What is a what if?

Nate Soares

MIRI

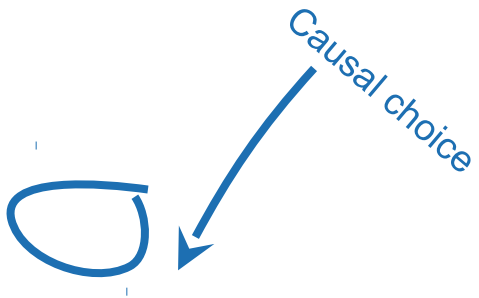
MACHINE INTELLIGENCE
RESEARCH INSTITUTE

Preferences aren’t enough

		Perfect Copy	
		Cooperate	Defect
You	Cooperate	2 2	0 3
	Defect	3 0	1 1

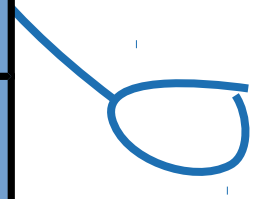
Preferences aren’t enough

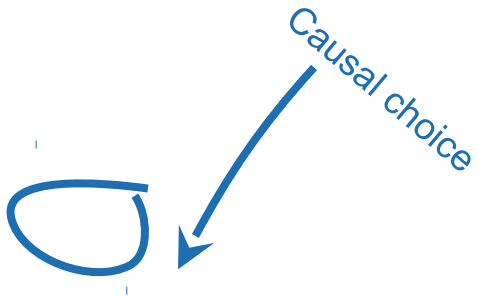
		Perfect Copy	
		Coope rate	Defect
You	Cooper ate	2 2	0 3
	Defect	3 0	1 1



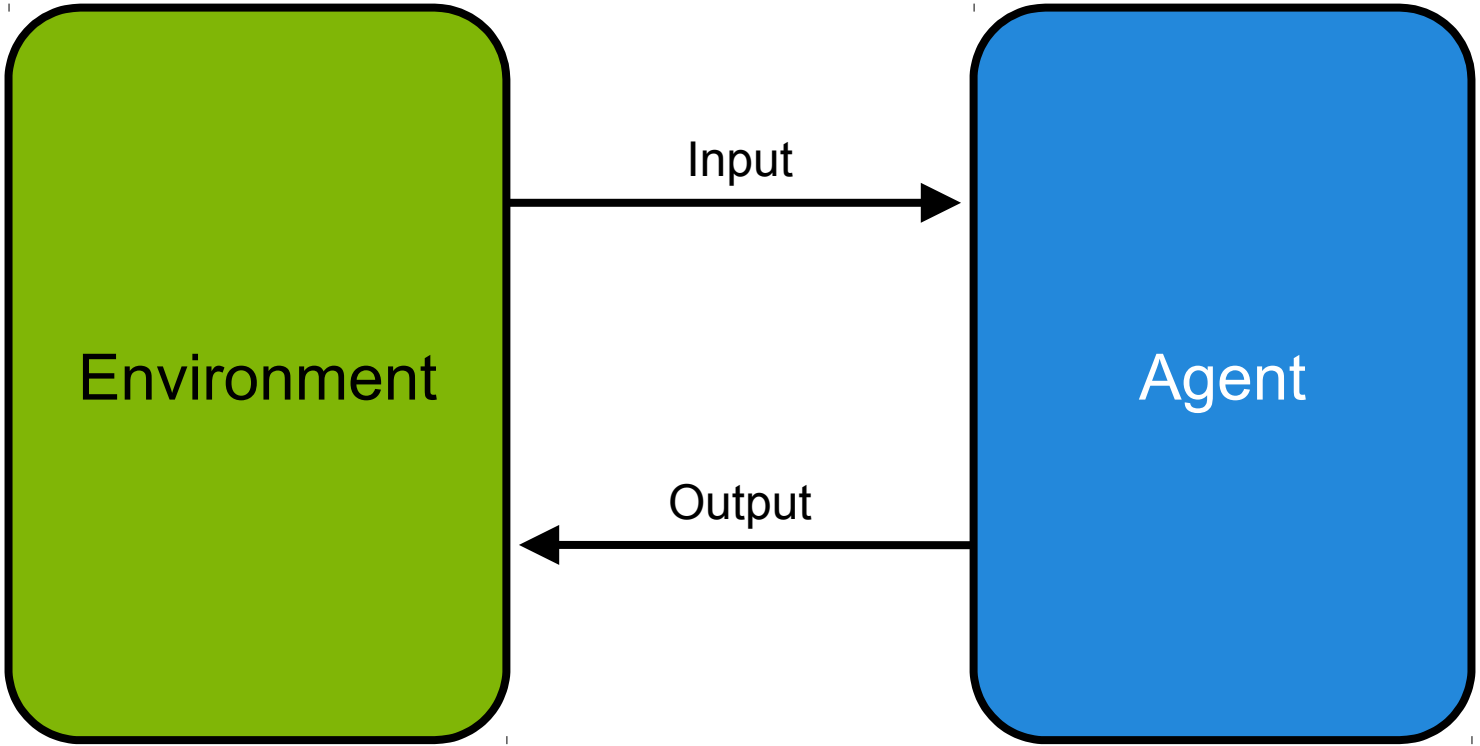
Preferences aren’t enough

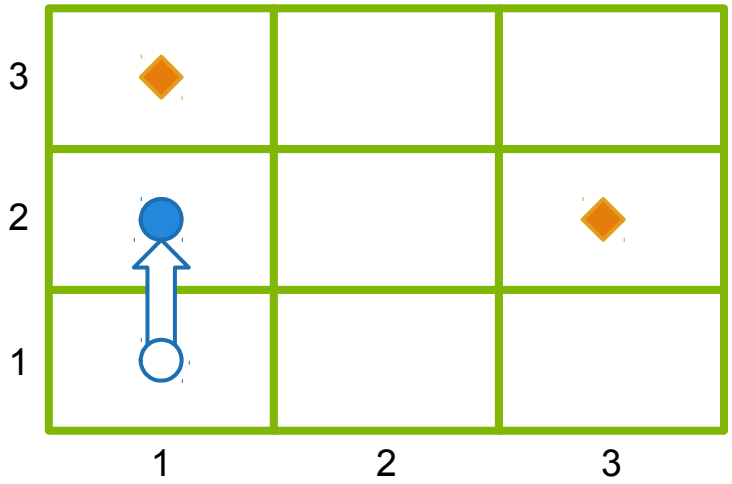
		Perfect Copy	
		Coope rate	Defect
You	Cooper ate	2 2	0 3
	Defect	3 0	1 1

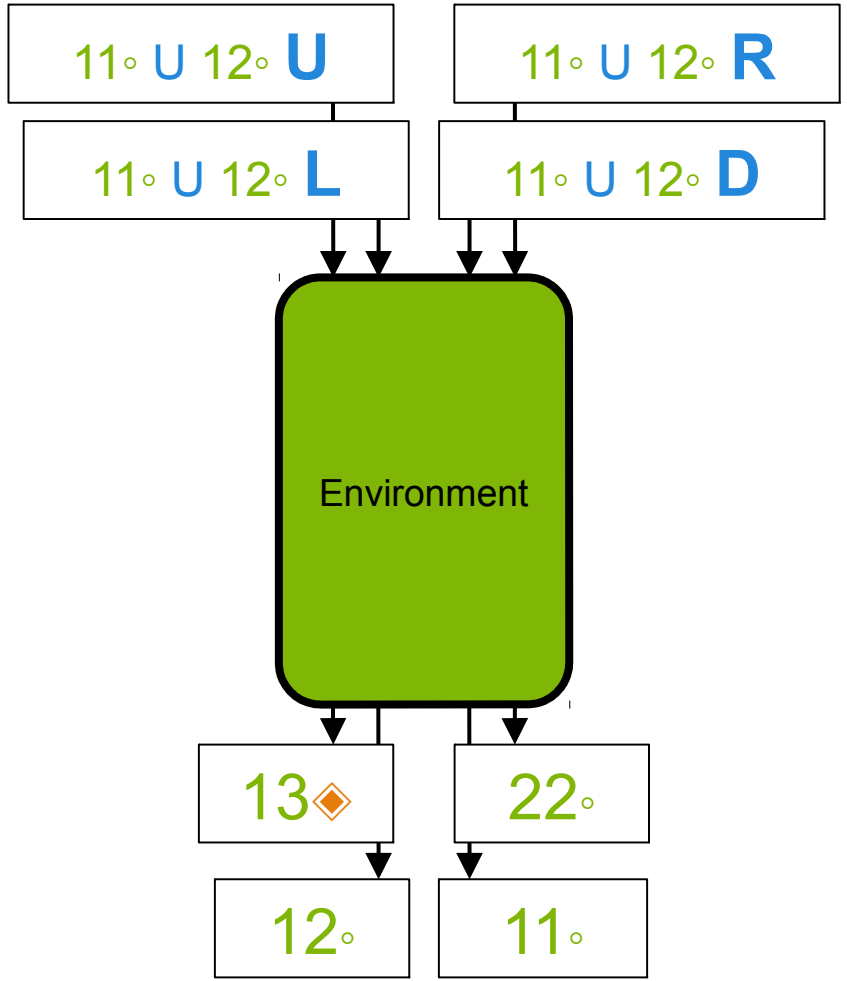


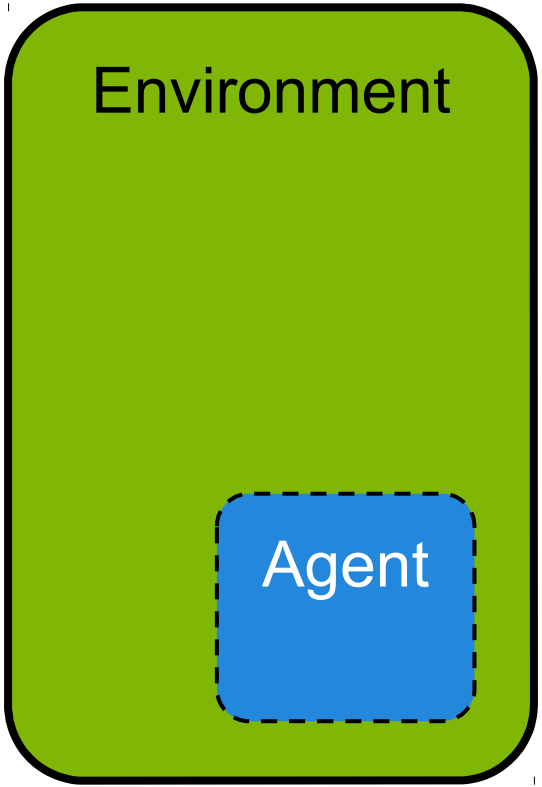


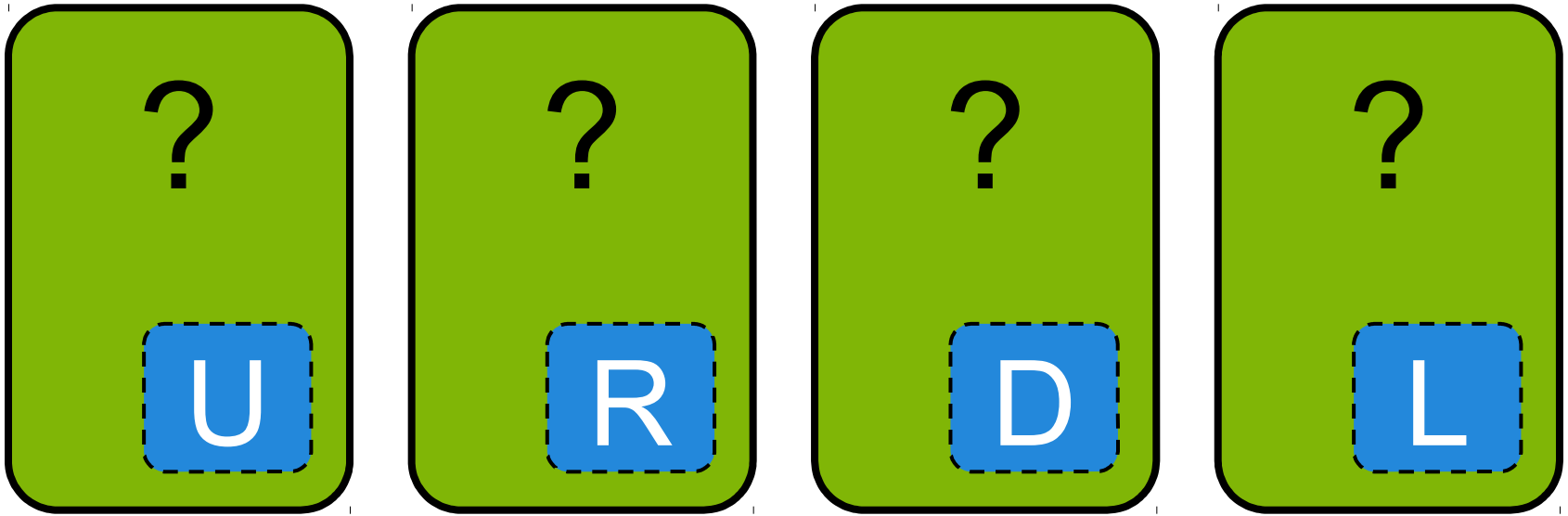
What is a what if?

