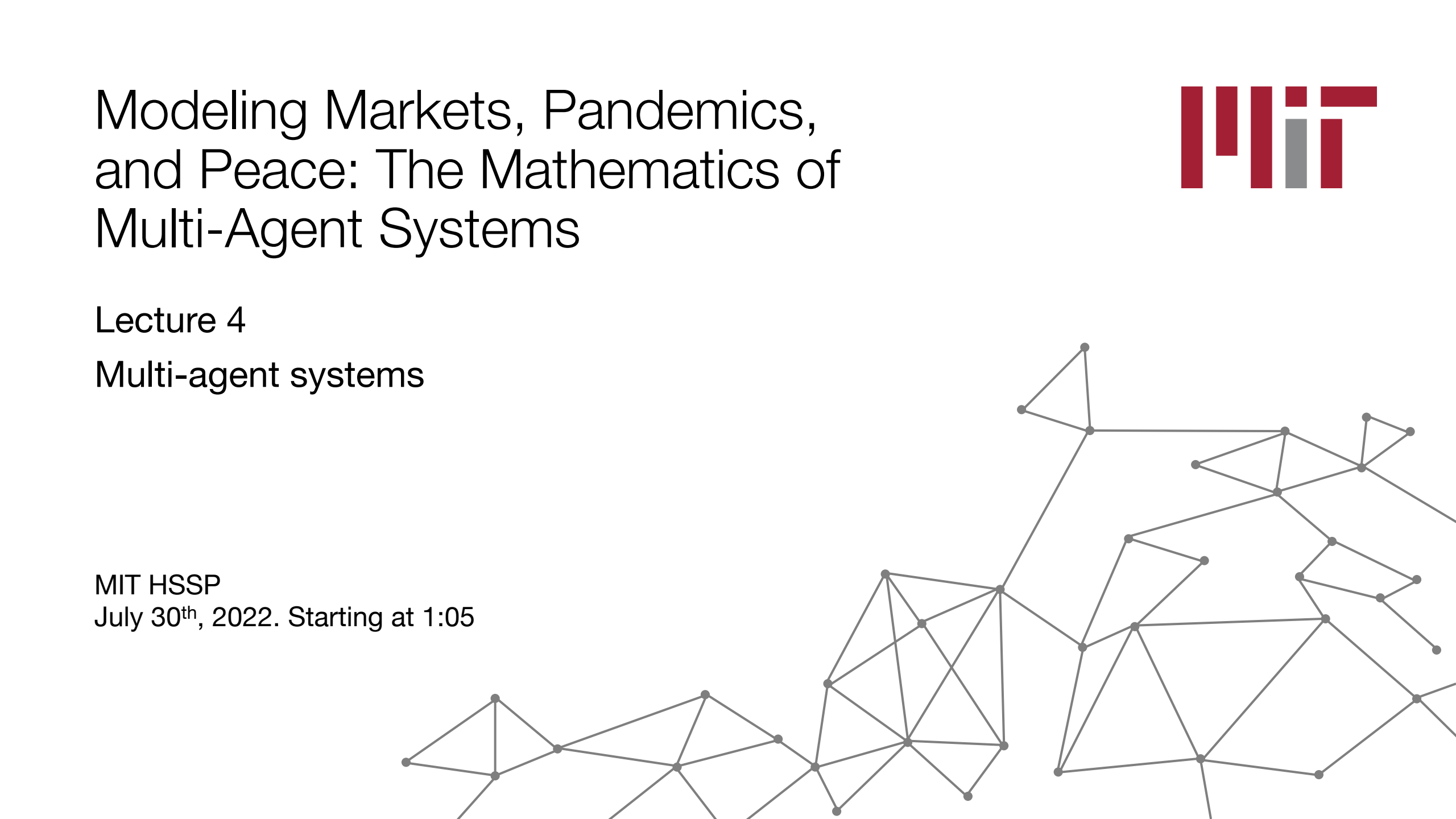
Modeling Markets, Pandemics, and Peace: The Mathematics of Multi-Agent Systems



Lecture 4

Multi-agent systems

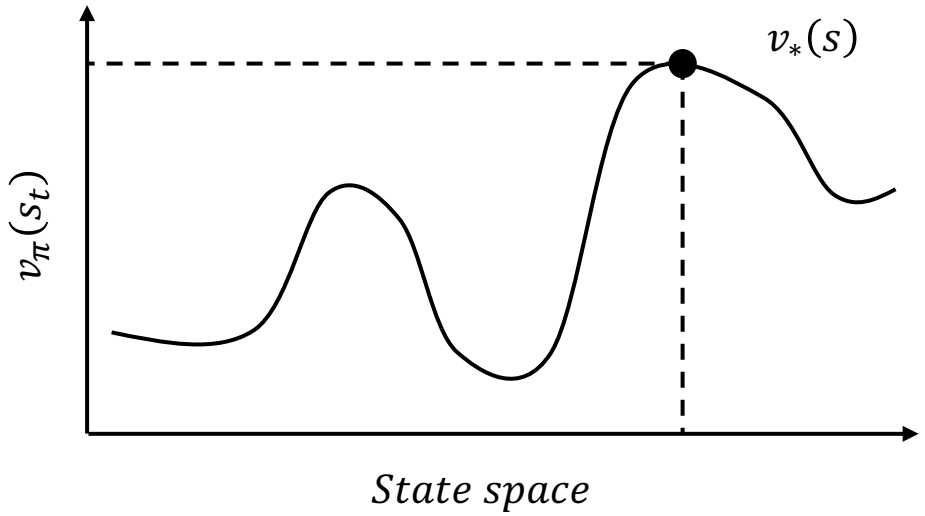
MIT HSSP
July 30th, 2022. Starting at 1:05

Recap: Expected utility hypothesis (RL)

We can program our reinforcement learning agent by telling it to maximize some utility,

$$v_\pi(s_t) = \mathbb{E}_\pi \left[\sum_{k=0}^{\infty} R_{t+k+1} | S = s_t \right] \qquad v_*(s) = \text{maximize}_\pi(v_\pi(s))$$

where we define the rewards (e.g., # apples eaten, money earned…) to shape its behavior.

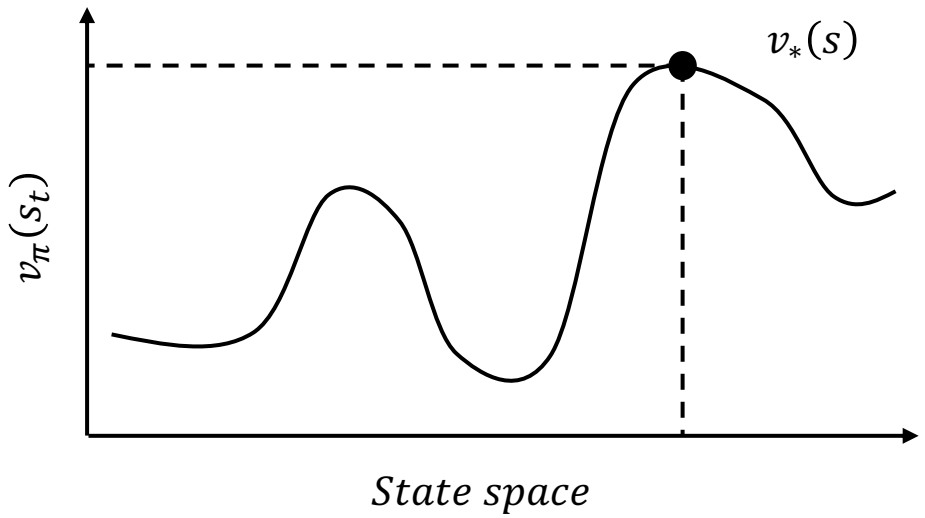


Recap: Expected utility hypothesis (social science)

To determine how actors behave, we can always define them to be utility maximizers

$$v_\pi(s_t) = \mathbb{E}_\pi\left[\sum_{k=0}^{\infty} u_{t+k+1} | S = s_t\right] \qquad v_*(s) = \text{maximize}_\pi(v_\pi(s))$$

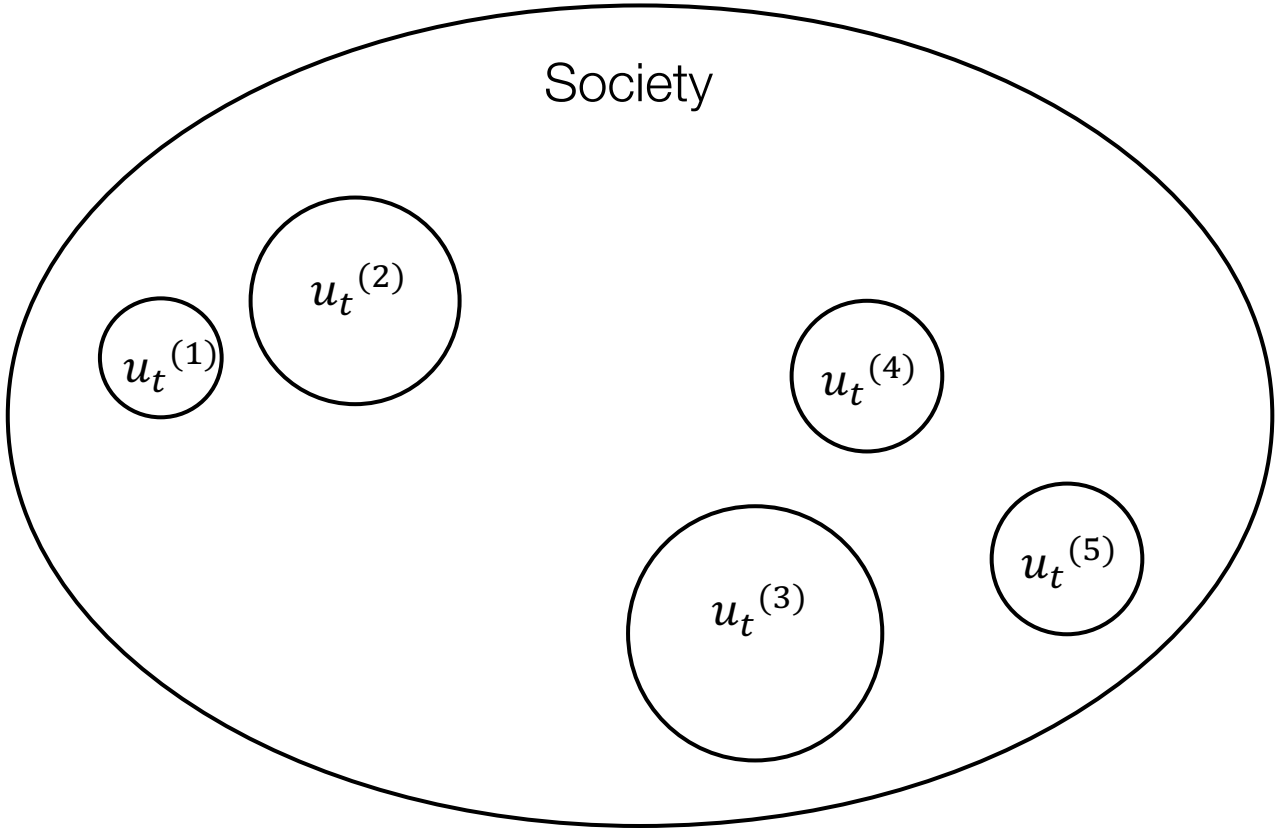
for some definition of $v_*(s)$ – as happiness, monetary value, or other rewards.



