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
- Tiling agent theory
- Logical uncertainty
- Decision Theory
- Corrigibility
- Value learning
Why decision theory?
Why decision theory?

We need some way to reason counterfactually.
"Les lignes de la main Artlibre".
Licensed under Free Art License via Wikimedia Commons.
http://goo.gl/5d6RK

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 \( a \).
Gene

Chew?

Ulcers

$
# Token Trade

<table>
<thead>
<tr>
<th></th>
<th>Give</th>
<th>Keep</th>
</tr>
</thead>
<tbody>
<tr>
<td>Give</td>
<td>($200, $200)</td>
<td>($0, $300)</td>
</tr>
<tr>
<td>Keep</td>
<td>($300, $0)</td>
<td>($100, $100)</td>
</tr>
</tbody>
</table>

Q: Text M5642 to 765-560-4177
Mirror Token Trade

Q: Text M5642 to 765-560-4177
CDT loses

Given some probability $p$ that $\text{TheirDecision}=\text{give}$

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

Because $300p + 100(1-p) > 200p$ regardless of the value of $p$
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
CDT agents would stop using CDT
No self-correction
No self-correction
No self-correction

Q: Text M5642 to 765-560-4177
Strange Blackmail

"USSR Blank Envelope 1989 - back" by Wesha. Own work. Licensed under Public domain via Wikimedia Commons.
http://goo.gl/oIOC9Y

Q: Text M5642 to 765-560-4177
Unrealistic?
Unrealistic?

Yes.
Unrealistic?

Yes.

But...
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.
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 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.
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
WHAT
Good counterfactual reasoning looks like?

WE MUST BUILD SOMETHING THAT WANTS WHAT WE WANT

BENEFICIAL VS ARBITRARY
WE DON'T UNDERSTAND REASONING w/ LOGICAL UNCERTAINTY

DOOM BY DEFAULT...

NO SELF-CORRECTION

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

HUMAN EMOTIONS CAN BE TAKEN INTO ACCOUNT

TOKEN TRADE
KEEP? TRAPE?
PALM

HOW THEY THINK YOU WILL ACT

NEWCOMBLIKE PROBLEMS

WE NEED THIS!!

WE WANT A BILLION CHANCE OF FAILURE
WE WON'T GET GOOD BEHAVIOR FOR FREE

TAP INTO ORIGINAL SOURCE CODE

ANY PAST SCENARIO CAN BECOME BLACKMAIL

INK FACTORY
you talk. we draw. it's awesome.

WHY AREN'T YOU RICH?

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

MIRI

BENEFICIAL IMPACT BEFORE ONE IS TURNED ON...

NATE SOORES

FUTURE

UNSTABLE & UNSATISFACTORY

NOT A GOOD THING FOR AI

GOALS VALUE

DECISION THEORY

WHO

HOW TO REASON COUNTERFACTUALLY

EVALUATE OUTCOMES

IMPLICATIONS OF DECISIONS

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