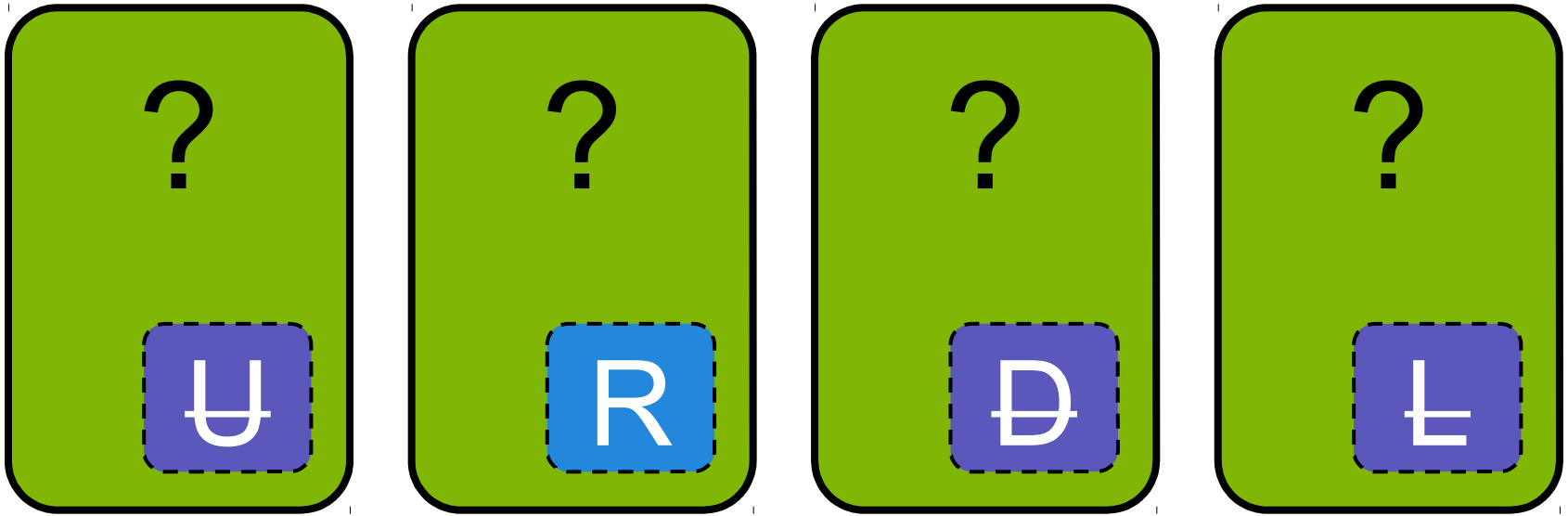

Environment
Input
Output
Agent


3
2
1
1
2
3


11◦ U 12◦ U
11◦ U 12◦ R
11◦ U 12◦ L
11◦ U 12◦ D
Environment
13
22◦
12◦
11◦

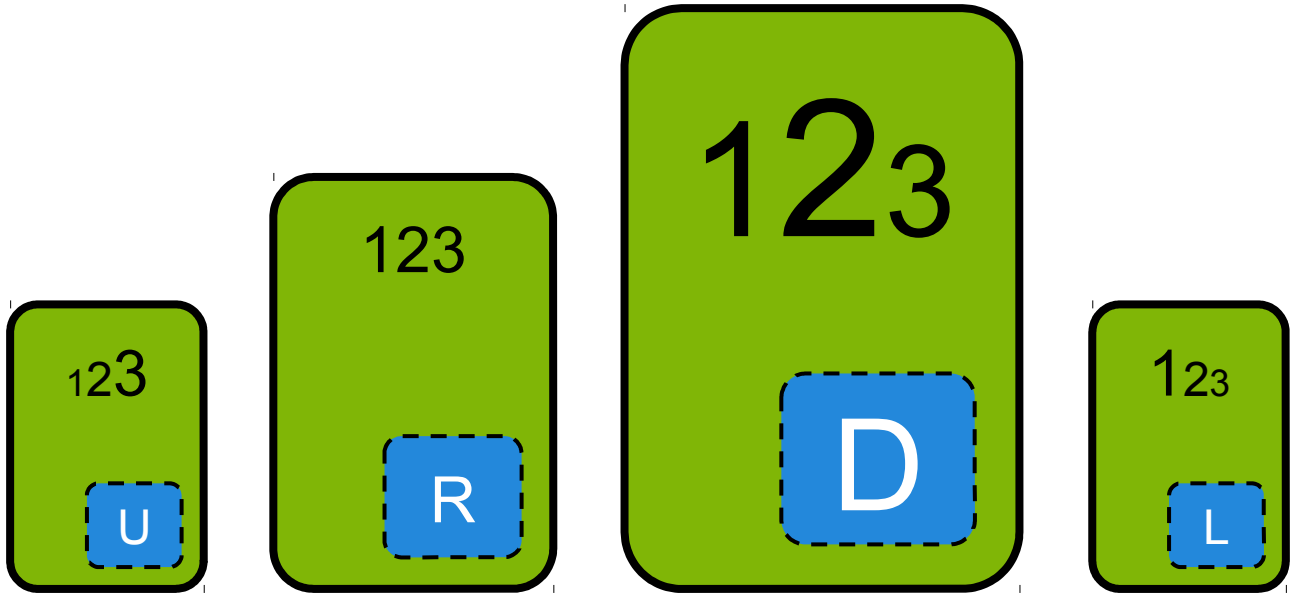

Environment
Agent


?
U
?
R
?
D
?
L

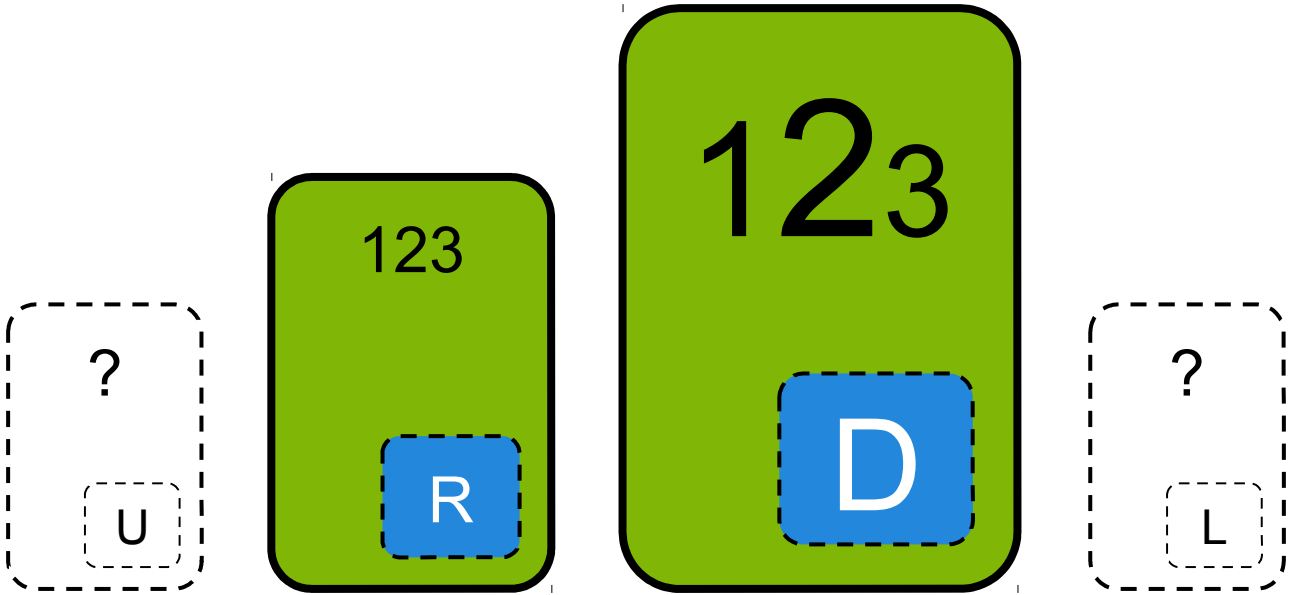

?
U
?
R
?
D
?
L

What is a what if?