We saw that we can't always explain human behavior using this hypothesis…

… but we can mostly fix this by modifying either u_t or the definition of $v_\pi(s_t)$!

What about multiple agents?



Each individual i maximizes their own utility over time,

$$v_\pi^{(i)}(s_t) = \mathbb{E}_\pi \left[\sum_{k=0}^{\infty} u_{t+k+1}^{(i)} | S = s_t \right]$$

Suppose you were in charge of this society. What should you maximize?

- Sum of individual $u_t^{(i)}$'s?
- Something else?

Social welfare functions

Utilitarian welfare

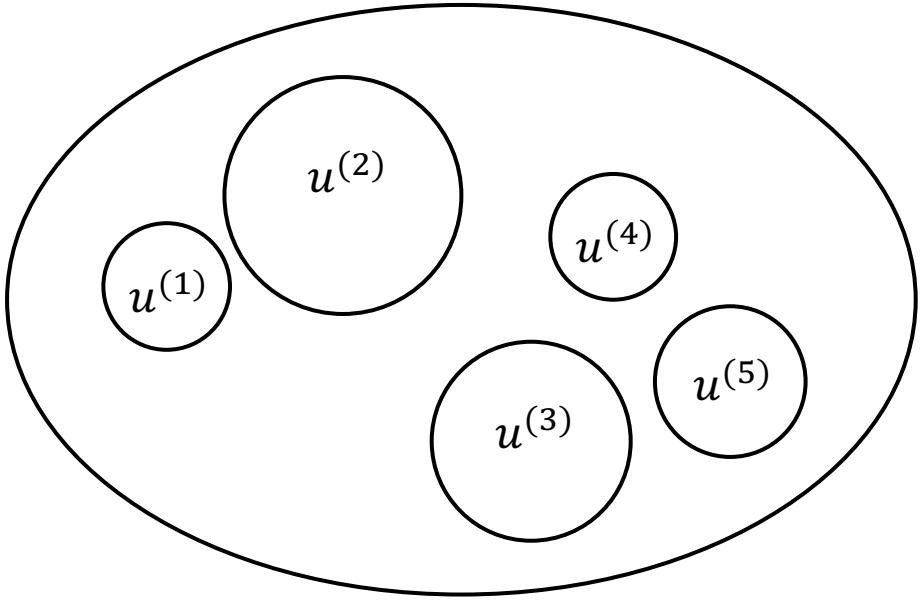
Maximize sum of individual utilities

$$U = \sum_i u^{(i)}$$

Rawlsian welfare

Maximize utility of least well-off individual

$$U = \min(u^{(1)}, u^{(2)}, \ldots, u^{(n)})$$



Someone may have to be worse off to maximize U

Pareto optimality

What if instead of maximizing some social utility U, our goal was to make people as well-off as possible without hurting anyone?

Pareto-optimal welfare: Outcome where no individual can be made better off without making at least one individual worse off.

Example: Suppose Alice and Bob are stranded on an island with 10 apples and 10 bananas. Alice prefers apples, but Bob is indifferent. Consider the following resource allocations:

- All the apples are given to Alice, and bananas to Bob. Is this Pareto-optimal? Yes.

- All the bananas are given to Alice, and apples to Bob. Is this Pareto-optimal? No.
  Alice can be made better off by trading one banana for Bob's apple.
  Bob's utility wouldn't change, so no one is worse-off.

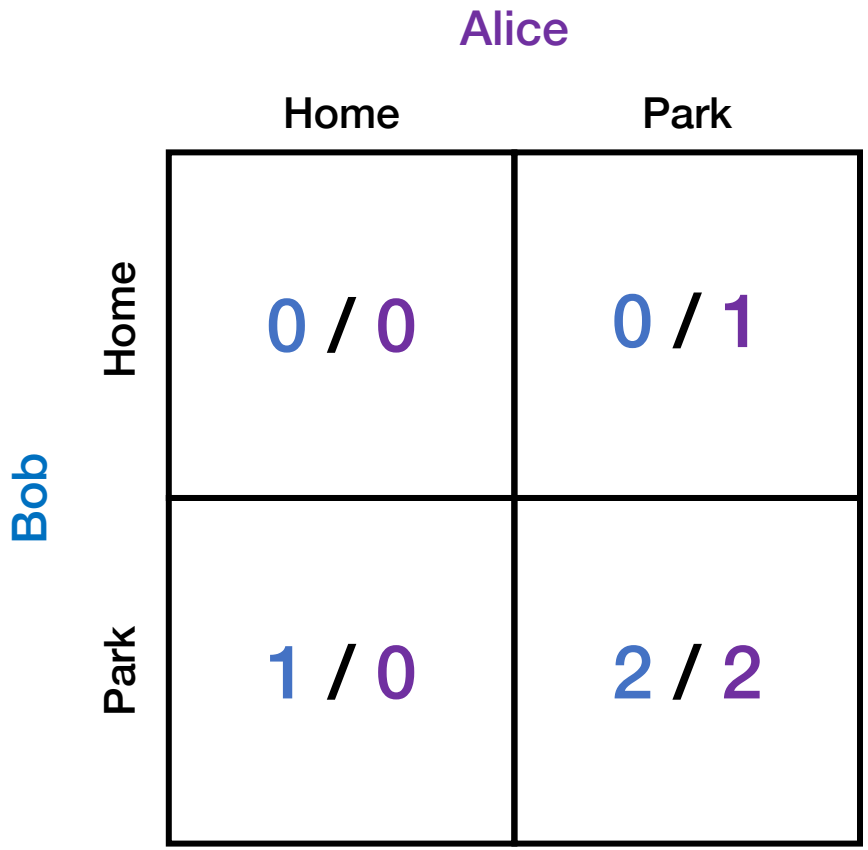
Introduction to game theory

Game theory is the study of multi-person decision problems (games).

Ex: “Meet-up game”
Suppose Alice and Bob want to meet up. They each can choose to either stay home (where they can’t see each other), or go to the park (where they might see each other). They each have the following utility function:

- Staying home = 0 utils
- Going to park alone = 1 util
- Going to park together = 2 utils

They cannot communicate their plans to each other ahead of time. What should they do?



Meet-up game (continued)

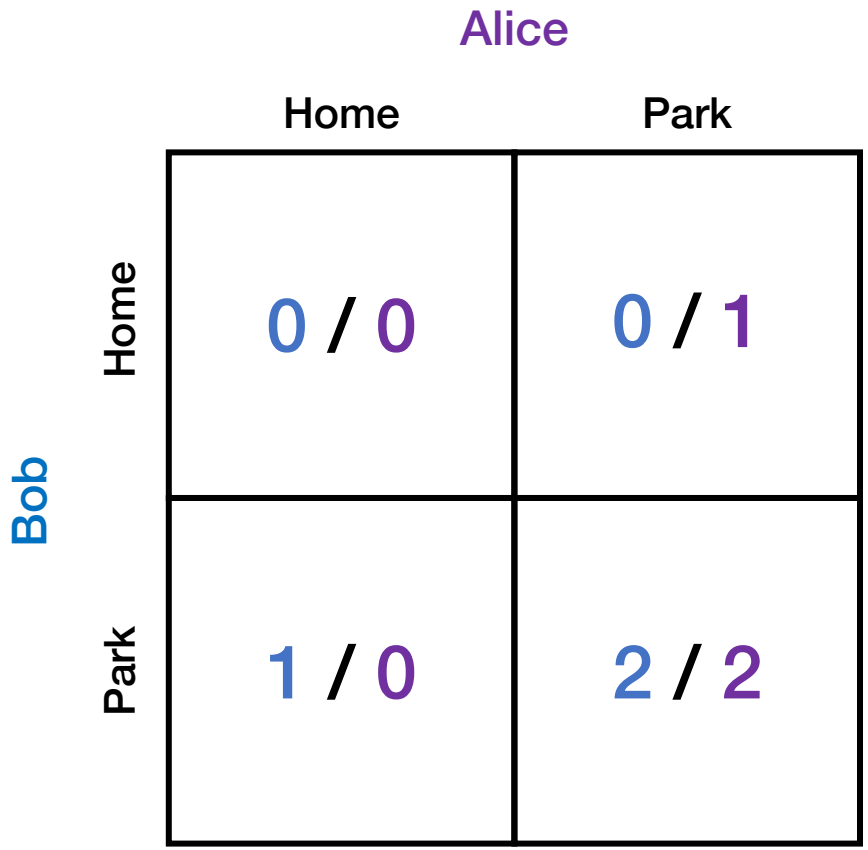
What is Alice's optimal strategy?
- If Bob chooses Home, she'd rather choose Park
- If Bob chooses Park, she'd rather choose Park

