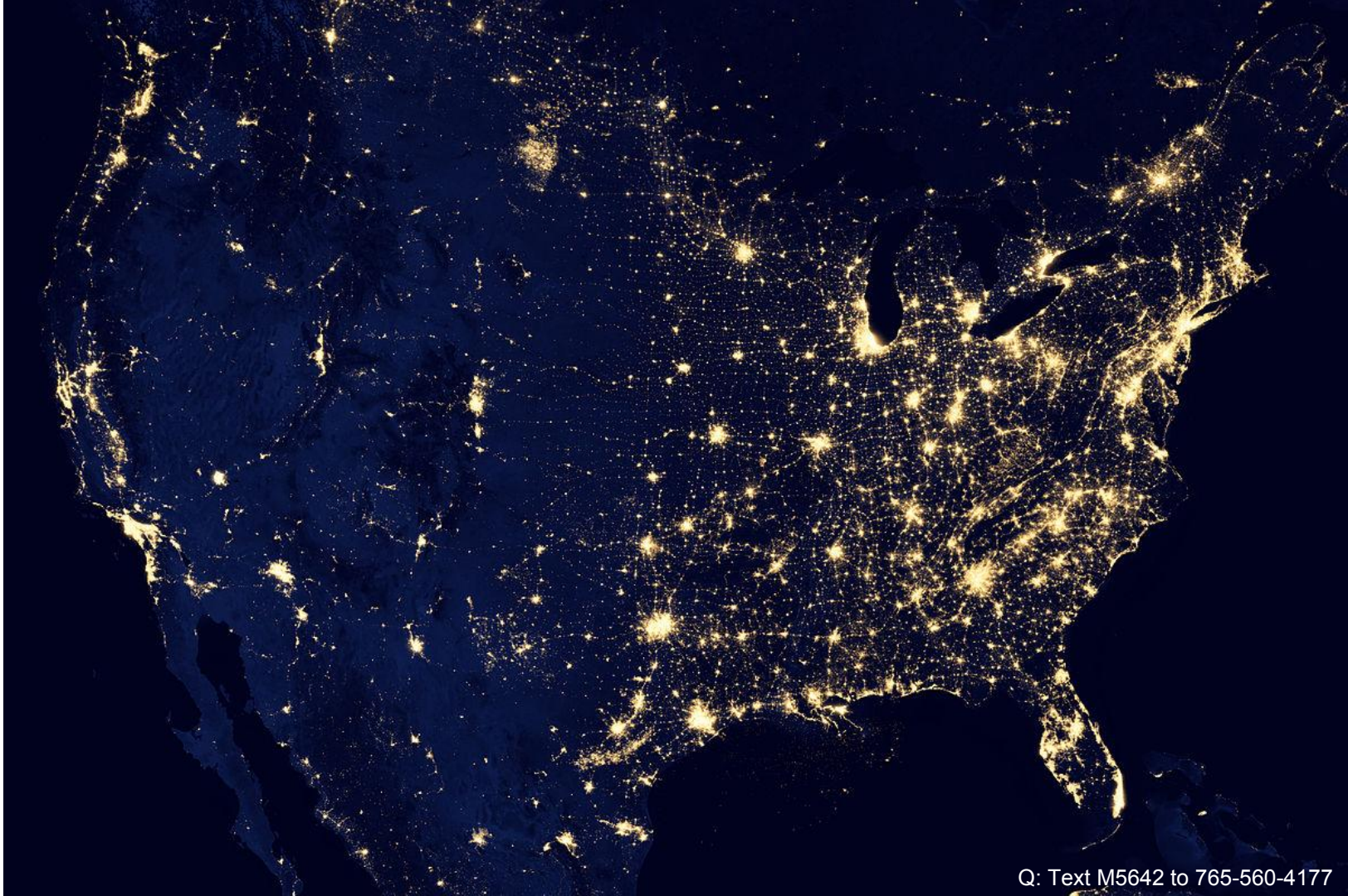
Why our current understanding of “rational choice” isn't good enough for superintelligence

Nate Soares

We do foundational mathematical research to ensure smarter-than-human artificial intelligence has a positive impact.

