What is a what if?

Nate Soares
Preferences aren’t enough

<table>
<thead>
<tr>
<th></th>
<th>Perfect Copy</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Cooper rate</td>
</tr>
<tr>
<td>You</td>
<td>2 2</td>
</tr>
<tr>
<td>Defect</td>
<td>3 0</td>
</tr>
</tbody>
</table>
Preferences aren’t enough

<table>
<thead>
<tr>
<th></th>
<th>Perfect Copy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coop rate</td>
<td>2 0 1 1</td>
</tr>
<tr>
<td>Defect</td>
<td>2 3 0 1</td>
</tr>
</tbody>
</table>

Causal choice
Preferences aren’t enough

<table>
<thead>
<tr>
<th></th>
<th>Perfect Copy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cooperate</td>
<td>2 2</td>
</tr>
<tr>
<td>Defect</td>
<td>3 0</td>
</tr>
</tbody>
</table>

Causal choice
What is a what if?
<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td>🟢</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>🟠</td>
<td></td>
<td>🟠</td>
</tr>
<tr>
<td>3</td>
<td>🟤</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Diagram:

- **U** (Up)
- **L** (Left)
- **R** (Right)
- **D** (Down)

Environment

- 11° U 12° U
- 11° U 12° R
- 11° U 12° L
- 11° U 12° D

Cells:

- 12°
- 11°
- 13°
- 22°
What is a what if?
Attempt #1: Evidential reasoning
Problem: Zero probability actions
Problem: Managing the news
Conditionals are not counterfactuals

Conditioning on “I take $1” is not the same as asking “what if I take $1?”
What is a what if?
Attempt #2: Causal reasoning

- Weather
- Forecast
- Umbrella?
- Umbrella?
Problem: Where’s the agent?
Problem: Logical links are neglected

Cooperate? \[\text{Identical}\] Clone

U
Problem: Logical links are neglected
Problem: Logical links are neglected
Causal counterfactuals aren’t idealized what ifs.

Humans take logical non-causal connections into account. Causal counterfactuals don’t.
What is a what if?
Attempt #3: Repair causal graphs
Problem: Generating the graph
Problem: Generating the graph
Problem: *Updating* the graph
Problem: *Updating* the graph

![Diagram]

- **You**
- **Opponent**
- **Yes!**
- **U**
What would happen if my algorithm had a different output?
“Logical decision theory”

“Timeless decision theory” (Yudkowsky 2009)
“Updateless decision theory” (Dai 2009)
“Causal decision theory” (Spohn 2012)
Prisoner’s dilemma vs clone

You

Clone

U
Prisoner’s dilemma vs clone

Cooperate?

You

Clone

U
Prisoner’s dilemma vs clone

- You: Cooperate? Yes
- Clone: Cooperate? Yes

Result: U
Newcomb’s problem
Newcomb’s problem

Onebox?

You

Predictor

U

U
Newcomb's problem
Death in Damascus

You

U

Death’s book

Death
Death in Damascus

1. You
2. Where?
3. Death's book
4. Death

U
Death in Damascus

Where?

Death’s book

You

Death

✓

U
What if logical links had come first?

Cooperate?

You

Clone

U
Well if it’s so great, why don’t you formalize it?
Opponent

My Algorithm

You

? U Opponent
We need a notion of *logical* counterfactuals.
If you know how assuming $A()=a$ affects $B()$ for arbitrary algorithms, you’re done.
If \texttt{Agent()} had output \texttt{a}, what output would \texttt{Environment()} have?
Counterpossibilities

def PLDT():
    map = {}
    for π in policies:
        if PA can prove “PLDT()=π \rightarrow \text{Environment}()=u”
            (for some u):
                map[π]=u
    return the π that maximizes map[π]
if I can prove

“Me()=π → Environment()=υ”

...
Counterpossibilities
def PLDT():
    map = {}
    for π in policies:
        if PA can prove "PLDT()=π → Environment()=u" (for some u):
            map[π]=u
    return the π that maximizes map[π]
What is a what if?
def PLDT():
    for π in policies:
        if PA can prove “PLDT()≠π”:
            return π

map = {}
for π in policies:
    if PA can prove “PLDT()=π → Environment()=u”
        (for some u):
        map[π]=u
return the π that maximizes map[π]
The Mad Newcomb Problem
1. Oh, they weren’t reasoning about me at all.
The Mad Newcomb Problem

1. Oh, they weren’t reasoning about me at all.
2. Huh, this reasoning is flawed.
1. Oh, they weren’t reasoning about me at all.
2. Huh, this reasoning is flawed.
3. This reasoning is about me if I hadn’t seen this proof.
1. Oh, they weren’t reasoning about me at all.
2. Huh, this reasoning is flawed.
3. This reasoning is about me if I hadn’t seen this proof.
4. Hey wait, this is accurate reasoning about me now!
It won’t work, but it will would have worked.
The Toxin Problem

Drink?

You

Toxin

Prediction of Intention

Payment

U
What exactly can you *logically* affect?
Environment

Agent
We need a better understanding of how to reason about what would happen if an algorithm did something it doesn’t.
Which logical relationships do we respect, and how?
Preferences aren’t enough

<table>
<thead>
<tr>
<th></th>
<th>Perfect Copy</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>You</strong></td>
<td></td>
</tr>
<tr>
<td>Coop rate</td>
<td>2 2</td>
</tr>
<tr>
<td>Defect</td>
<td>3 0</td>
</tr>
<tr>
<td><strong>Coop rate</strong></td>
<td>2 2</td>
</tr>
<tr>
<td>Defect</td>
<td>3 0</td>
</tr>
</tbody>
</table>

Causal choice
What is a what if?