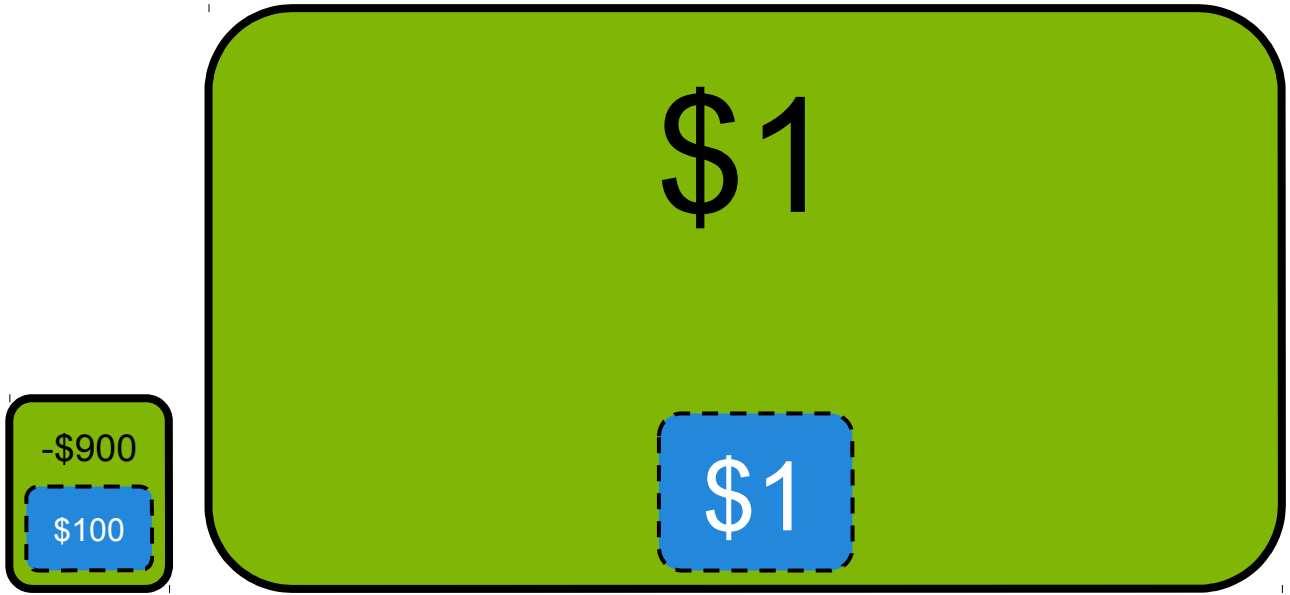
Attempt #1: Evidential reasoning



Problem: Zero probability actions



Problem: Managing the news

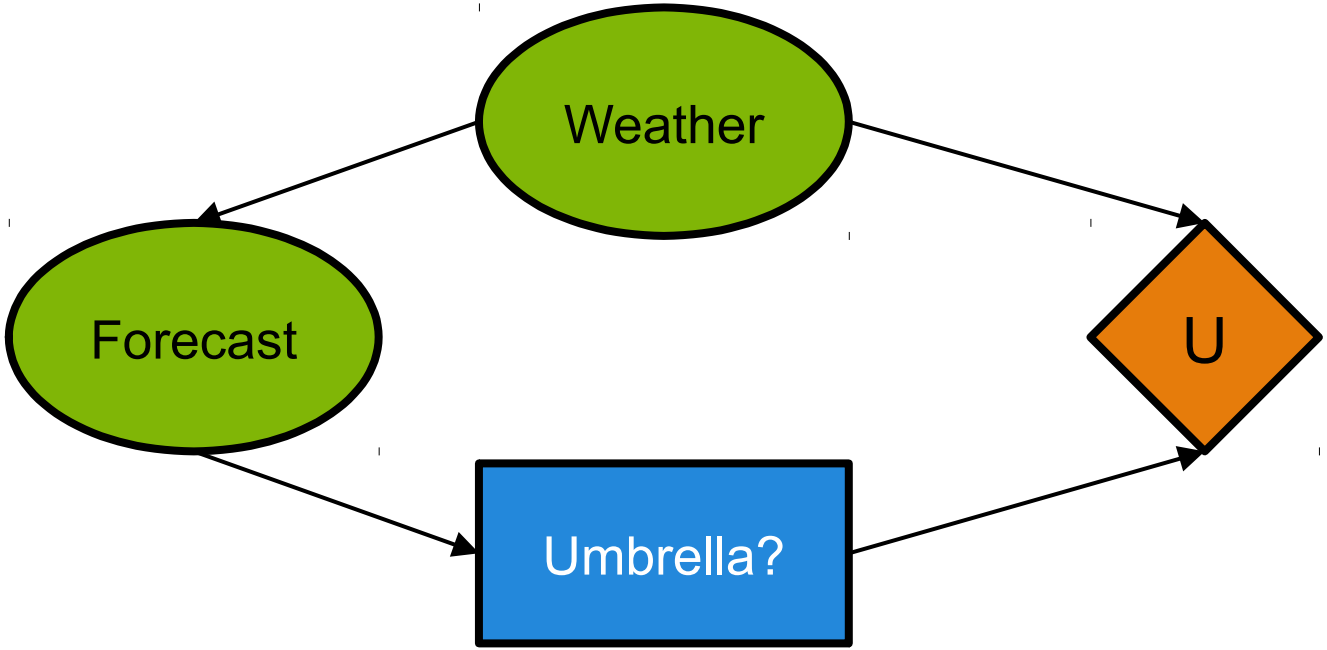


Conditionals are not counterfactuals

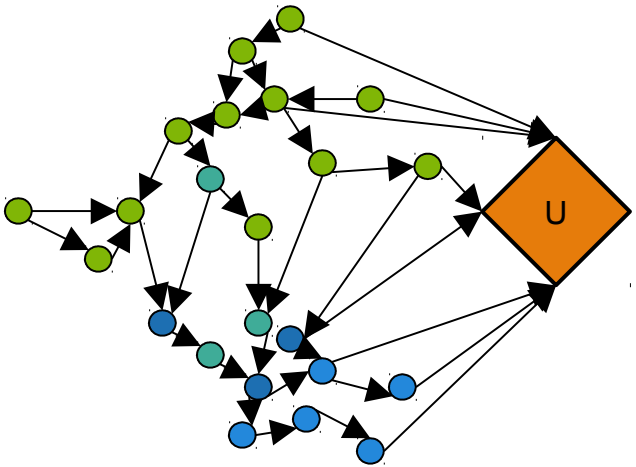
Conditioning on “I take $1” is not the same as asking “what if I take $1?”

What is a what if?

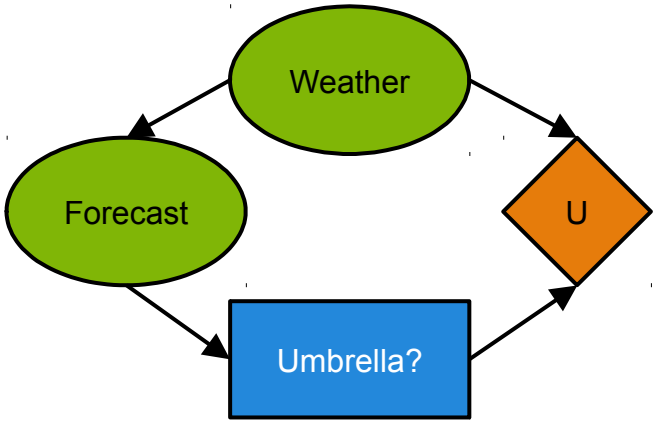
Attempt #2: Causal reasoning



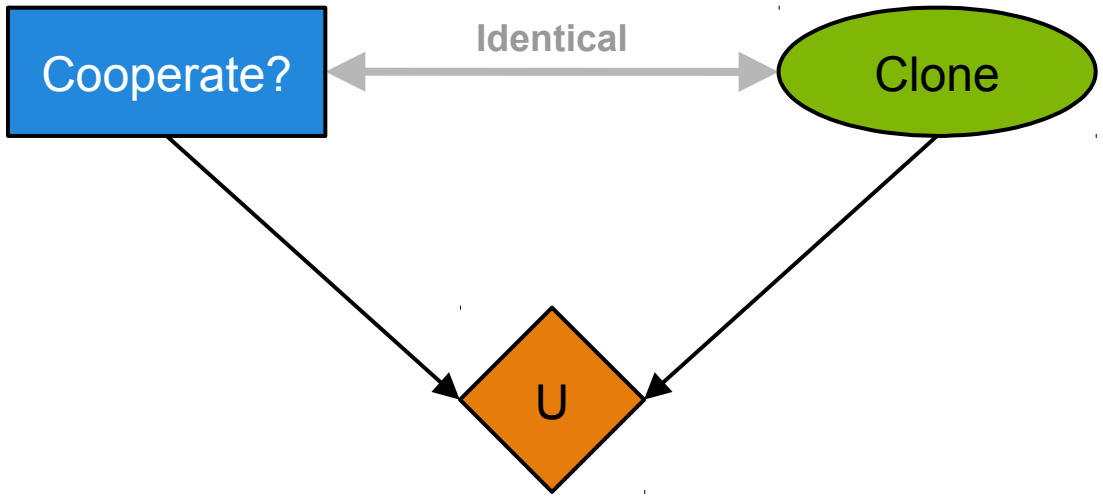
Problem: Where's the agent?