Alice should always choose Park

What is Bob's optimal strategy?
- If Alice chooses Home, he'd rather choose Park
- If Alice chooses Park, he'd rather choose Park

Bob should always choose Park

No player has any reason to deviate from (Park, Park).
Therefore, (Park, Park) is a **Nash Equilibrium.**



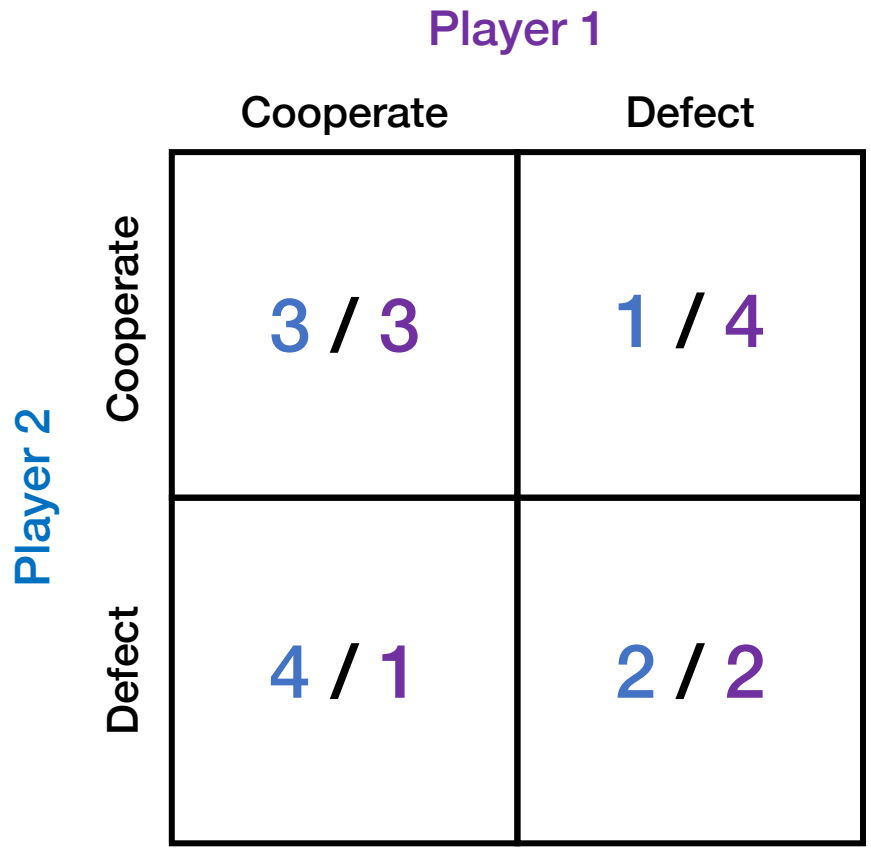
Prisoner's dilemma

Suppose two criminals have been arrested and they are held in separate cells so they can't communicate.

Each criminal has the choice to cooperate with the other criminal by keeping quiet, or defect and tell on the other criminal to the police. Each criminal has the following utility function:

- Both cooperate = 3 utils (i.e. save 3 years in prison)
- They cooperate but the other defects = 1 util
- They defect but the other cooperates = 4 utils
- Both defect = 2 utils

What are the Nash equilibria (i.e. outcomes where no player wants to deviate)?

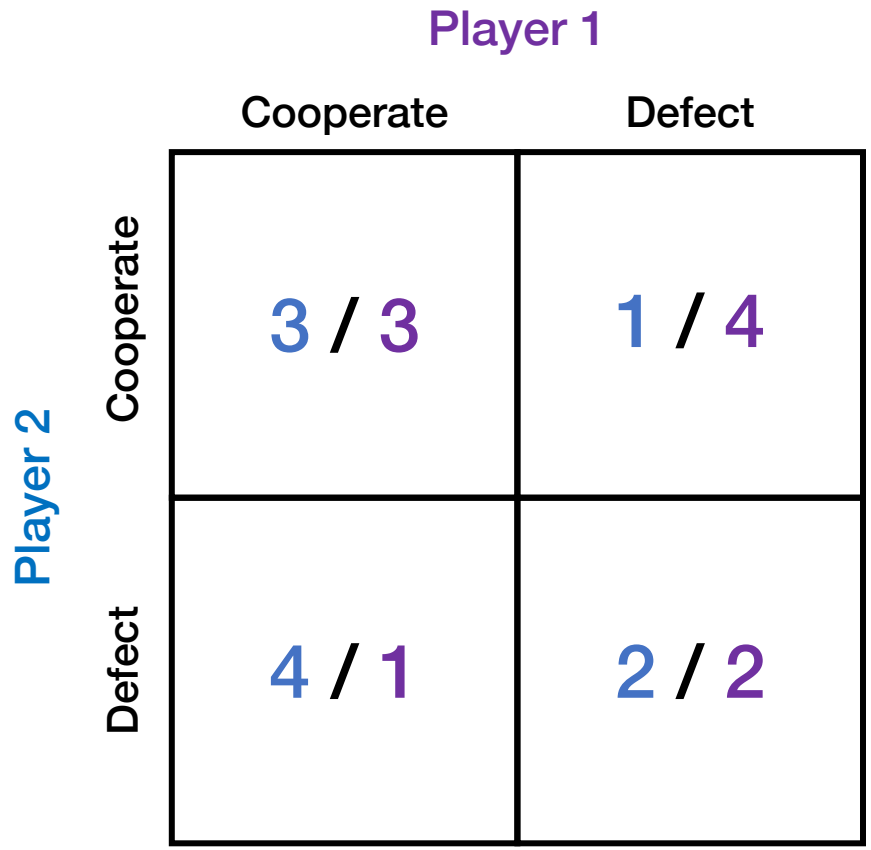


Prisoner's dilemma

What are the Nash equilibria?

- [Cooperate, Cooperate]:
  Since player 2 cooperates, player 1 would rather defect. Not a Nash equilibrium.
- [Cooperate, Defect]:
  Since player 1 cooperates, player 2 would rather defect. Not a Nash equilibrium.
- [Defect, Cooperate]:
  Since player 2 cooperates, player 1 would rather defect. Not a Nash equilibrium.
- [Defect, Defect]:
  Neither player would deviate $\Longrightarrow$ Nash equilibrium!

Note: Nash equilibrium $\neq$ Pareto optimality!
(i.e., individual optimum $\neq$ social optimum)

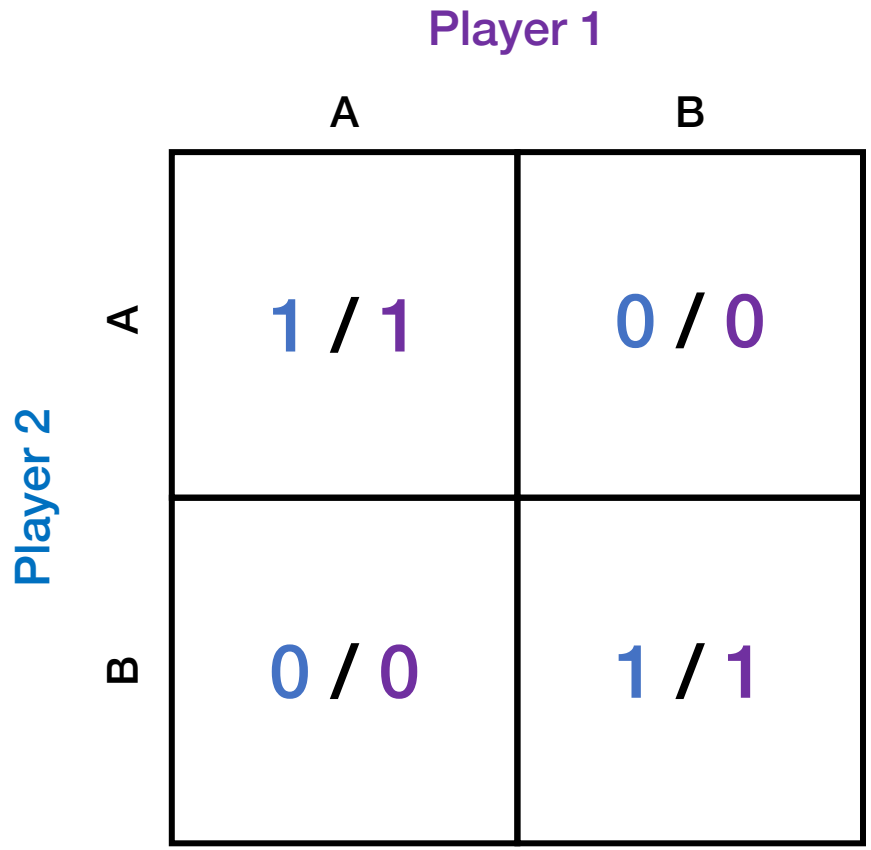


Coordination game

What are the Nash equilibria?

- [A, A]:
  Neither player would deviate ⟹ Nash equilibrium!
- [A, B]:
  Since player 2 chooses B, player 1 would rather choose B. Not a Nash equilibrium.
- [B, A]:
  Since player 2 chooses A, player 1 would rather choose A. Not a Nash equilibrium.
- [B, B]:
  Neither player would deviate ⟹ Nash equilibrium!

The two Nash equilibria are (A, A) and (B, B).