Questions?

Text M5642 + your question to 765-560-4177



City Lights of the United States 2012
by NASA Earth Observatory.
http://goo.gl/7rvKLr
for GeoTIFF original.
http://goo.gl/pKdQwM
for quicker access to the jpeg.
Licensed under Public domain
via Wikimedia Commons
http://goo.gl/W9Xjdx

Q: Text M5642 to 765-560-4177

- Tiling agent theory
- Logical uncertainty
- **Decision Theory**
- Corrigibility
- Value learning

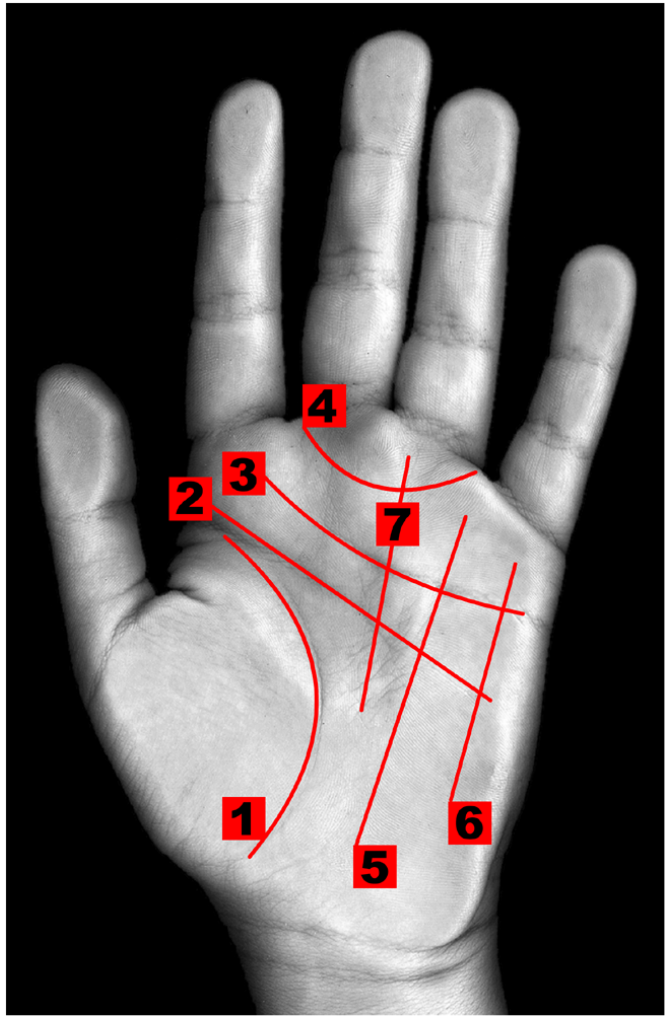
Q: Text M5642 to 765-560-4177

Why decision theory?