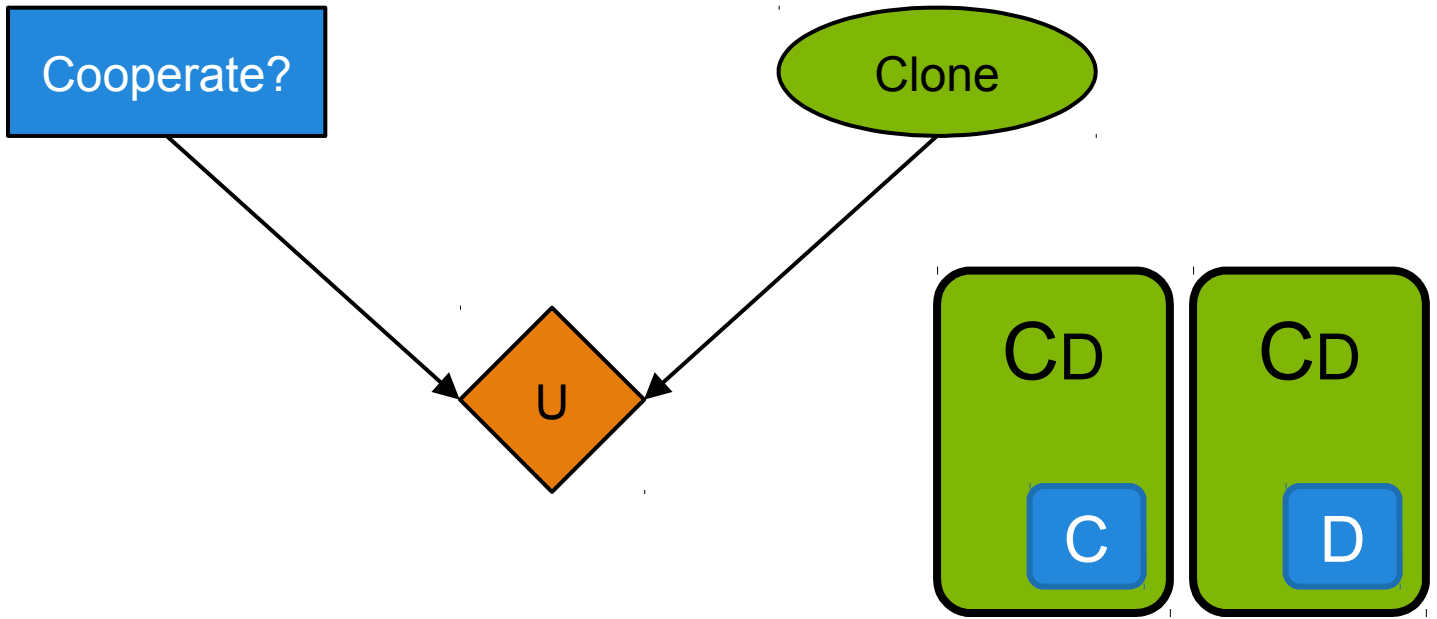
≠



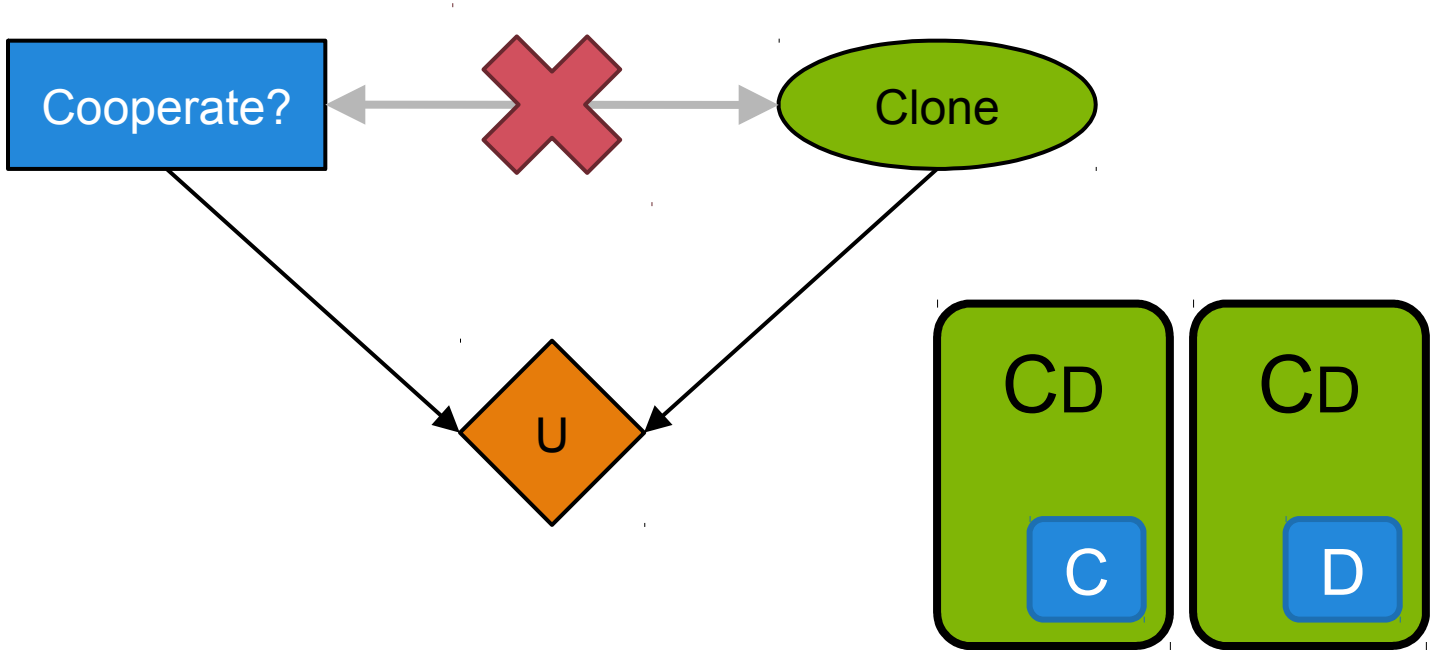
Problem: Logical links are neglected



Problem: Logical links are neglected



Problem: Logical links are neglected

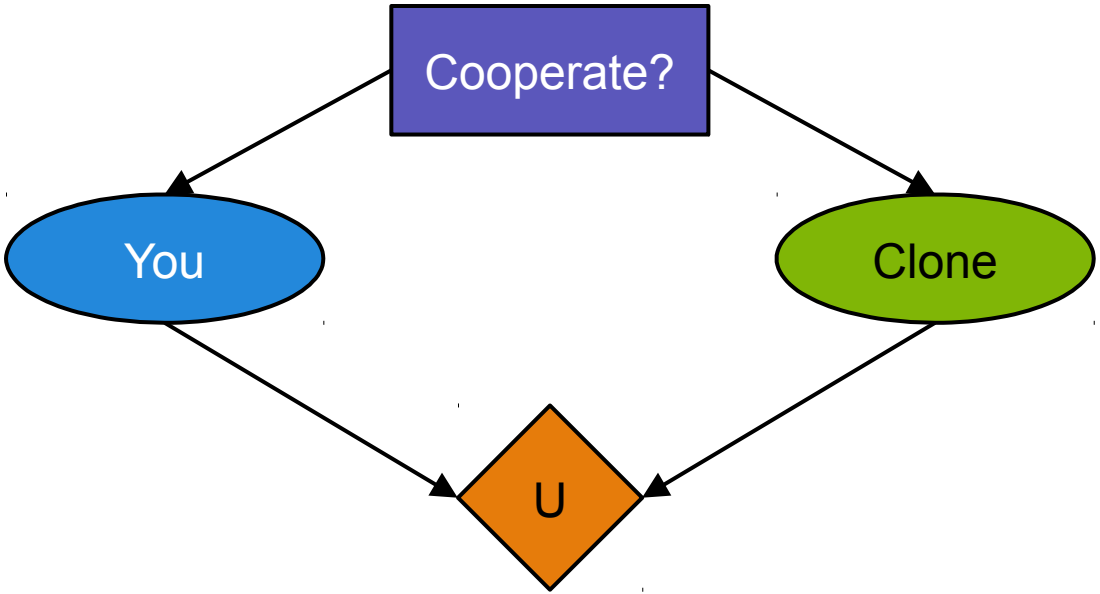


Causal counterfactuals aren’t idealized what ifs.

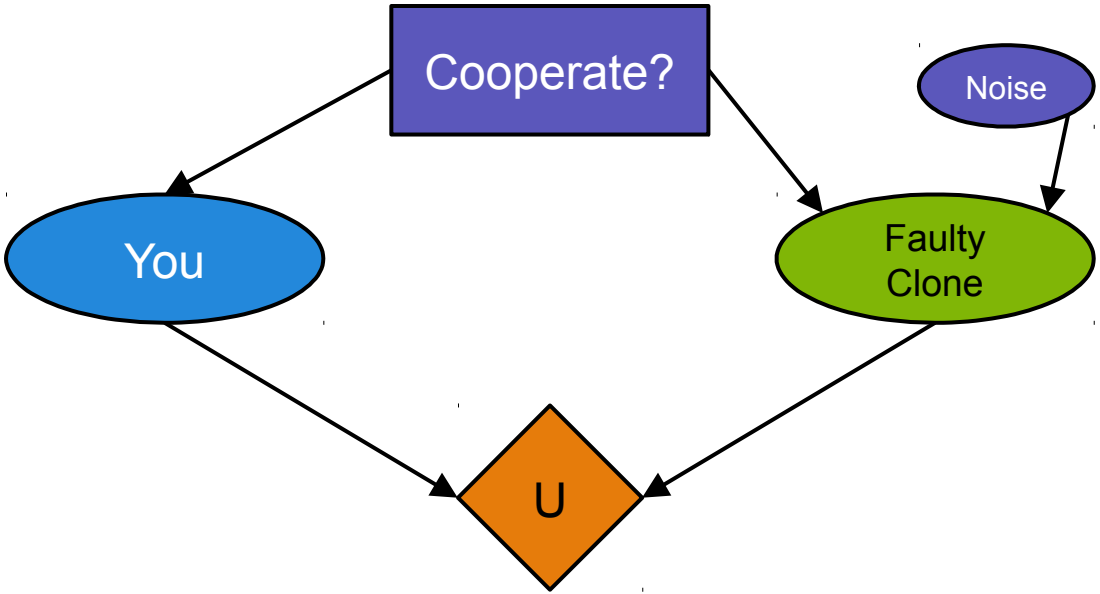
Humans take logical non-causal connections into account. Causal counterfactuals don’t.

What is a what if?