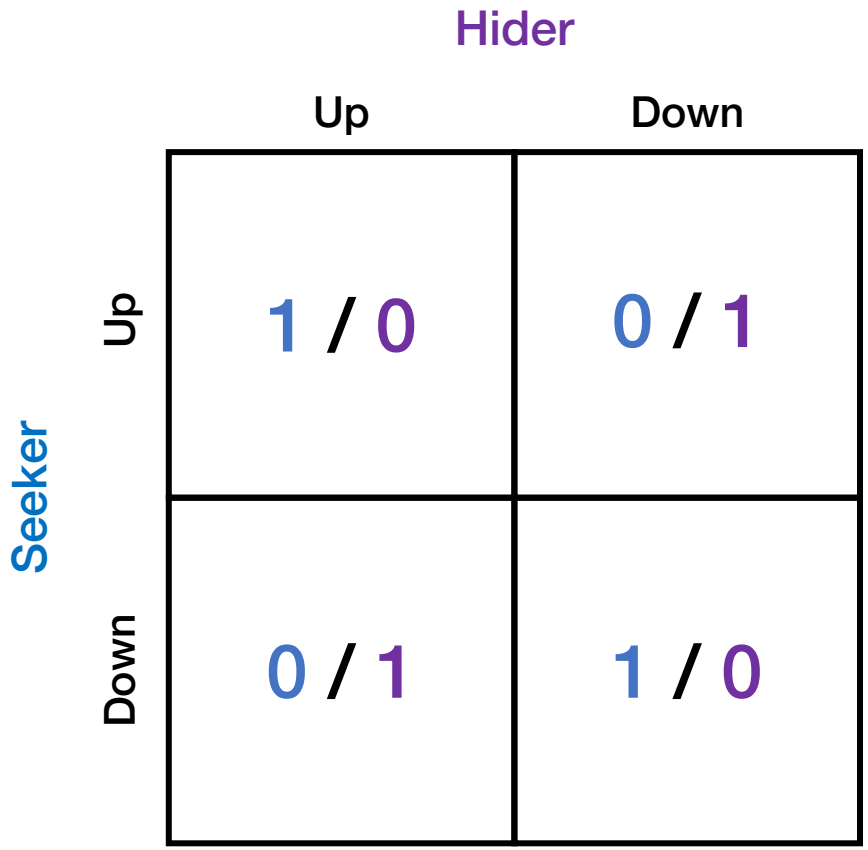


Hide-and-seek

What are the Nash equilibria?

- [Up, Up]:
  Since seeker goes Up, hider would rather go Down.
  Not a Nash equilibrium.
- [Up, Down]:
  Since hider goes Up, seeker would rather go Up.
  Not a Nash equilibrium.
- [Down, Up]:
  Since hider goes Down, seeker would rather go Up.
  Not a Nash equilibrium.
- [Down, Down]:
  Since seeker goes Down, hider would rather go Up.
  Not a Nash equilibrium.

What now?



Hide-and-seek

What if players could have **mixed strategies** (don't have to stick to one strategy 100% of the time)?

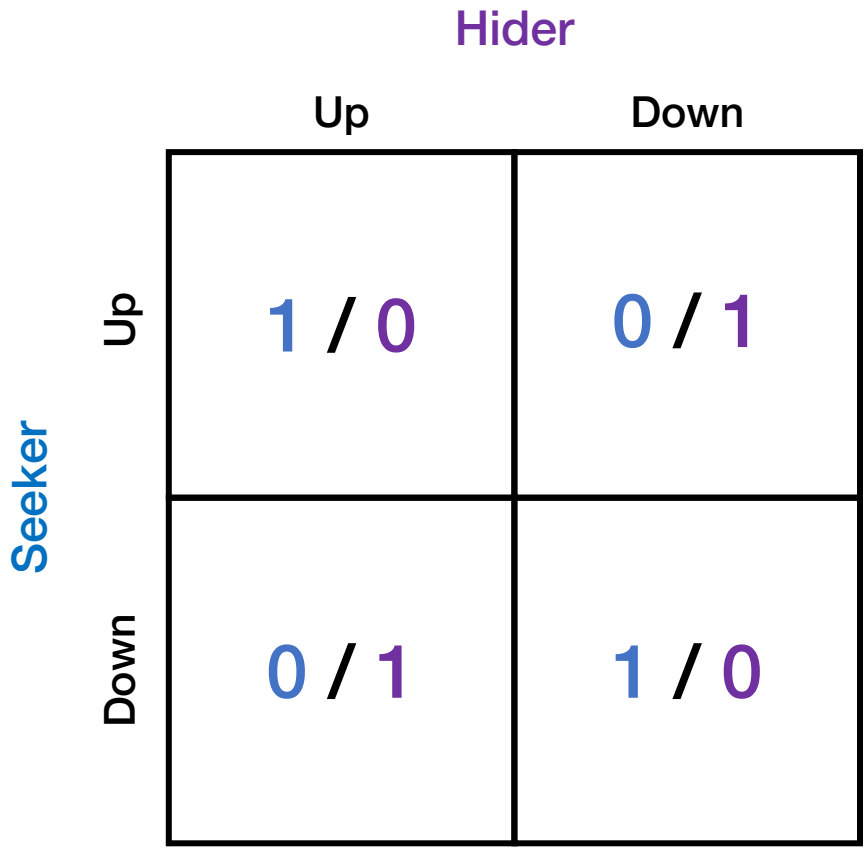
Ex: Suppose the hider's strategy was to go Up > 50% of the time, i.e. (0.75 Up, 0.25 Down).

Then the seeker will decide to always go Up.

But then the hider will decide to always go Down!

Both players end up deviating if they play one action over 50% of the time.

There are **no pure-strategy Nash equilibria**, but there is **one mixed-strategy Nash equlibirium**:
[(0.5 Up, 0.5 Down), (0.5 Up, 0.5 Down)]

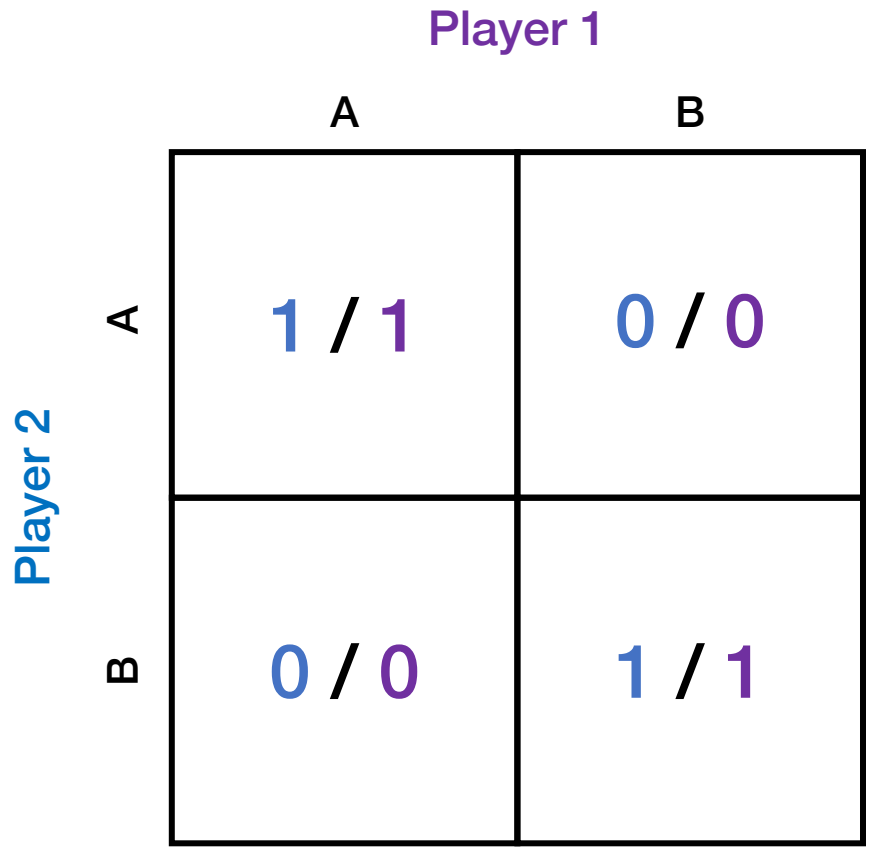


Recall: coordination game

What are the pure-strategy Nash equilibria?
[P1 = A, P2 = A]
[P1 = B, P2 = B]

What are the mixed-strategy Nash equilibria?
[P1 = (0.5A, 0.5B), P2 = (0.5A, 0.5B)]

Check: If P1 plays (0.5A, 0.5B), P2's expected utility is ½ no matter what strategy P2 uses (i.e. no reason to deviate).
By symmetry, P1 doesn't deviate either.



A 3-choice game

What are the pure-strategy Nash equilibria?
There are none.

What are the mixed-strategy Nash equilibria?
[P1 = ($\frac{1}{3}$ A, $\frac{1}{3}$ B, $\frac{1}{3}$ C), P2 = ($\frac{1}{3}$ A, $\frac{1}{3}$ B, $\frac{1}{3}$ C)]

What game does this remind you of?

Rock-paper-scissors

Player 2 \ Player 1	Rock	Paper	Scissors
Rock	0 / 0	-1 / 1	1 / -1
Paper	1 / -1	0 / 0	-1/ 1
Scissors	-1 / 1	1 / -1	0 / 0

Nash's existence theorem

Theorem: Every finite game (i.e. each player's strategy set is finite) has a pure- or mixed-strategy Nash equilibrium.

Proof: Beyond the scope of this class.

There is always a stable solution to a finite non-cooperative game that does not require external enforcement!

Why does cooperation arise in the real world?

Tragedy of the commons

Player 2 \ Player 1	Cut emissions	No action
Cut emissions	3 / 3	1 / 4
No action	1 / 4	2 / 2

Country A

Country B

In the Nash equilibrium, neither countries take action and keep increasing emissions until we reach climate catastrophe.

Are we doomed by the structure of the game? Does this mean that the only rational play is the Nash equilibrium, so we can at least get the best out of the situation?

An incomplete picture

Elinor Ostrom, Nobel Prize in Economics (2009)

People in local communities manage shared resources, like fishing waters, pastures, and forests all the time! They do so by establishing mechanisms and cooperative rules to stop the degradation of nature.

1) People can build commitment devices and institutions to guard resources

2) Games are not one-off; prolonged interactions encourage cooperation

3) We can couple games to other issues (e.g., in negotiations) to ensure cooperation

Rational choice theory is appealing because it shows that we are helpless and morally inculpable. It argues that humans are not evil but merely caught in the tragedy of politics.