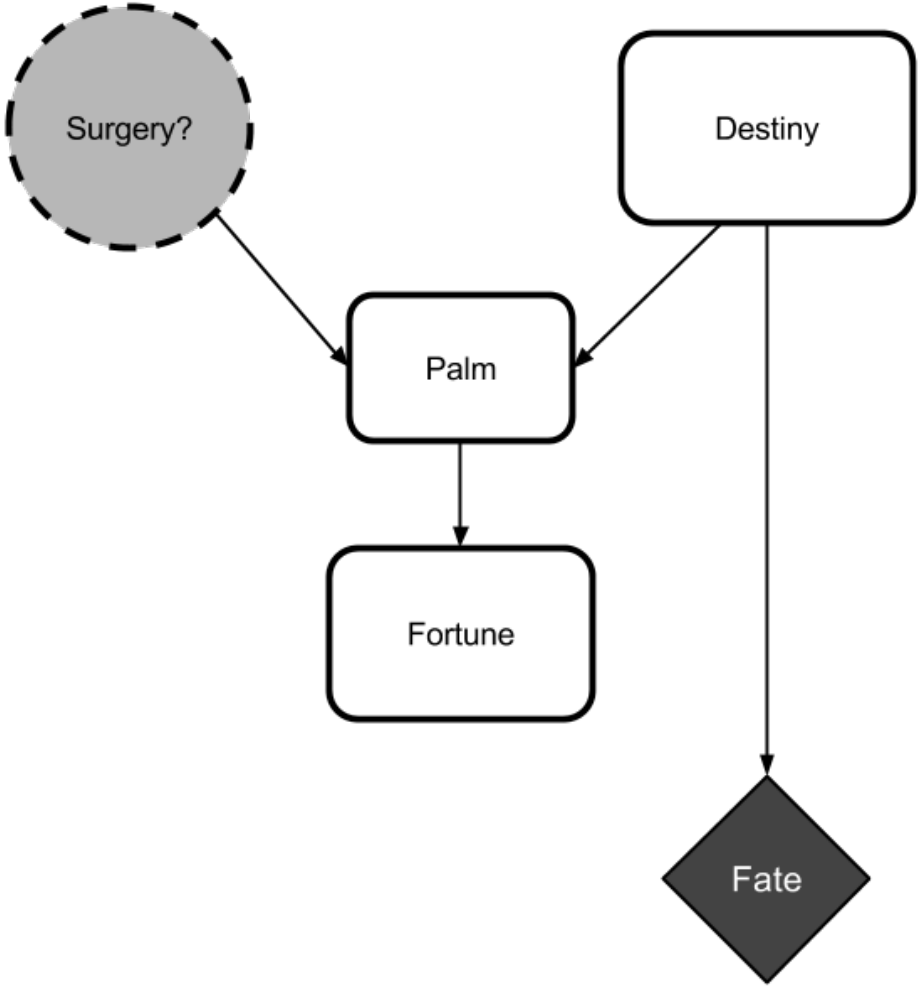
Q: Text M5642 to 765-560-4177

Why decision theory?

We need some way to reason counterfactually.



Q: Text M5642 to 765-560-4177



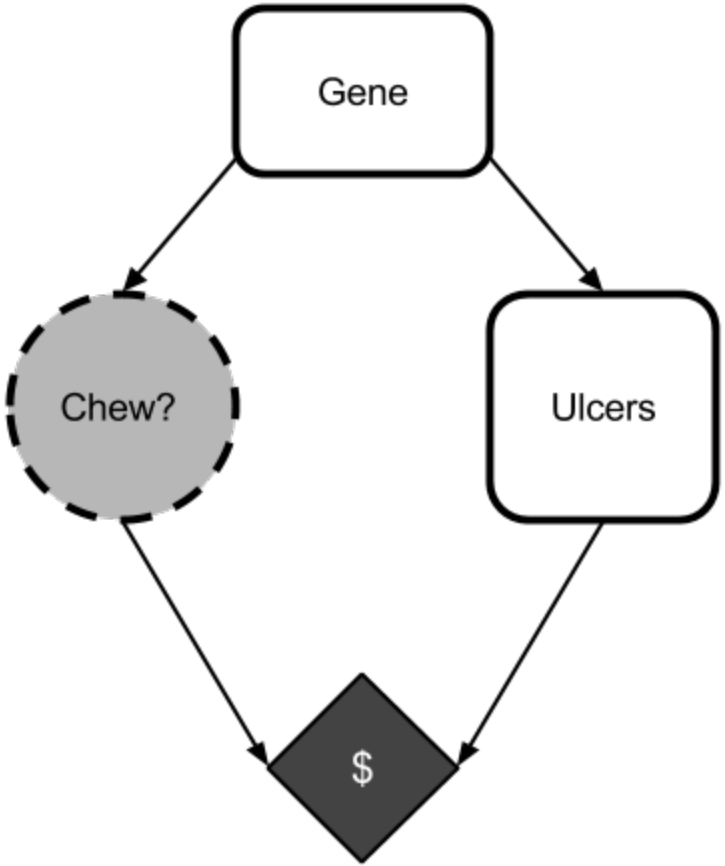
Q: Text M5642 to 765-560-4177

Causal Decision Theory (CDT)