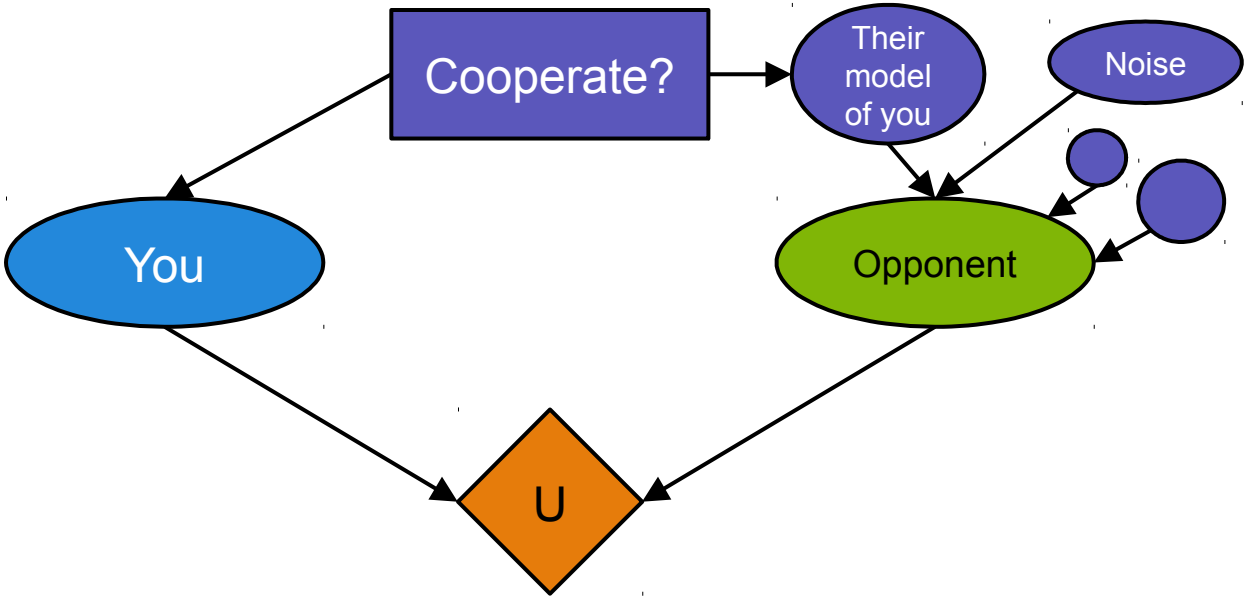
Attempt #3: Repair causal graphs



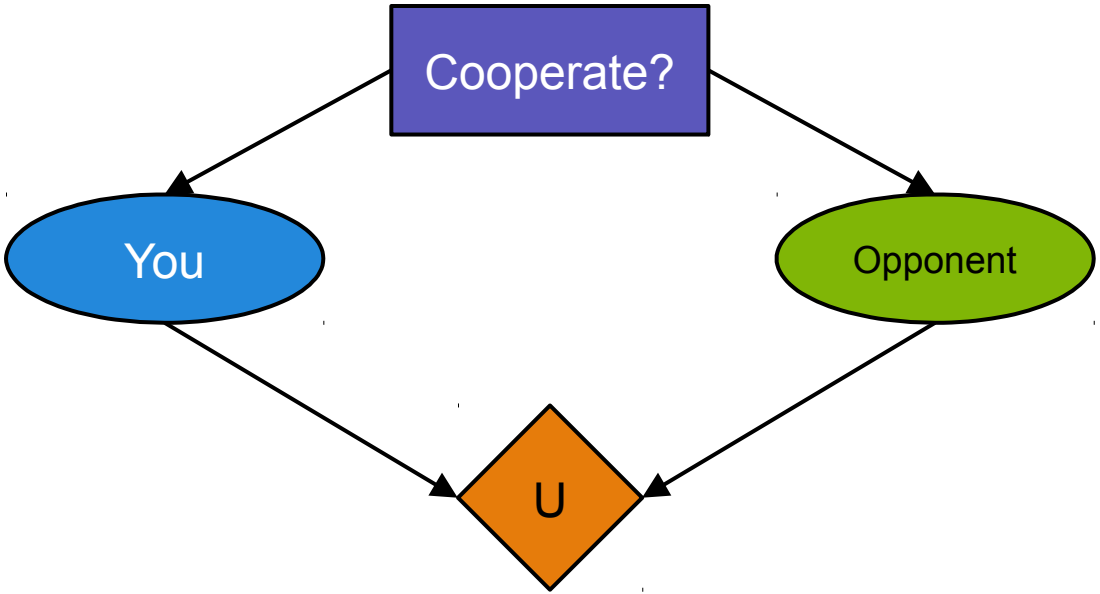
Problem: Generating the graph



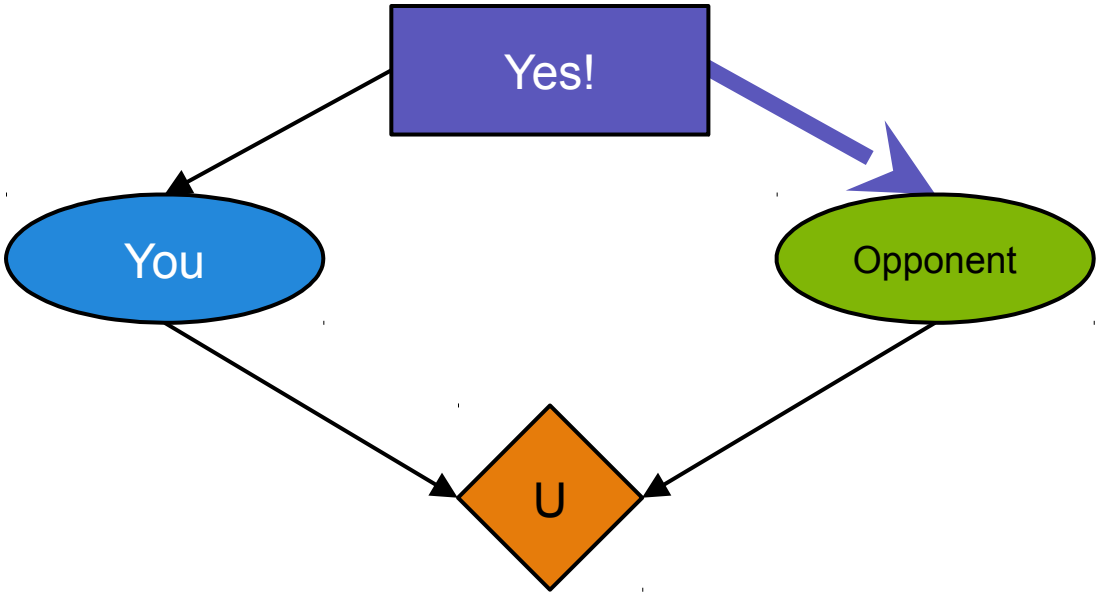
Problem: Generating the graph



Problem: *Updating* the graph



Problem: *Updating* the graph

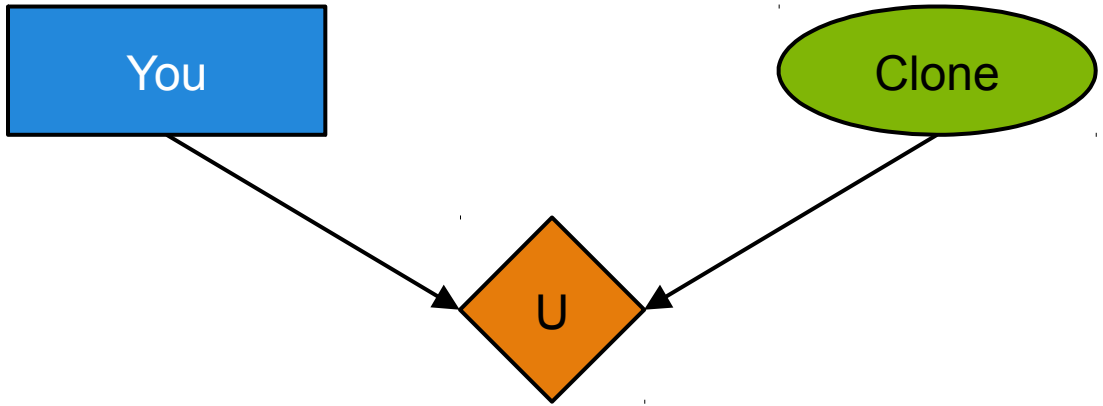


What would happen if my algorithm had a different output?

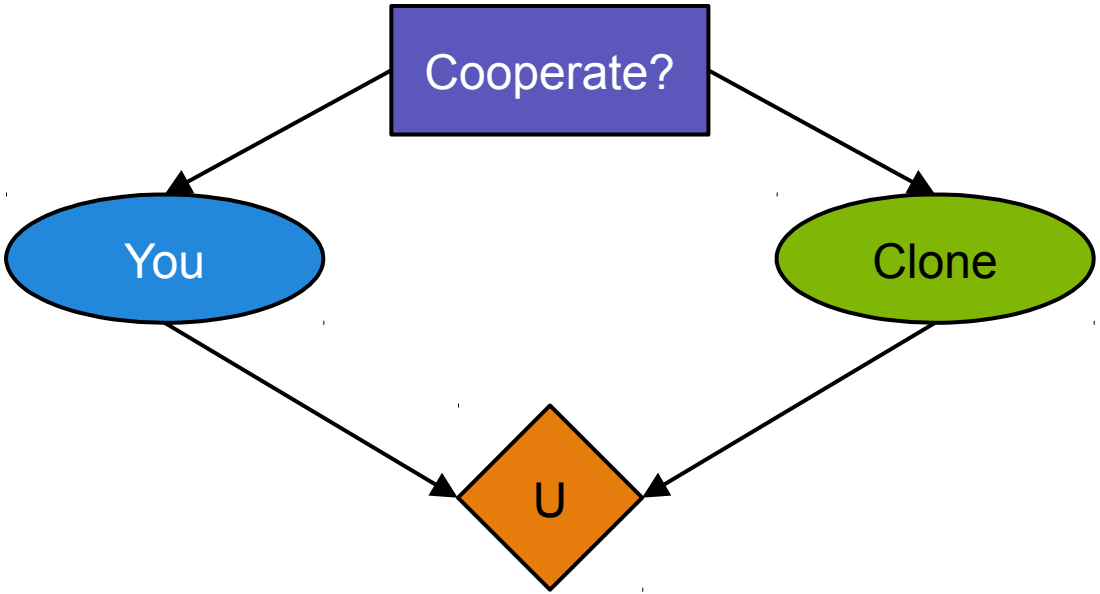
“Logical decision theory”

“Timeless decision theory” (Yudkowsky 2009)
“Updateless decision theory” (Dai 2009)
“Causal decision theory” (Spohn 2012)