Game theory is not wrong, rather, they “are special models that utilize extreme assumptions rather than general theories.” – Ostrom, Governing the Commons

Recall from lecture 3: utility discounting

People have **present-bias**: they put more value on immediate reward than long-term reward



We modelled this via **quasi-hyperbolic discounting,** where the agent maximizes

$$u(r_0) + \beta\delta u(r_1) + \beta\delta^2 u(r_2) + \beta\delta^3 u(r_3) \ldots = \sum_{t=0}^{\infty} \beta\delta^t u(r_t),$$

where $\delta \leq 1$is the **long-term discount factor** (usually ≤ 1),

and $0 \leq \beta \leq 1$ is the **short-term discount factor.**

What if the agent was able to commit to a decision about the future today?

Example: quasi-hyperbolic discounting

Sihao has 3 days to complete slides for his HSSP class: $t = 0,1,2$. The instant cost of completing the slides increases each day as follows:

- Cost at $t = 0$: -18 utils
- Cost at $t = 1$: -24 utils
- Cost at $t = 2$: -30 utils

If he hasn't completed the slides during days $t = 0$ or $t = 1$, he must complete them on day $t = 2$.

Suppose Sihao is a quasi-hyperbolic discounter with $\delta = 1$, $\beta = \frac{1}{2}$, maximizing $u_0 + \frac{1}{2}u_1 + \frac{1}{2}u_2$.

When would he complete the slides?

t=0

$$u_{\text{complete at t=0}} = -18 + \frac{1}{2}(0) + \frac{1}{2}(0) = -18$$

$$u_{\text{complete at t=1}} = 0 + \frac{1}{2}(-24) + \frac{1}{2}(0) = \boxed{-12}$$

$$u_{\text{complete at t=2}} = 0 + \frac{1}{2}(0) + \frac{1}{2}(-30) = -15$$

t=1

$$u_{\text{complete at t=1}} = -24 + \frac{1}{2}(0) = -24$$

$$u_{\text{complete at t=2}} = 0 + + \frac{1}{2}(-30) = \boxed{-15}$$

t=2

$$u_{\text{complete at t=2}} = \boxed{-30}$$

Time-dependent game

At t=0, Sihao is playing a game with his future self, maximizing discounted utility: $u_0 + \frac{1}{2}u_1 + \frac{1}{2}u_2$.

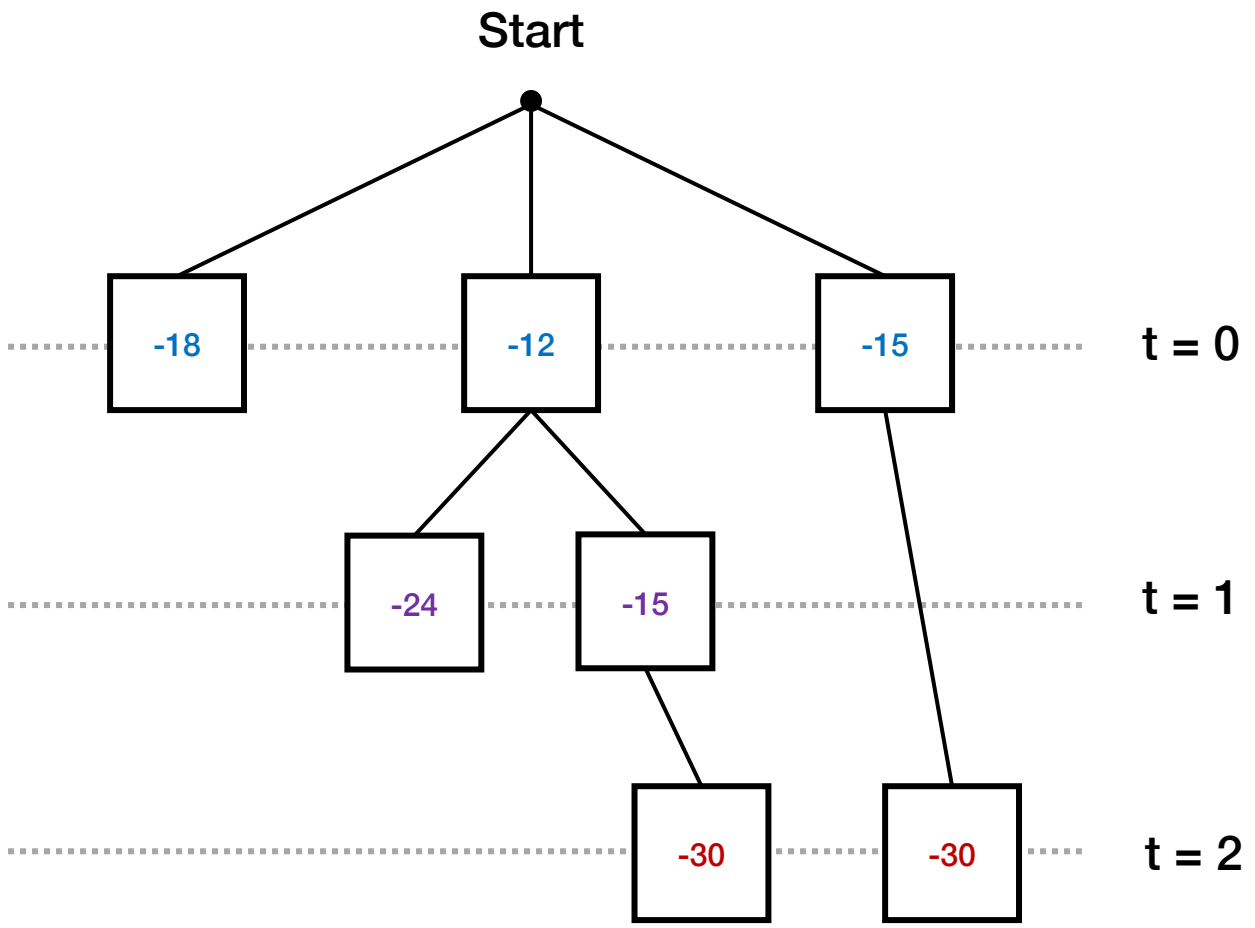
- Cost at $t = 0$: -18 utils
- Cost at $t = 1$: -24 utils
- Cost at $t = 2$: -30 utils

What is the Nash equilibrium?

(t=1, t=2, t=2) ⟹ -30 utils

What is the Pareto-optimal outcome?

(t=0, …, …) ⟹ -18 utils



There are three players (t = 0, 1, 2) each optimizing for themselves

Example: quasi-hyperbolic discounting

Sihao has 3 days to complete slides for his HSSP class: $t = 0,1,2$. The instant cost of completing the slides increases each day as follows:

- Cost at $t = 0$: -18 utils
- Cost at $t = 1$: -24 utils
- Cost at $t = 2$: -30 utils

If he hasn't completed the slides during days $t = 0$ or $t = 1$, he must complete them on day $t = 2$.

Suppose Sihao is a quasi-hyperbolic discounter with $\delta = 1$, $\beta = \frac{1}{2}$, maximizing $u_0 + \frac{1}{2}u_1 + \frac{1}{2}u_2$.

Now suppose he can get Julia to force him to complete the slides whenever he decides at t=0.

t=0

$$u_{\text{complete at t=0}} = -18 + \frac{1}{2}(0) + \frac{1}{2}(0) = -18$$

$$u_{\text{complete at t=1}} = 0 + \frac{1}{2}(-24) + \frac{1}{2}(0) = \boxed{-12}$$

$$u_{\text{complete at t=2}} = 0 + \frac{1}{2}(0) + \frac{1}{2}(-30) = -15$$

t=1

Julia forces him to complete them at t=1

$$u_{\text{complete at t=1}} = -24 + \frac{1}{2}(0) = \boxed{-24}$$

~~$u_{\text{complete at t=2}} = 0 + + \frac{1}{2}(-30) = -15$~~

t=2

Commitment devices

A **commitment device** is a way for an agent to restrict their future choice set, typically by making certain choices more expensive.

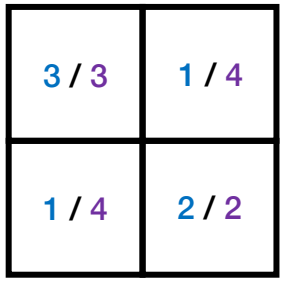
Ex: Buying long-term gym membership

Ex: Buying junk food in small packages rather in bulk

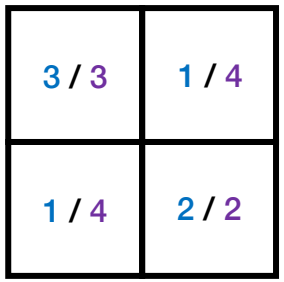
Ex: Rotating Savings and Credit Associations (peer-to-peer banking)

Finite repeated prisoner's dilemma

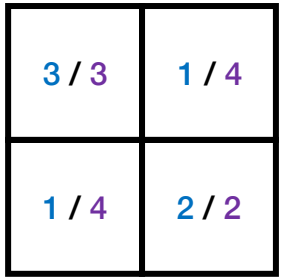
If we cannot have cooperation in the prisoner's dilemma, what if it is played multiple times? Players should, in principle, consider future games when deciding their initial strategy.



Game 1

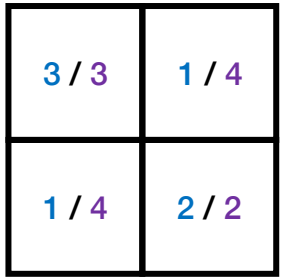


Game 2



Game 3

.....