1. Identify your action node **A**
2. Identify the available actions *Acts*
3. Identify your payoff node **U**
4. For each action *a* in *Acts*
   - Set **A**=*a* by overwriting **A** with a function that always returns *a*
   - Evaluate the expectation of **U** given that **A**=*a*
5. Take the action *a* with the highest associated value of **U**

How do you construct counterfactuals?

CDT prescribes considering action a by considering what would happen if, instead of being you, you were a simple function that always chose a.



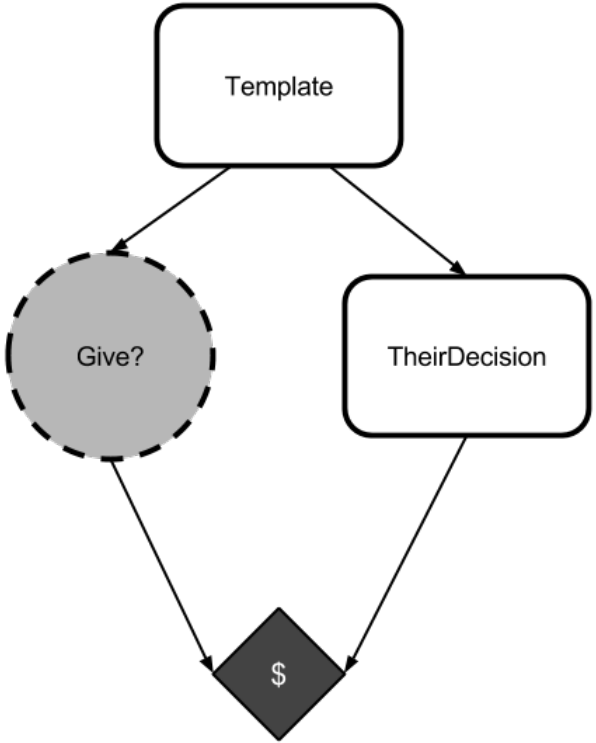
Q: Text M5642 to 765-560-4177

Token Trade

	Give	Keep
Give	($200, $200)	($0, $300)
Keep	($300, $0)	($100, $100)

Q: Text M5642 to 765-560-4177

Mirror Token Trade

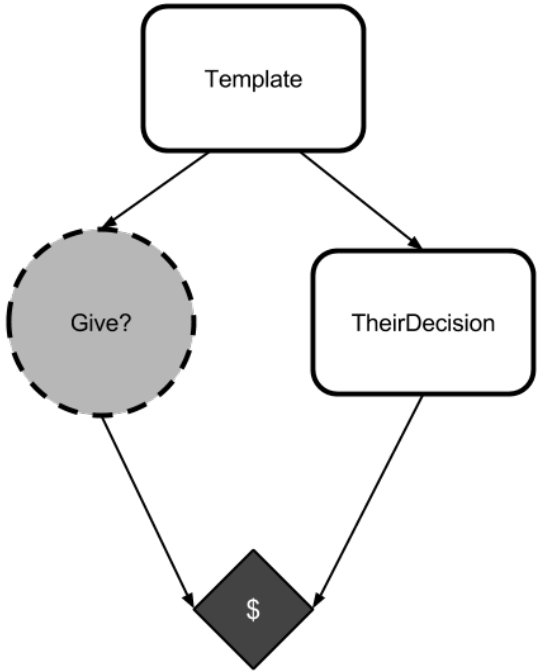


Q: Text M5642 to 765-560-4177

CDT loses

Given some probability p that **TheirDecision**=*give*

1. The action node is **`Give?`**
2. The actions are *`give`* and *`keep`*.
3. The payoff node is **$**.
4. If **`Give?`**=*`give`* then **$**=200p
5. If **`Give?`**=*`keep`* then **$**=300p + 100(1-p)
6. Take the action *`keep`*



Because 300p + 100(1-p) > 200p regardless of the value of p

Q: Text M5642 to 765-560-4177

Unfair game?