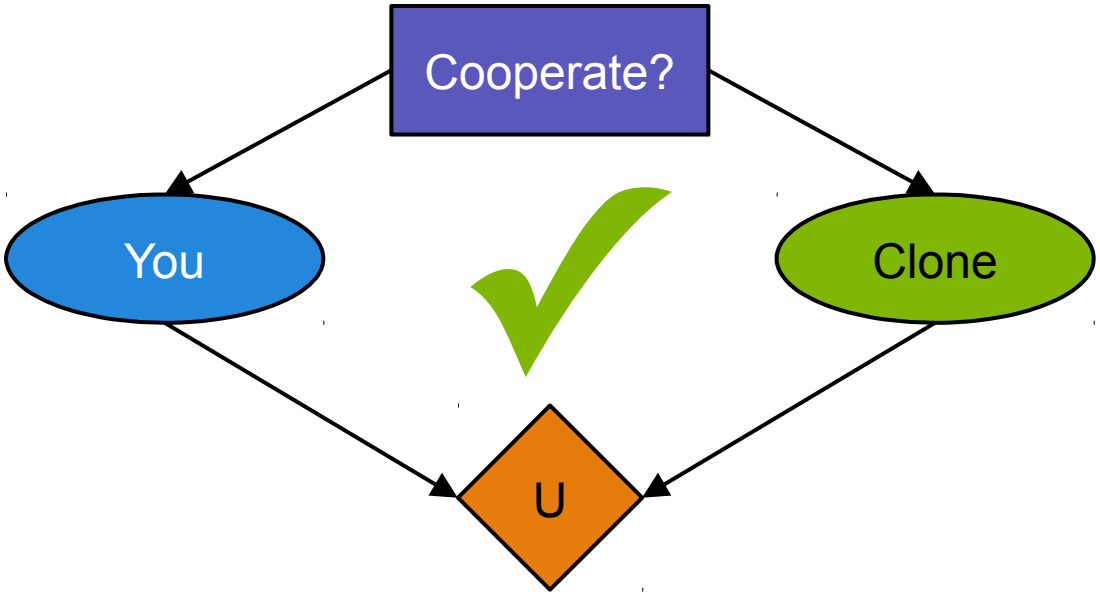
Prisoner’s dilemma vs clone



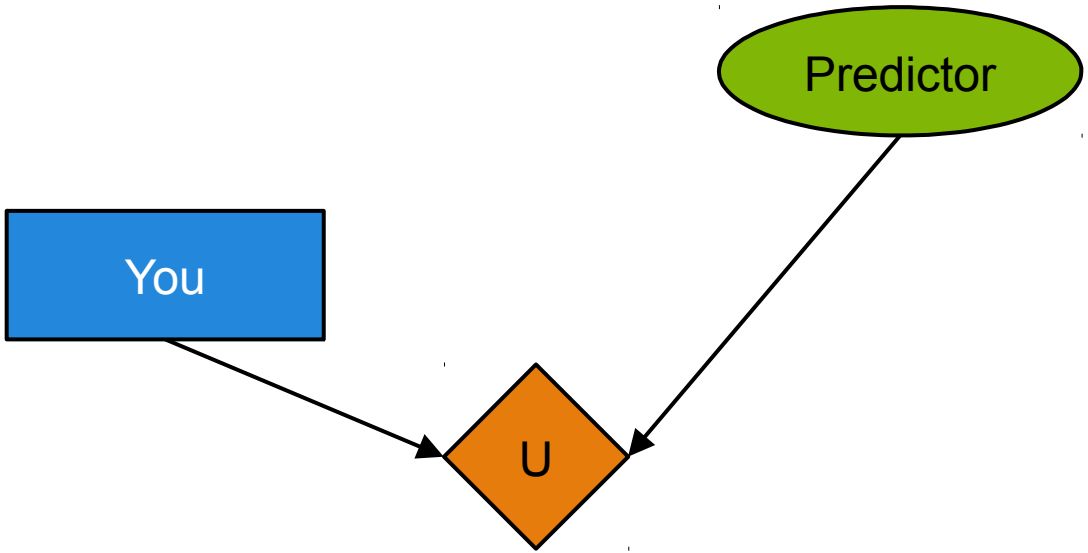
Prisoner’s dilemma vs clone



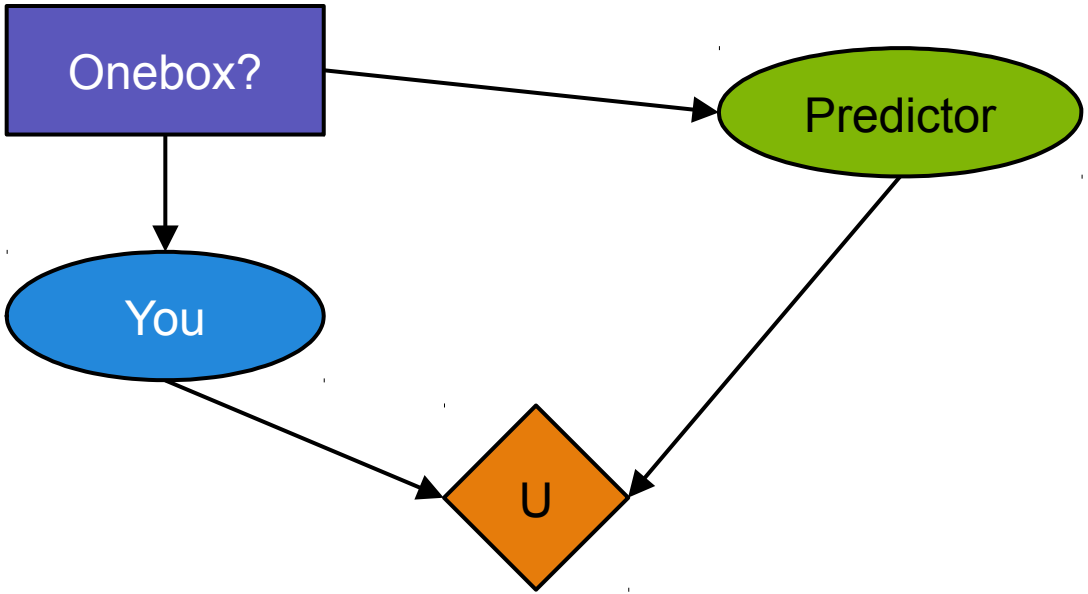
Prisoner’s dilemma vs clone



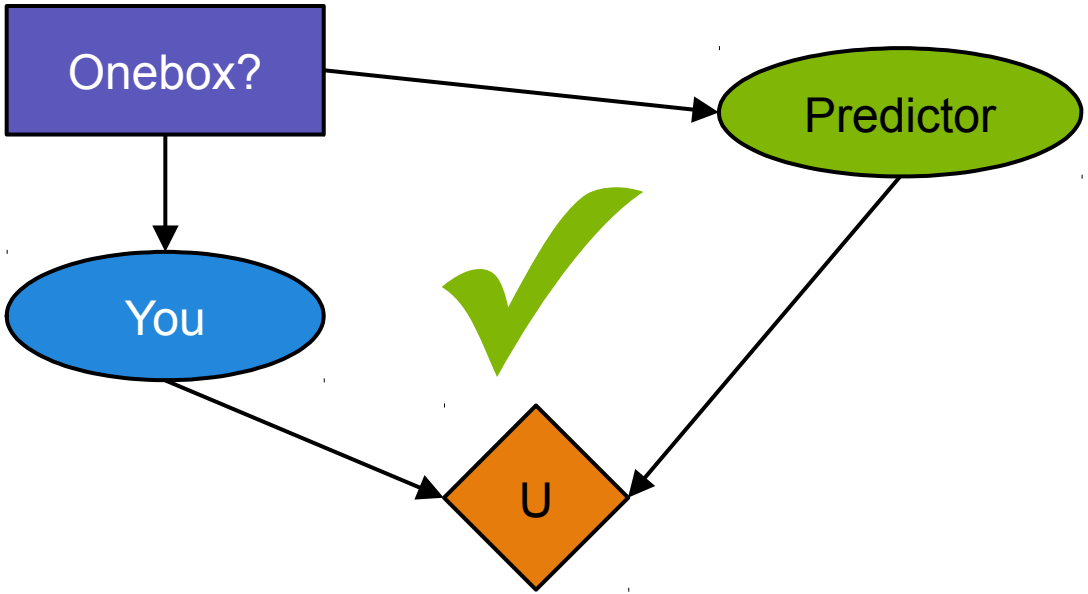
Newcomb’s problem



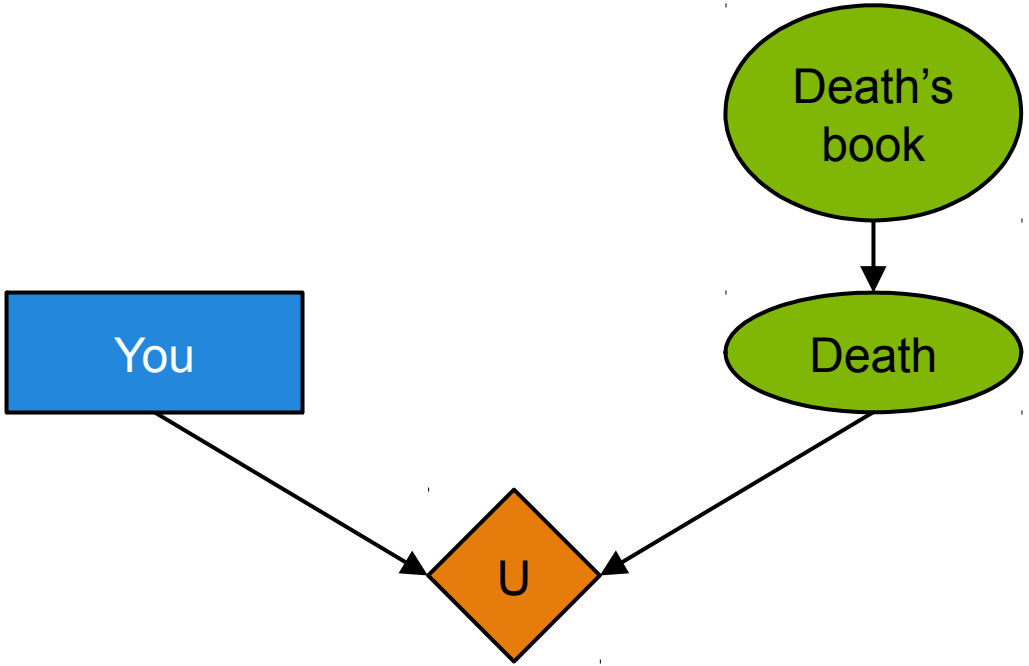
Newcomb's problem



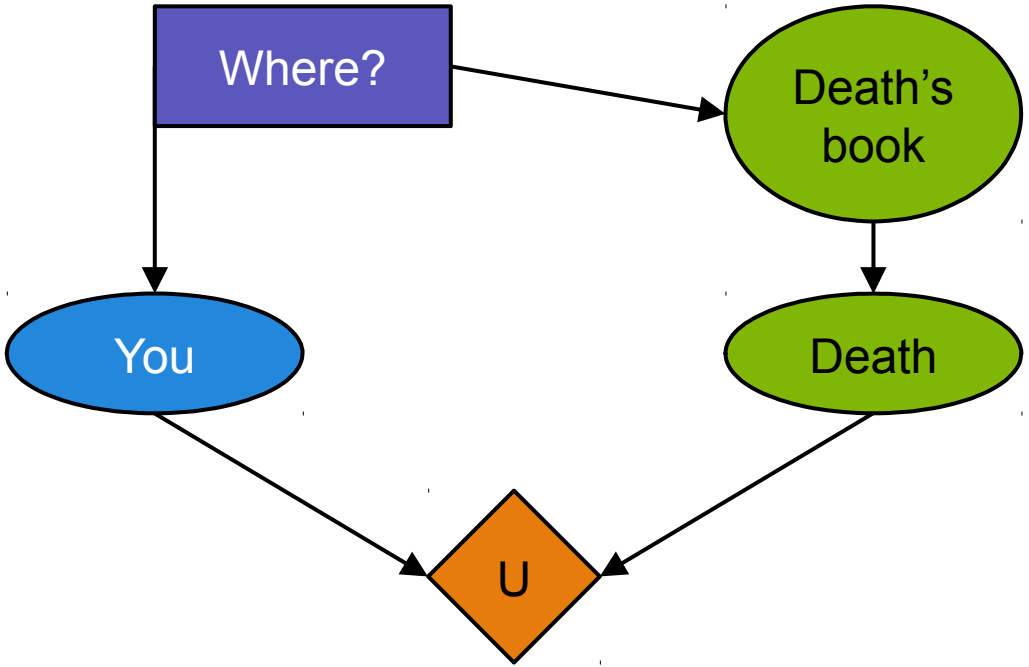
Newcomb's problem



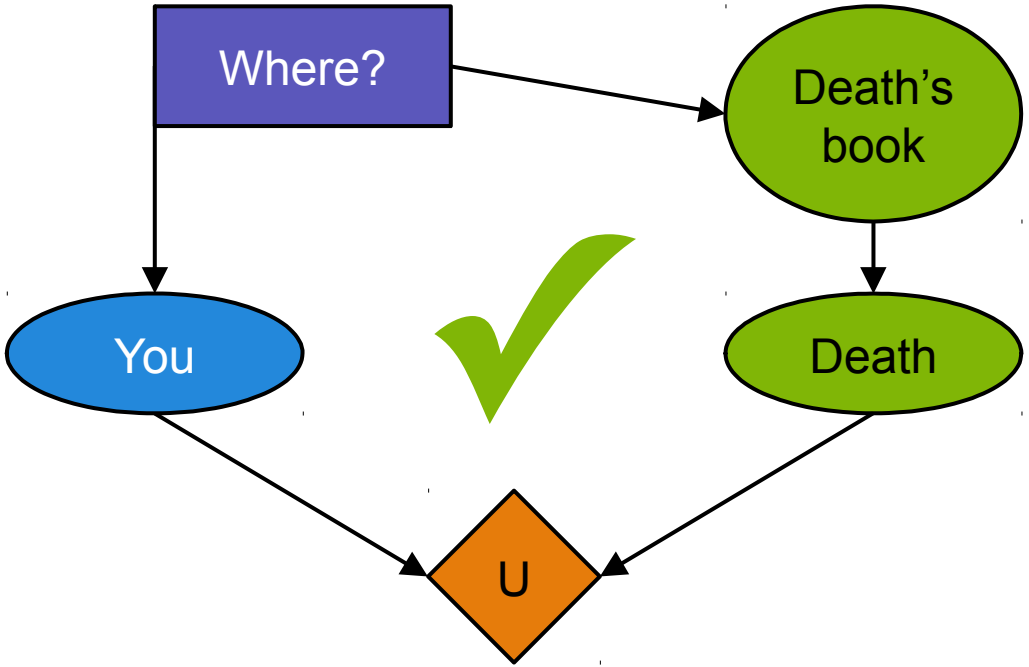
Death in Damascus



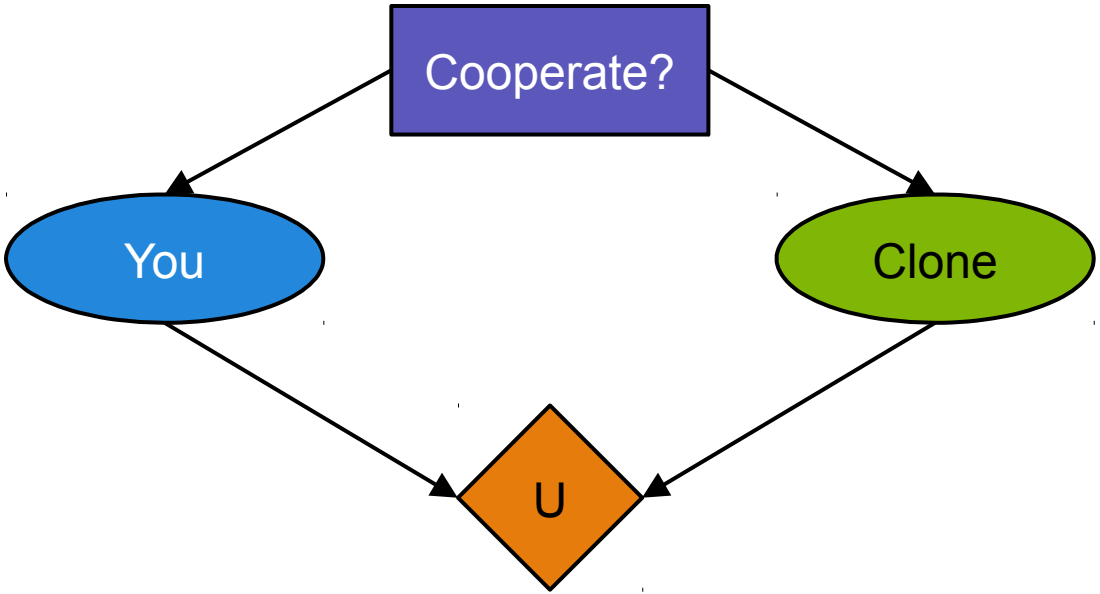