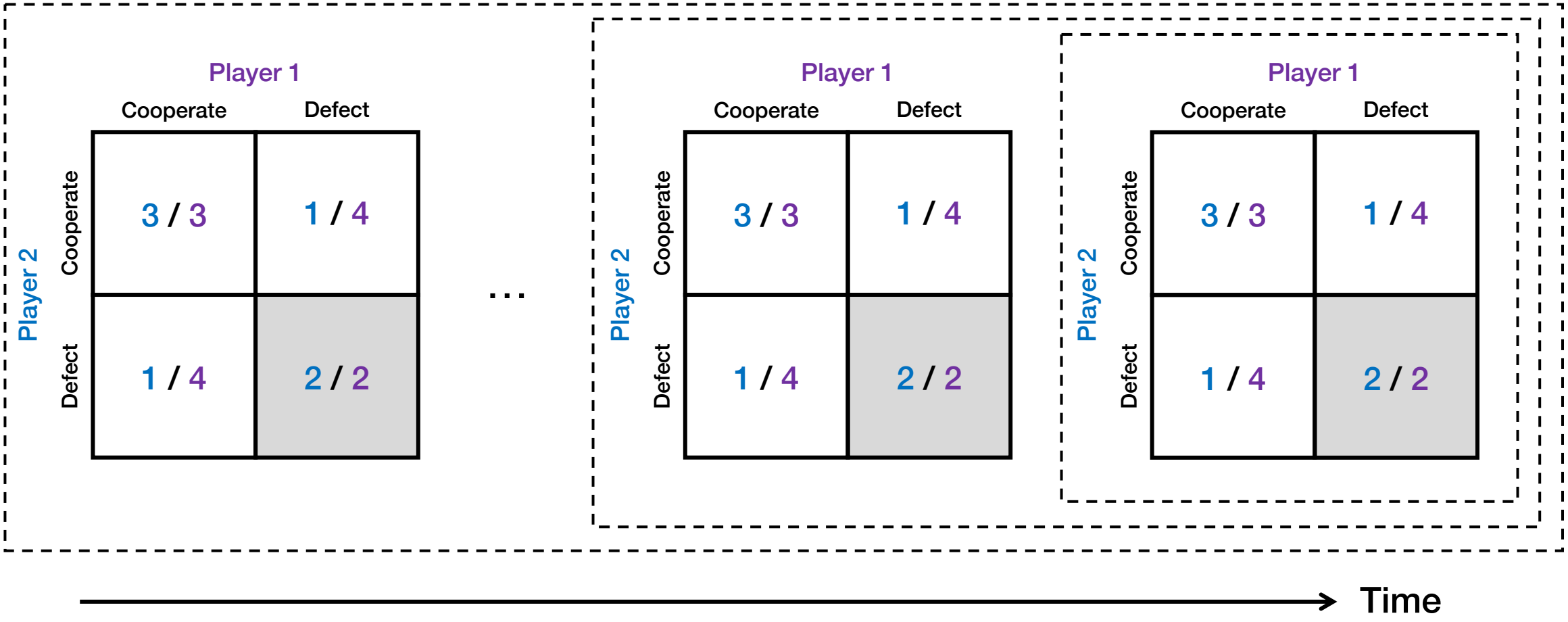


Final game

Theorem: in the final subgame, players must play a Nash equilibrium. Since all prior payoffs are locked in, players should only worry about optimizing their actions for that game.

Bellman's equation	"If I know the shortest path from Boston to DC runs through New York, then once I get to New York, I should just follow the shortest path from New York to DC."

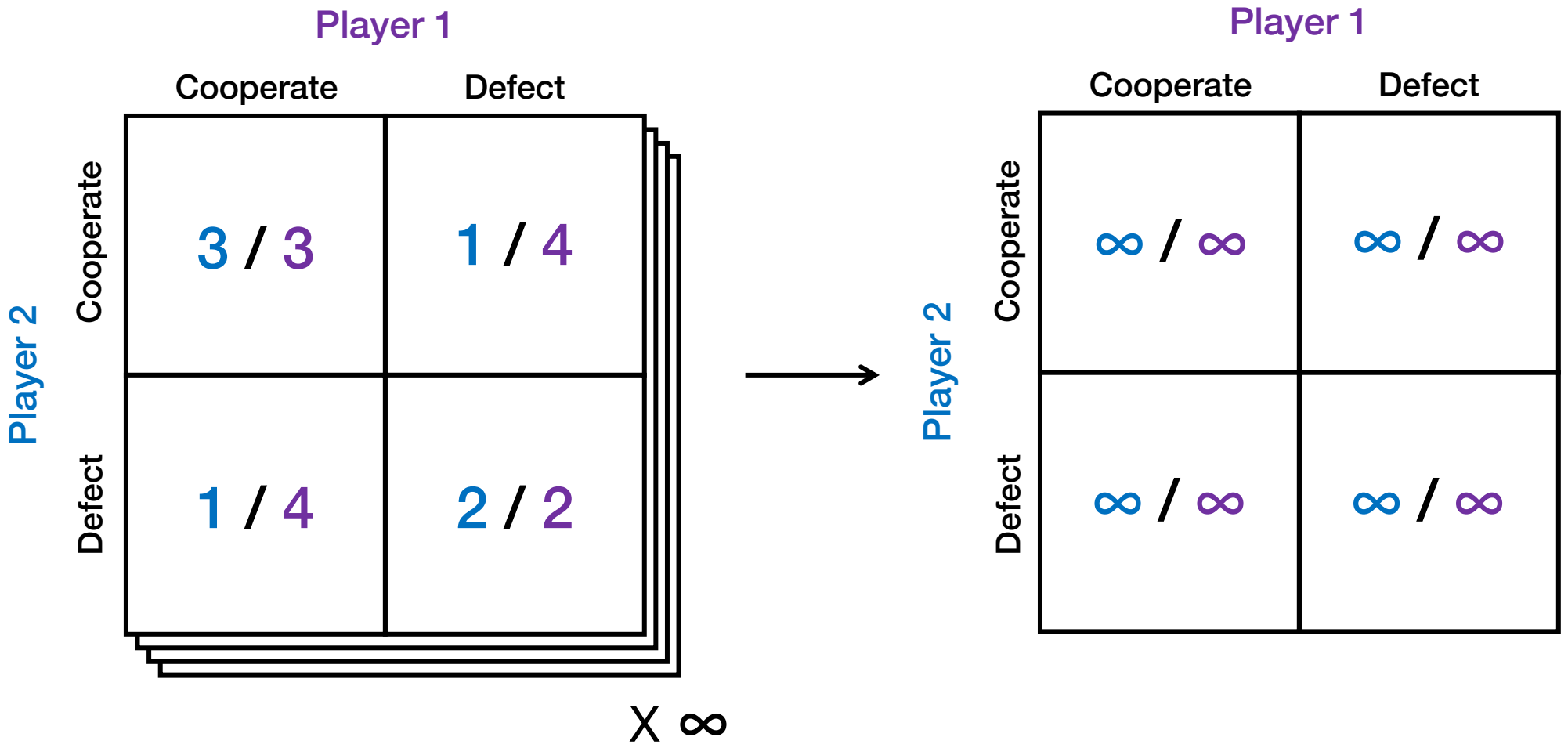
Finite repeated prisoner's dilemma



No cooperation possible! Being locked into mutual defection at the last stage poisons the well.

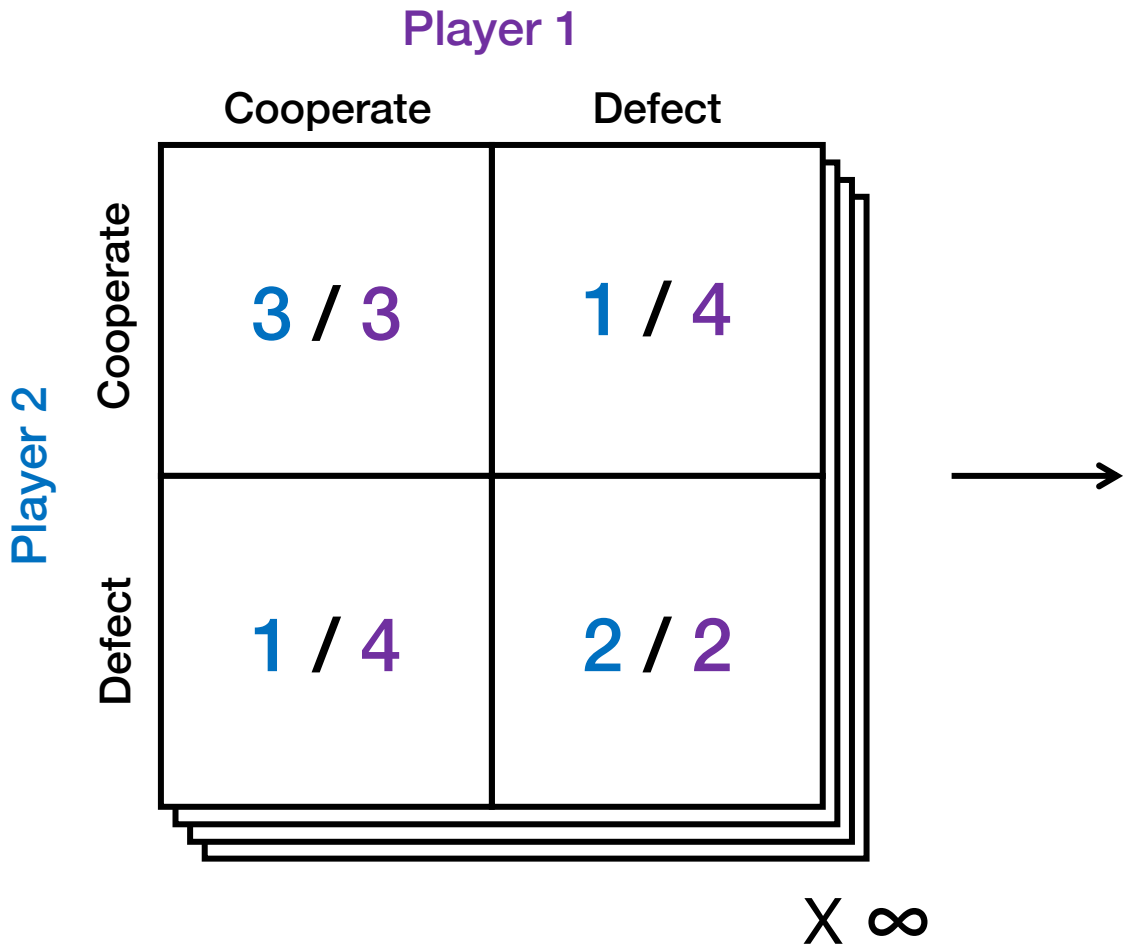
Infinite repeated prisoner's dilemma

Instead of playing the game a known, finite number of turns, let's create an uncertain future