Unfair game?

Fair enough for me.

Unfair game?

Fair enough for me.

Why ain’t you rich?

Leaky scenarios

Known as “Newcomblike problems”

Leaky scenarios

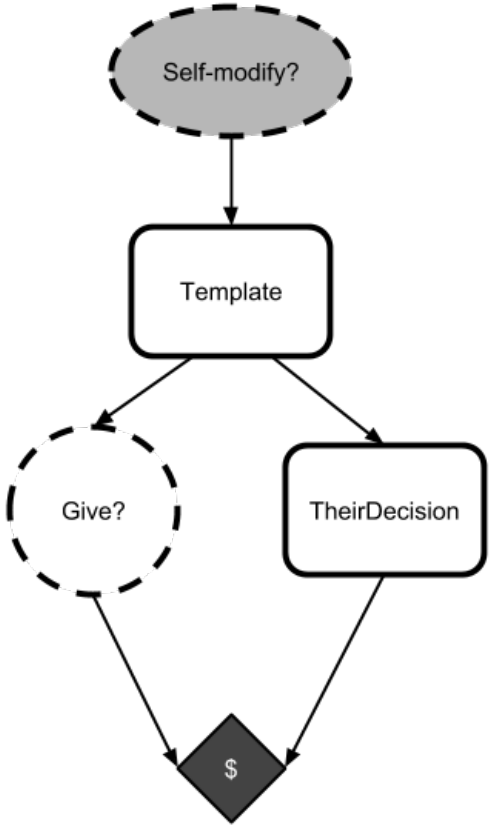
Known as “Newcomblike problems”

These scenarios are the norm.