Death in Damascus



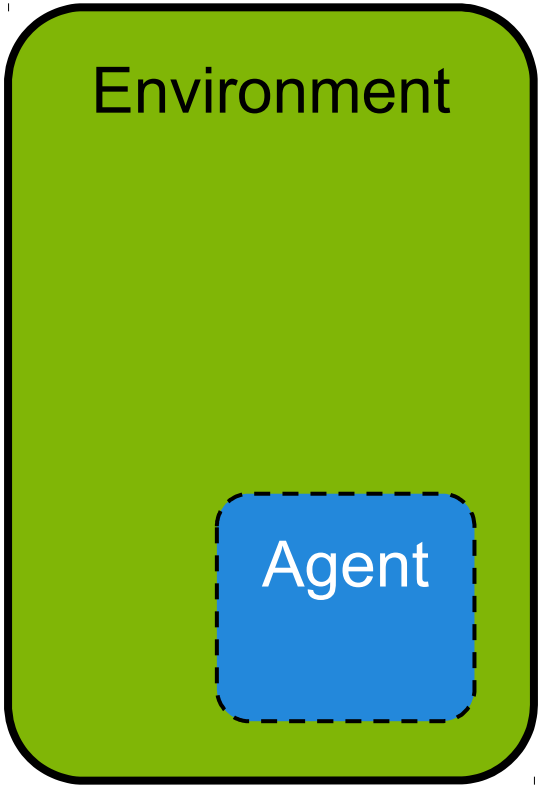
Death in Damascus

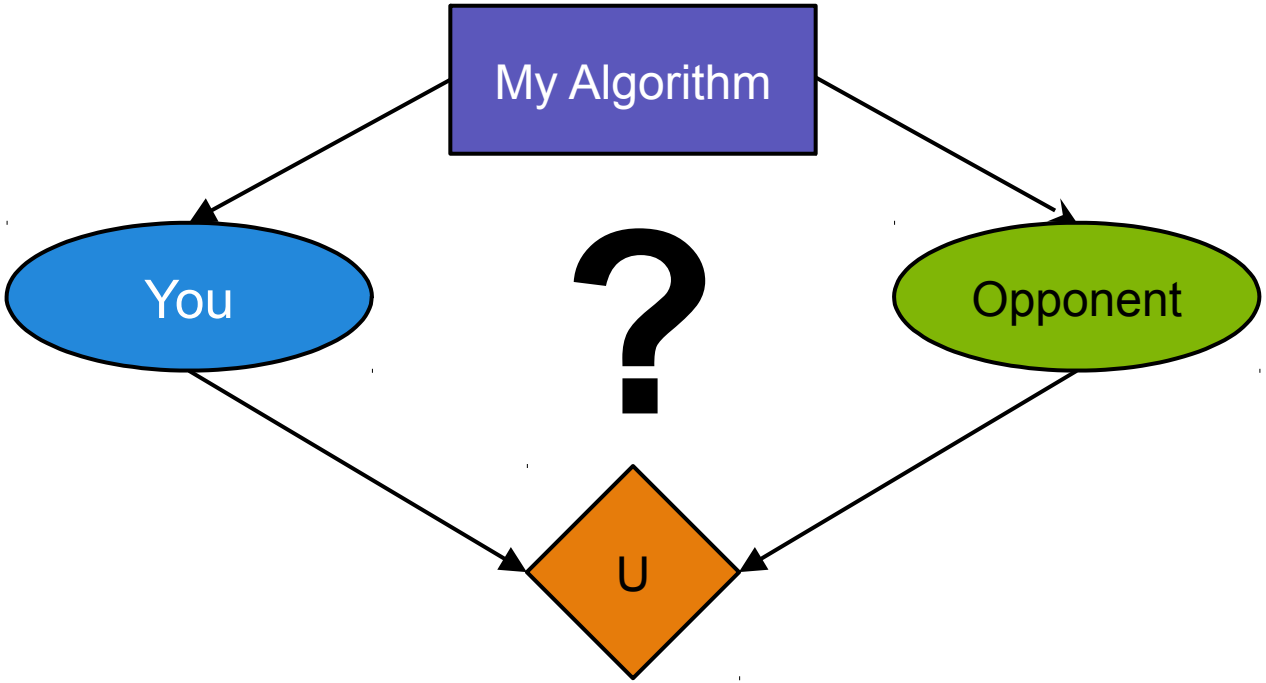


What if logical links had come first?



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


Environment
Agent


My Algorithm
You
?
Opponent
U

We need a notion of *logical* counterfactuals

If you know how assuming A()=a affects B() for arbitrary algorithms, you're *done*.

Counterpossibilities

If `Agent()` had output `a`, what output would `Environment()` have?