Infinite repeated prisoner's dilemma

Instead of playing the game a known, finite number of turns, let's create an uncertain future

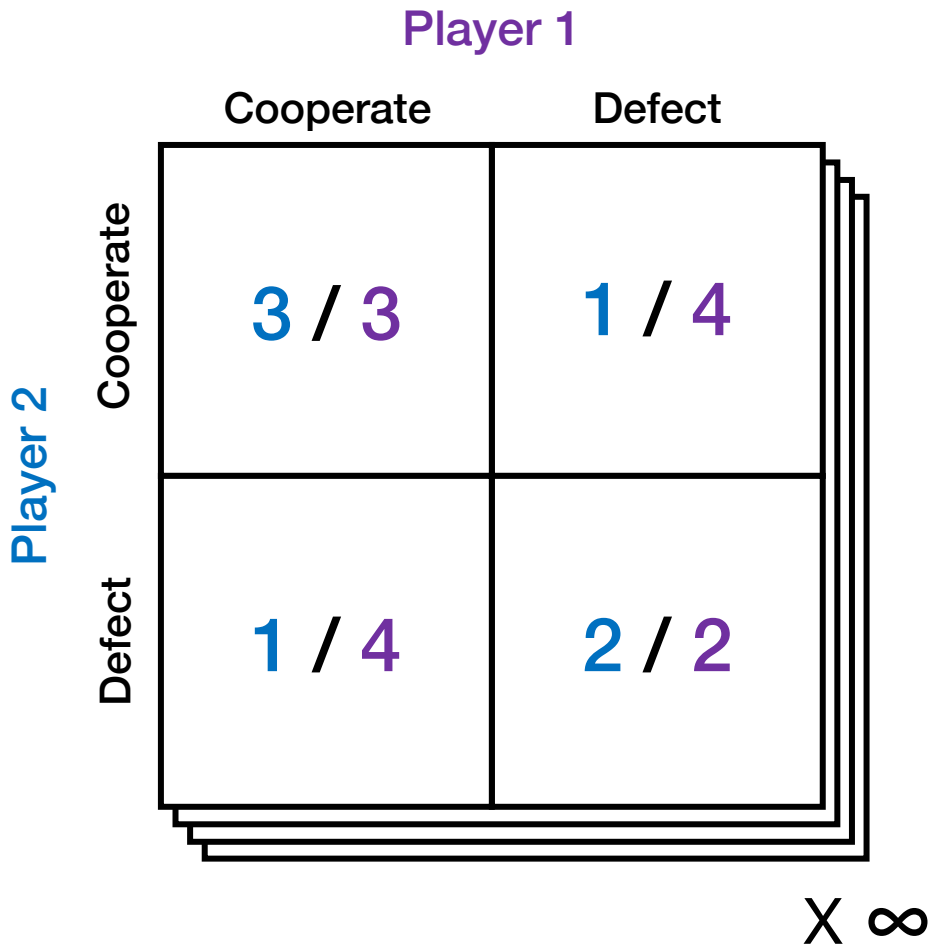


Solution: add a discount factor so the series converges, and to account for the possibility of game termination

$$u + \delta u + \delta^2 u + \delta^3 u \ldots = \frac{u}{1-\delta}$$

Infinite repeated prisoner's dilemma

Grim trigger: If anyone defects at any point, defect forever. Otherwise, cooperate.



Both players cooperate: $3 + \delta 3 + \delta^2 3 + \delta^3 3 \ldots$

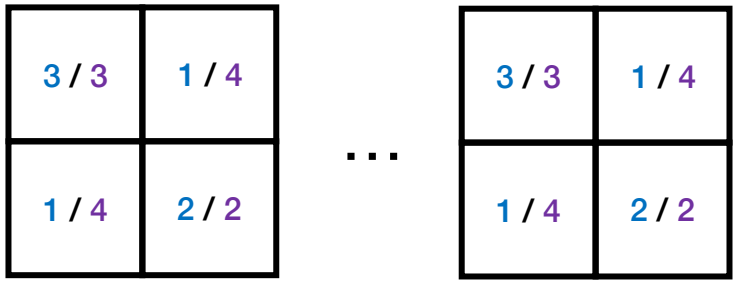
Cooperate

Player 1 tempted to defect first: $4 + \delta 2 + \delta^2 2 + \delta^3 2 \ldots$

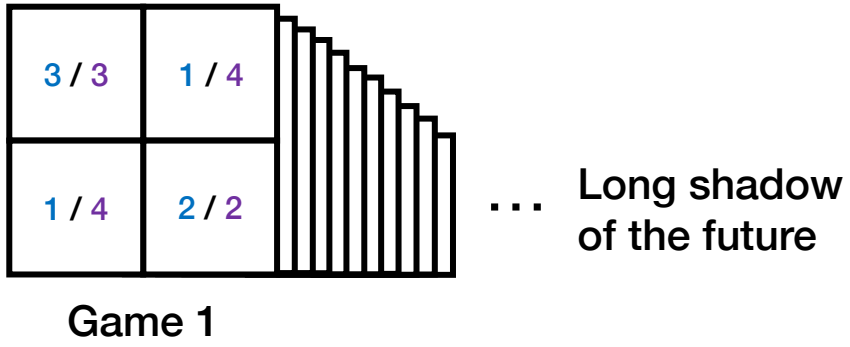
Defect Grim trigger

Cooperation if: $\frac{3}{1-\delta} \geq 4 + \frac{2\delta}{1-\delta}$ $\delta \geq 1/2$

Since this is an infinite game, after making the first move, the game looks the same at the second move.
Cooperation is always profitable if $\delta \geq 1/2$!



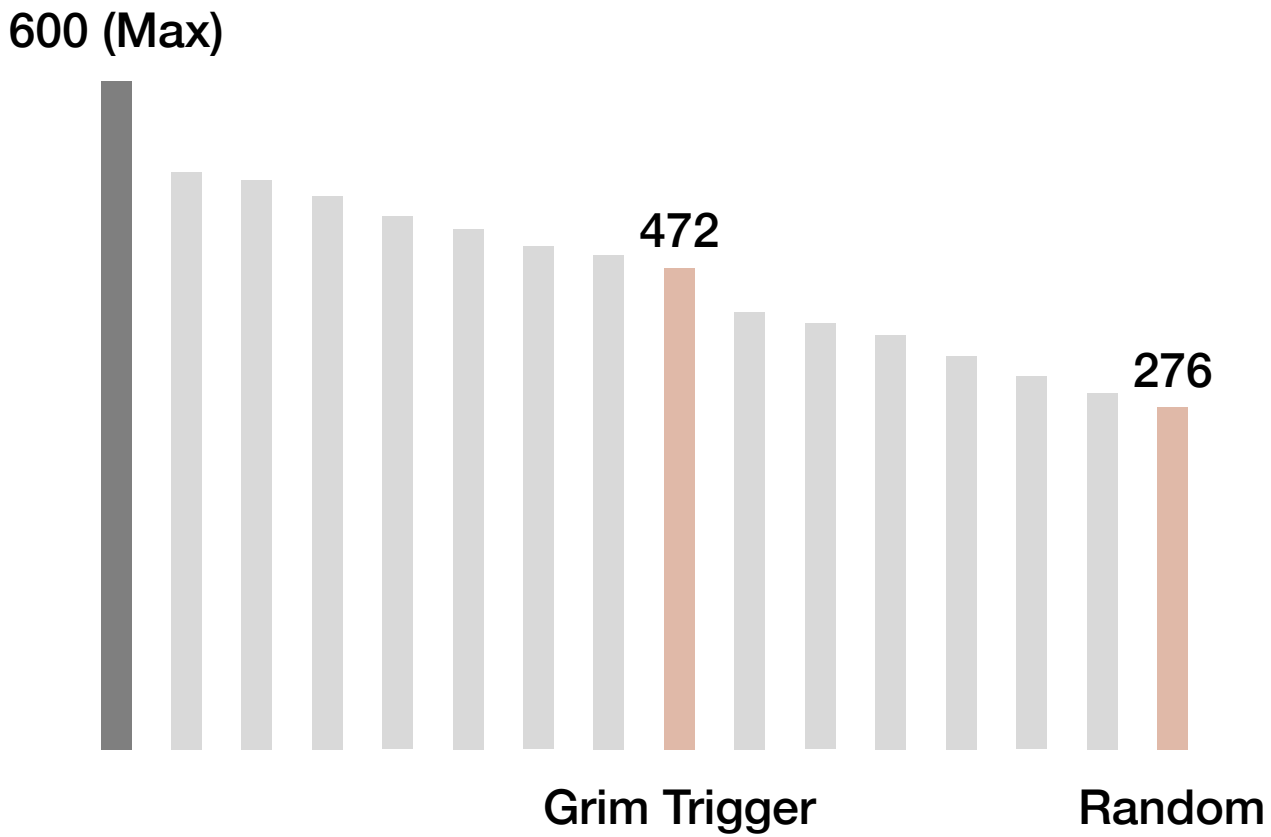
Finite prisoner's dilemma: no cooperation possible



Infinite prisoner's dilemma: cooperation!