CDT agents would stop using CDT

Q: Text M5642 to 765-560-4177

CDT agents would stop using CDT



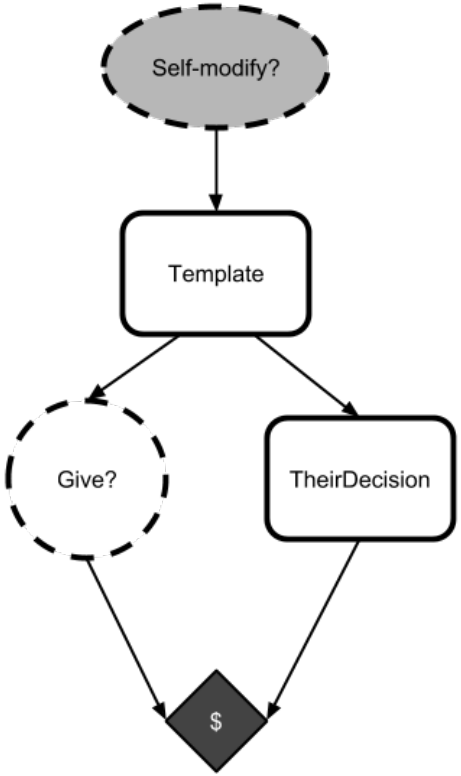
Q: Text M5642 to 765-560-4177

Q: Text M5642 to 765-560-4177

No self-correction

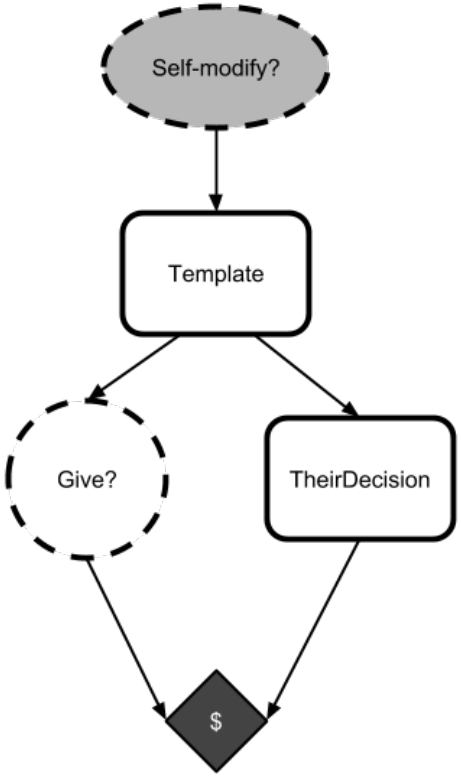
Q: Text M5642 to 765-560-4177

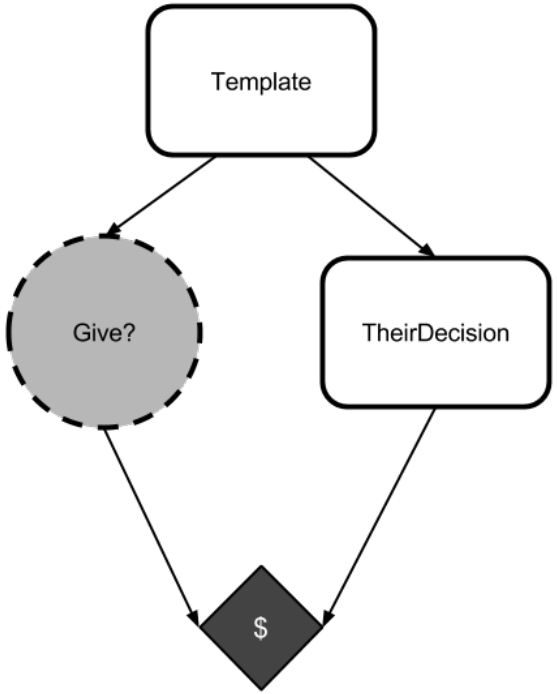
No self-correction



Q: Text M5642 to 765-560-4177

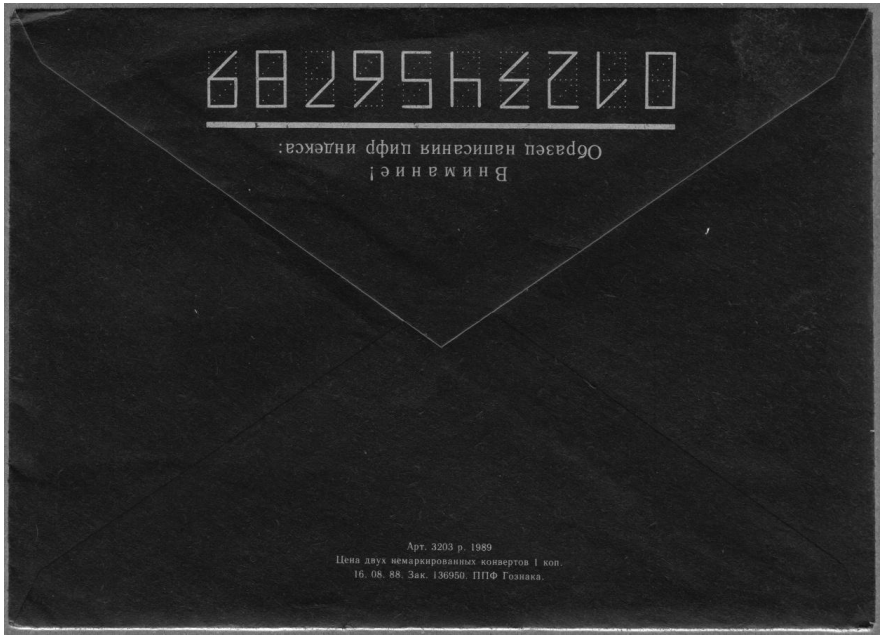
No self-correction





Q: Text M5642 to 765-560-4177

Strange Blackmail



Q: Text M5642 to 765-560-4177

Unrealistic?