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]
```

Counterpossibilities

if I can prove
“Me()=π → Environment()=u”
...

Counterpossibilities

Counterpossibilities

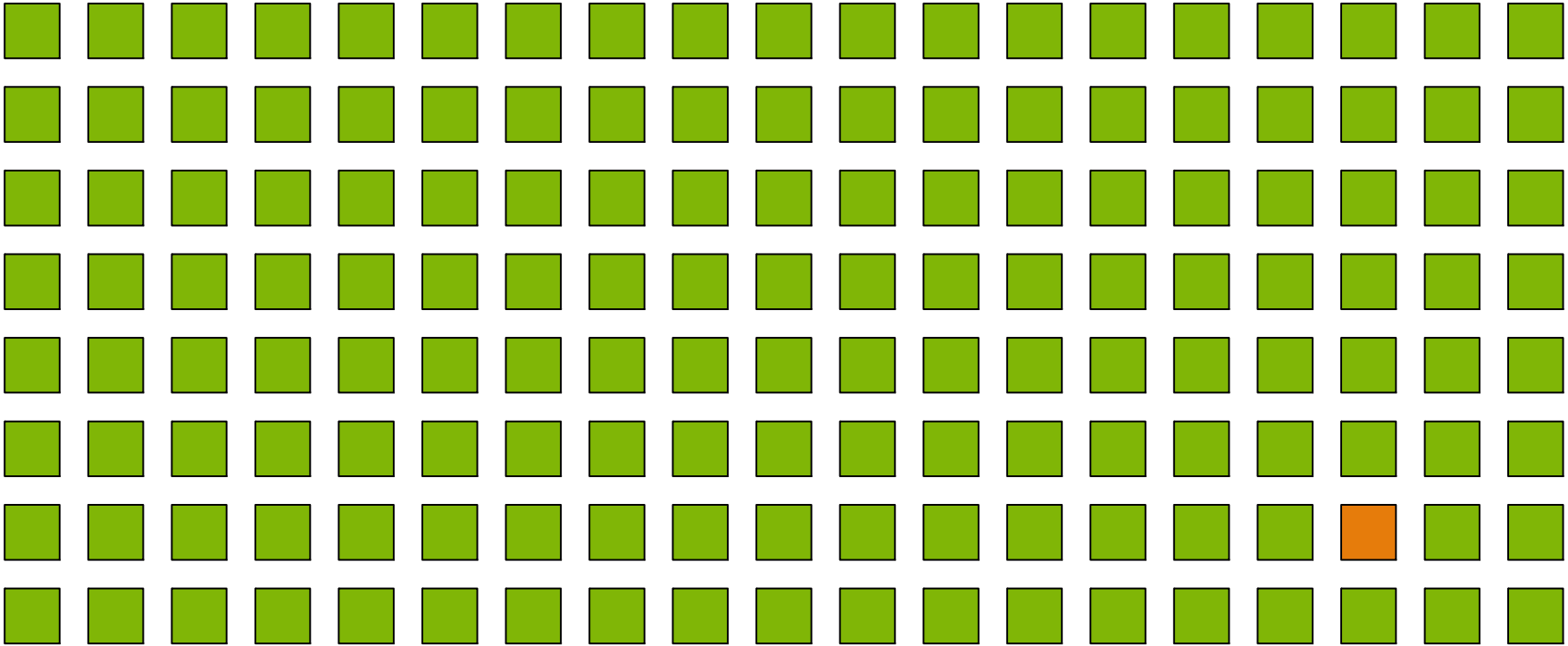
```
def PLDT():
  map = {}
  for π in policies:
    if PA can prove “PLDT()=π → Environment()=u”
         (for some u):
      map[π]=u
  return the π that maximizes map[π]
```

What is a what if?

Commandeer logical control

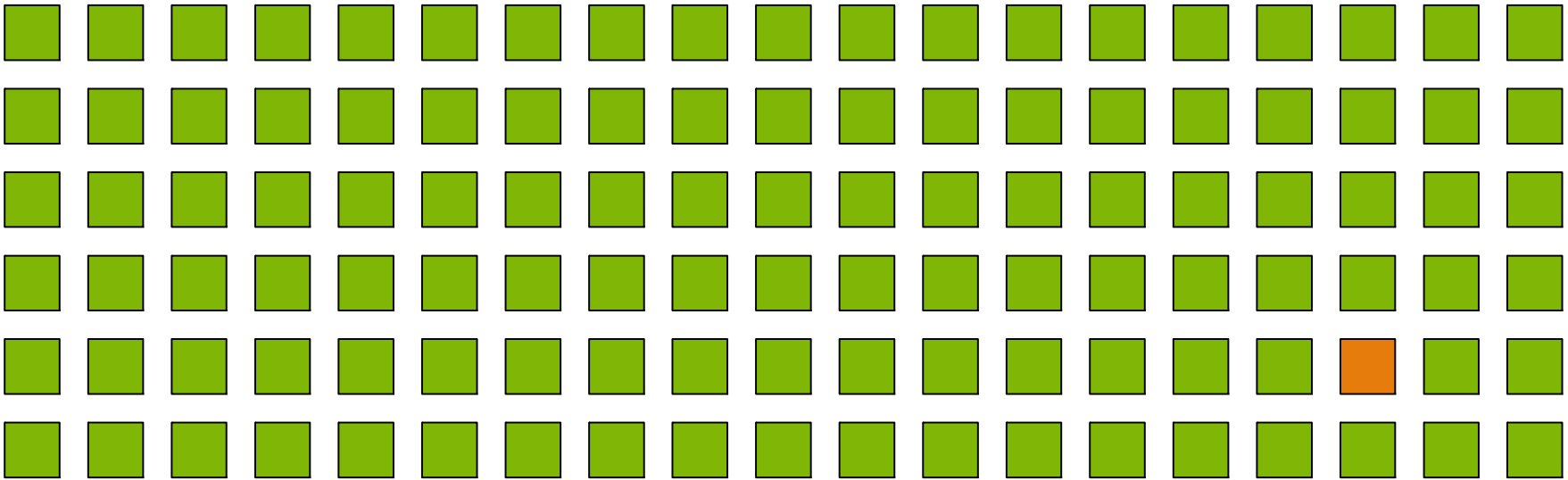
```
def PLDT():
  for π in policies:
    if PA can prove “PLDT()≠π”:
      return π
  map = {}
  for π in policies:
    if PA can prove “PLDT()=π → Environment()=u”
         (for some u):
      map[π]=u
  return the π that maximizes map[π]
```

The Mad Newcomb Problem



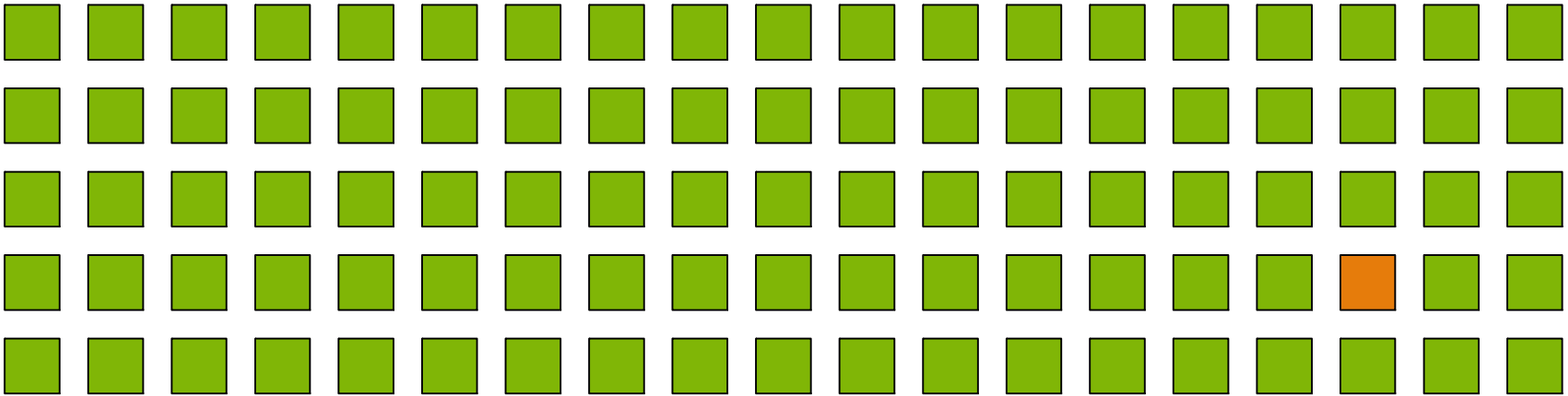
The Mad Newcomb Problem

1. Oh, they weren't reasoning about me at all.



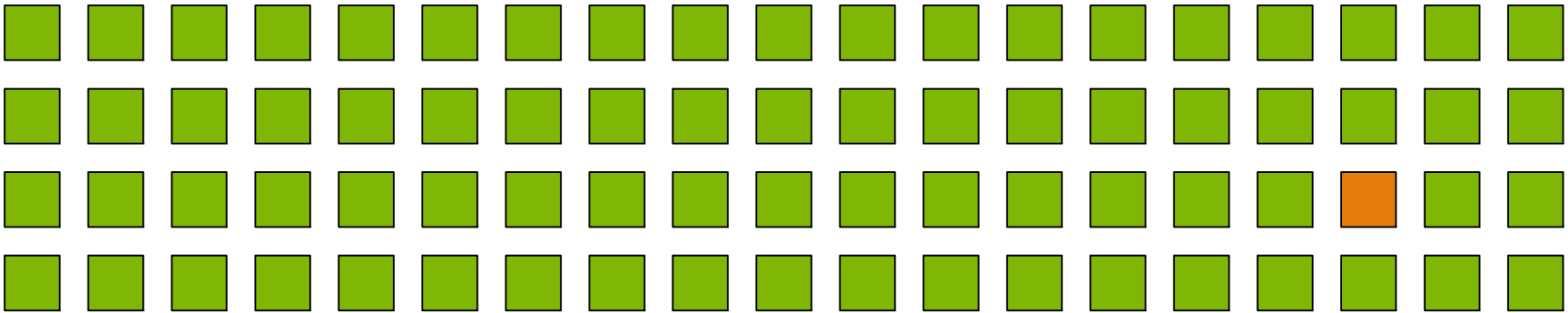
The Mad Newcomb Problem

1. Oh, they weren't reasoning about me at all.
2. Huh, this reasoning is flawed.



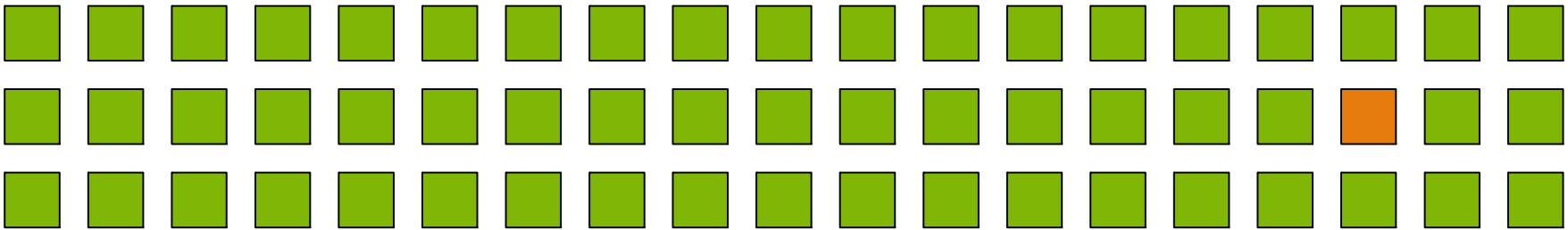
The Mad Newcomb Problem

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.



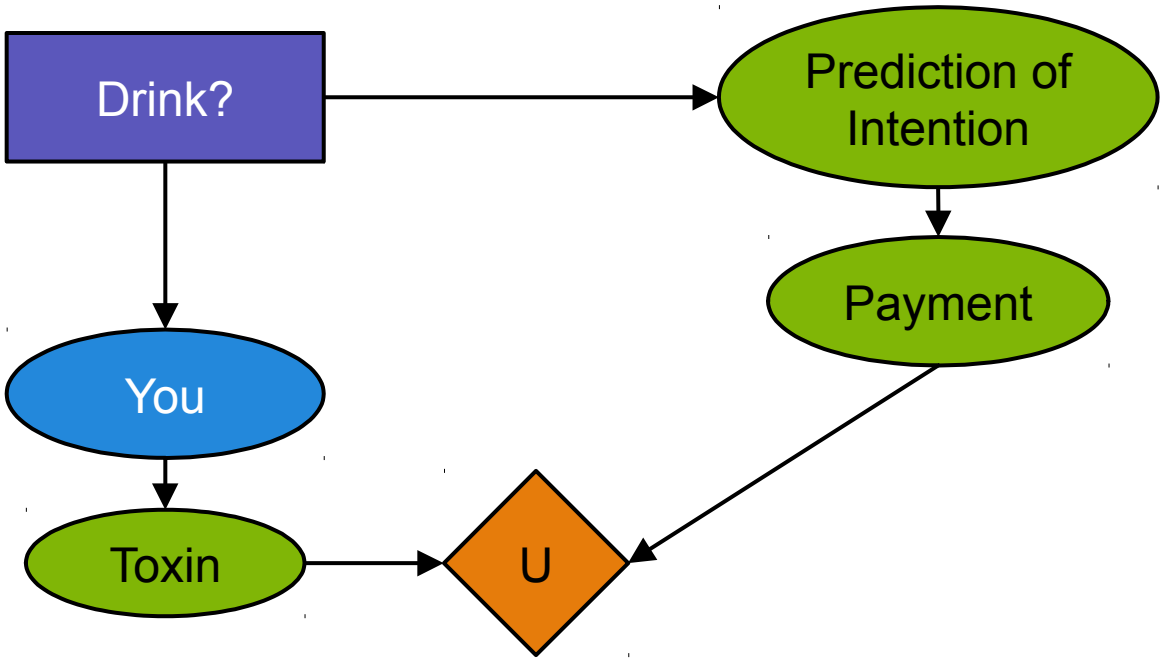
The Mad Newcomb Problem

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!

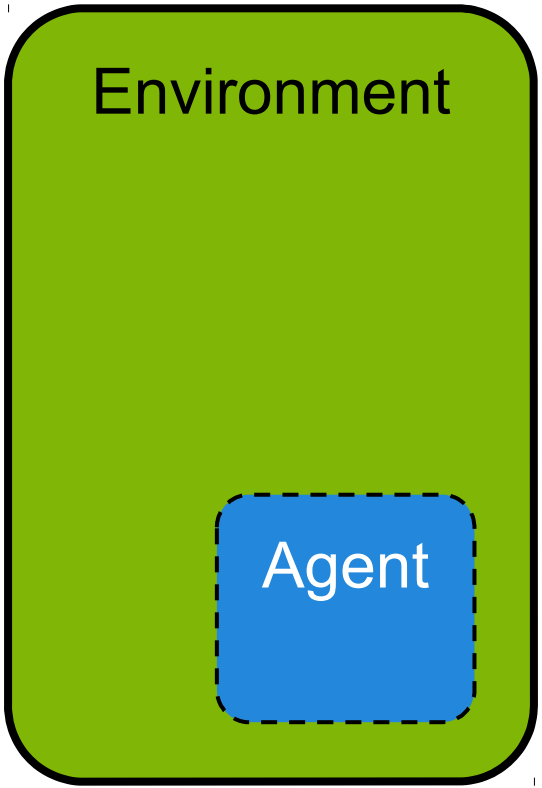


It *won't* work, but it will *would have* worked.

The Toxin Problem

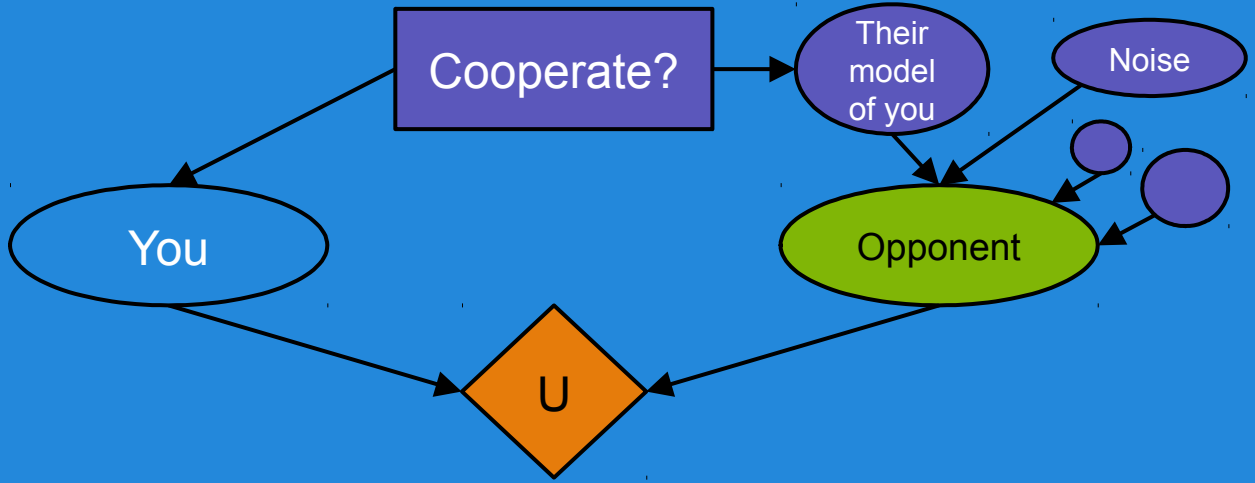


What exactly can you *logically* affect?


Environment
Agent

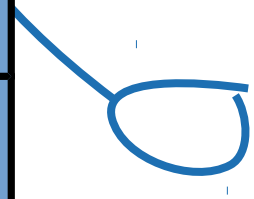
We need a better understanding of how to reason about what would happen if an algorithm did something it doesn’t.

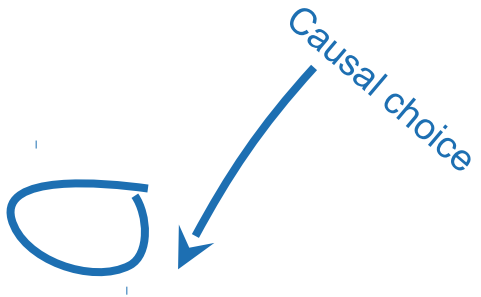
Which logical relationships do we respect, and how?



Preferences aren't enough

		Perfect Copy	
		Coope rate	Defect
You	Cooper ate	2 2	0 3
	Defect	3 0	1 1





What is a what if?