Q: Text M5642 to 765-560-4177

Unrealistic?

Yes.

Q: Text M5642 to 765-560-4177

Unrealistic?

Yes.

But...

Q: Text M5642 to 765-560-4177

We don’t know what we’re doing

We don’t know what we’re doing

- How do you reason as if your action is connected to the reasoning of others?
  - We actually do have a way to solve this particular problem, but the solution is imperfect.

Q: Text M5642 to 765-560-4177

We don’t know what we’re doing

- How do you reason as if your action is connected to the reasoning of others?
  - We actually do have a way to solve this particular problem, but the solution is imperfect.
- What does good counterfactual reasoning look like?
  - And how does this affect your reasoning about how others are reasoning about you?

Q: Text M5642 to 765-560-4177

We don’t know what we’re doing

- How do you reason as if your action is connected to the reasoning of others?
  - We actually do have a way to solve this particular problem, but the solution is imperfect.
- What does good counterfactual reasoning look like?
  - And how does this affect your reasoning about how others are reasoning about you?
- *We don’t yet understand an algorithm that knowably converges on a good decision making procedure.*

Q: Text M5642 to 765-560-4177

Doom by default

- Tiling agent theory
- Logical uncertainty
- **Decision Theory**
- Corrigibility
- Value learning

Doom by default

What formal reasoning system could an intelligent agent use to gain very high confidence in similar systems?

Doom by default

Probability theory assumes we know all consequences of everything we know. How could an agent reason reliably under logical uncertainty?

Doom by default

Intelligent agents have, by default, strong instrumental incentives to preserve their goals, by manipulation or deception if necessary. How do we avoid these?

Doom by default

It is not enough to build something that *understands* what we want. We must build something that *wants* what we want.

Doom by default

We won’t get good behavior for free.

Dawn or doom?

It depends entirely upon whether we can figure out how to build a beneficial superintelligent system before we figure out how to build an arbitrary one.

Nate Soares

WHY ain't You rich?

NATE SOARES

BEHAVIORAL ECONOMICS PLAYS A PART...

WE DON'T KNOW WHAT WE ARE DOING...

THIS MIGHT BE FLAWED...

- GOALS
- VALUE
- DECISION THEORY → HOW TO REASON COUNTERFACTUALLY

UNSTABLE & UNSATISFACTORY

causal (CDT) DECISION THEORY

- EVALUATE OUTCOMES
- IMPLICATIONS OF DECISIONS

NOT A GOOD THING FOR AI

WHAT

GOOD COUNTERFACTUAL REASONING LOOK LIKE?

HUMAN EMOTIONS CAN BE TAKEN INTO ACCOUNT

TOKEN TRADE

KEEP?

TRADE?

PALM

WHICH WAY?

HOW THEY THINK YOU WILL ACT

"NEWCOMBLIKE" PROBLEMS

FATE

BUILDING A SUPER INTELLIGENCE W/CDT WILL MAKE THE SI MODIFY

NO SELF-CORRECTION

WE MUST BUILD SOMETHING THAT WANTS WHAT WE WANT

BENEFICIAL VS ARBITRARY

WE DON'T UNDERSTAND REASONING W/ LOGICAL UNCERTAINTY

DOOM BY DEFAULT...

WE NEED THIS!!

$\frac{1}{1\text{BILLION}}$ CHANCE OF FAILURE

I'M WORRIED ABOUT US MAKING AI BEFORE WE SOLVE PROBLEMS

WE WON'T GET GOOD BEHAVIOR FOR FREE

REWARD

BOMB

TAP INTO ORIGINAL SOURCE CODE

ON / OFF

ANY PAST SCENARIO CAN BECOME BLACK MAIL

MIRI

BENEFICIAL IMPACT BEFORE ONE IS TURNED ON...

Future

WHY OUR CURRENT UNDERSTANDING OF RATIONAL CHOICE ISN'T GOOD ENOUGH FOR SUPERINTELLIGENCE