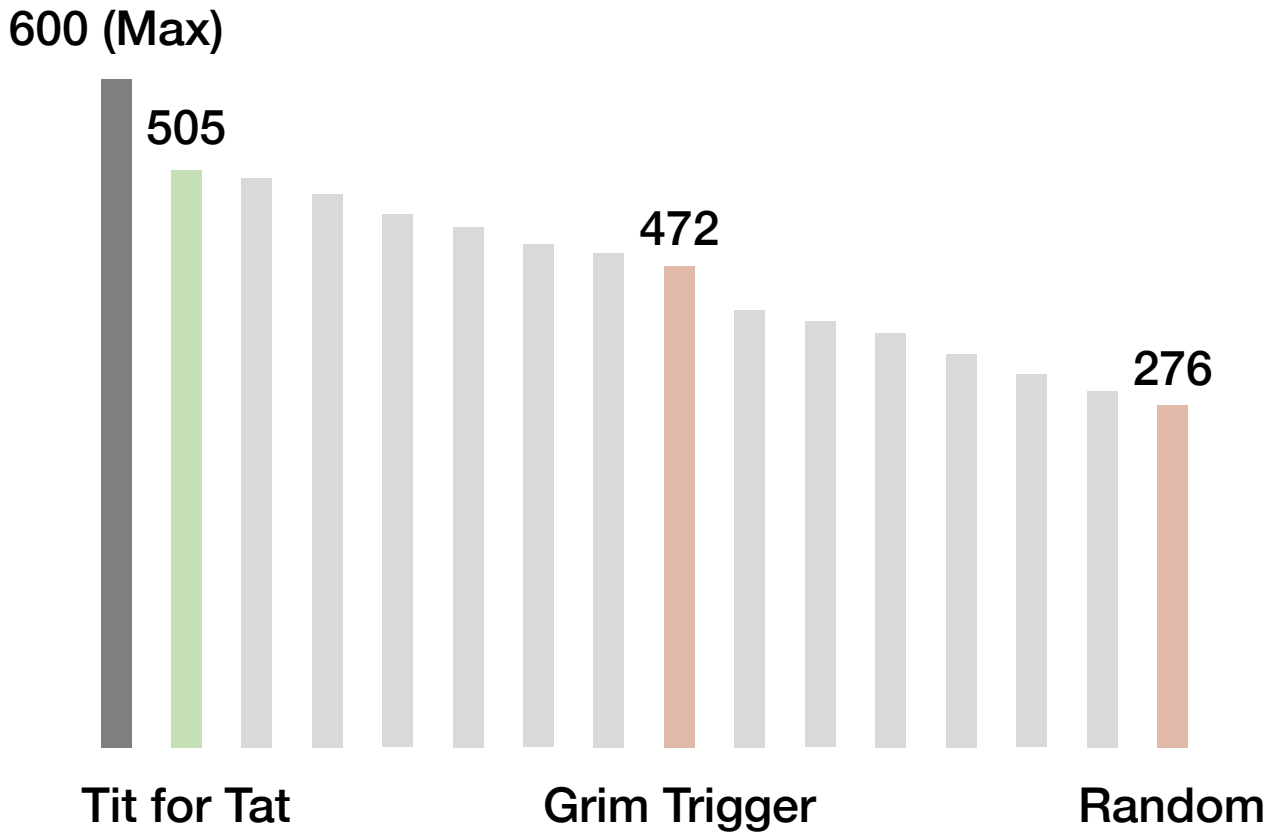
Axelrod's tournament

But… do these strategies work when you don't know what you're up against, and there is no way to guarantee to other players that you will play the Grim Trigger?



Axelrod's tournament

But… do these strategies work when you don't know what you're up against, and there is no way to guarantee to other players that you will play the Grim Trigger?



Grim Trigger (punish forever)

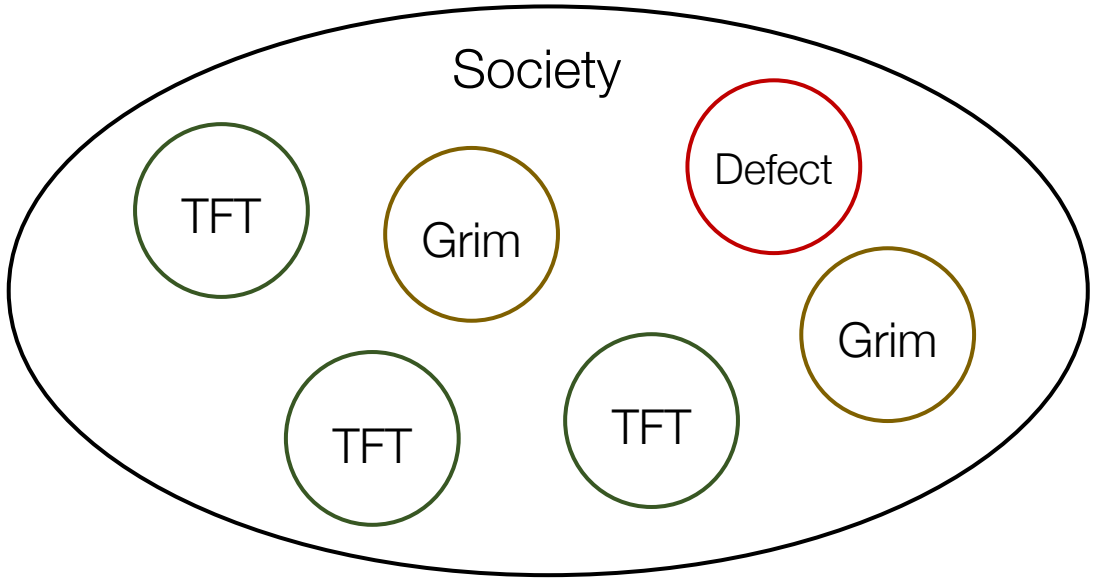
P1	3	3	1	2	4	2	2
P2	3	3	4	2	1	2	2

Tit for Tat (punish, but forgive)

P1	3	3	1	2	4	3	3
P2	3	3	4	2	1	3	3

The evolution of cooperation

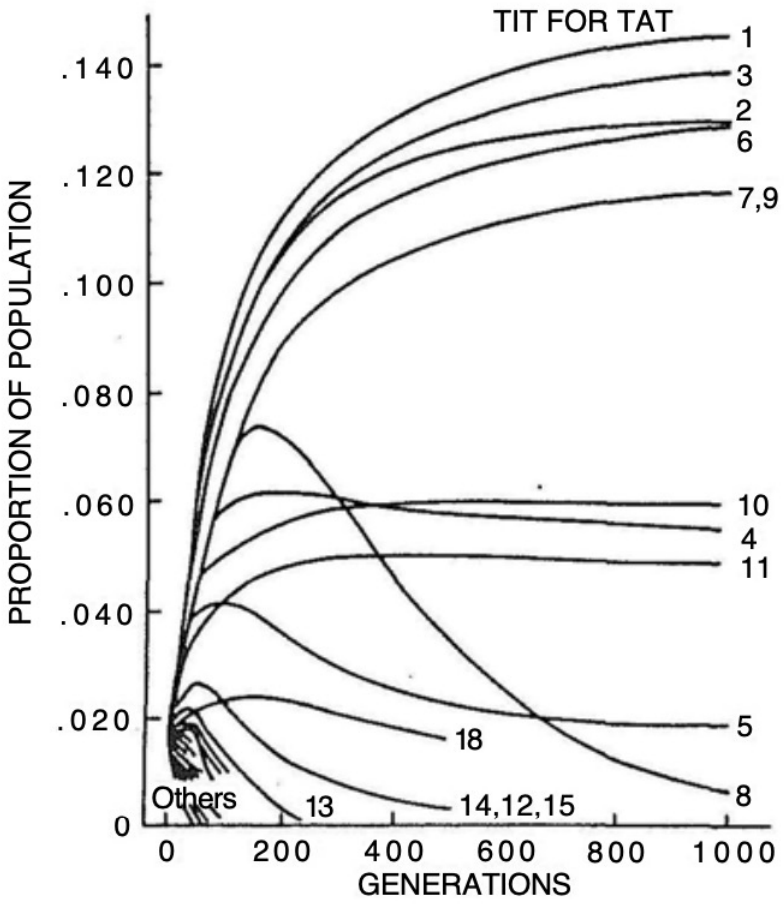
Let's make things more exciting: sum up the score at the end, and make the strategy "reproduce" according to how much they win.



Cooperation can begin with small clusters and thrive in neighborhoods that are "nice," protecting themselves from invasion. **But they can also go extinct with bad neighbors!**

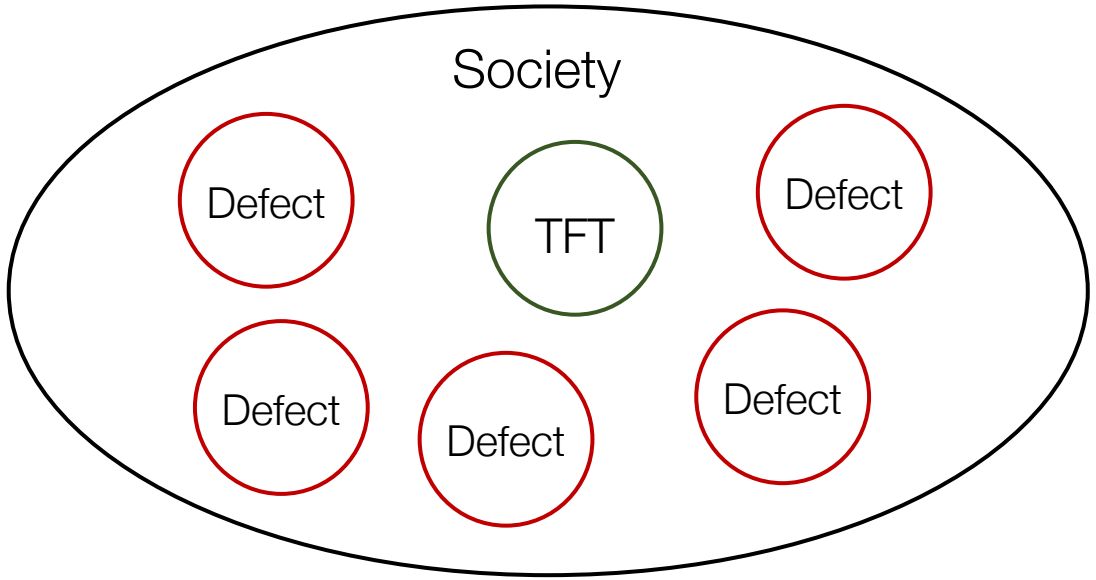
FIGURE 2

Simulated Ecological Success of the Decision Rules



The evolution of cooperation

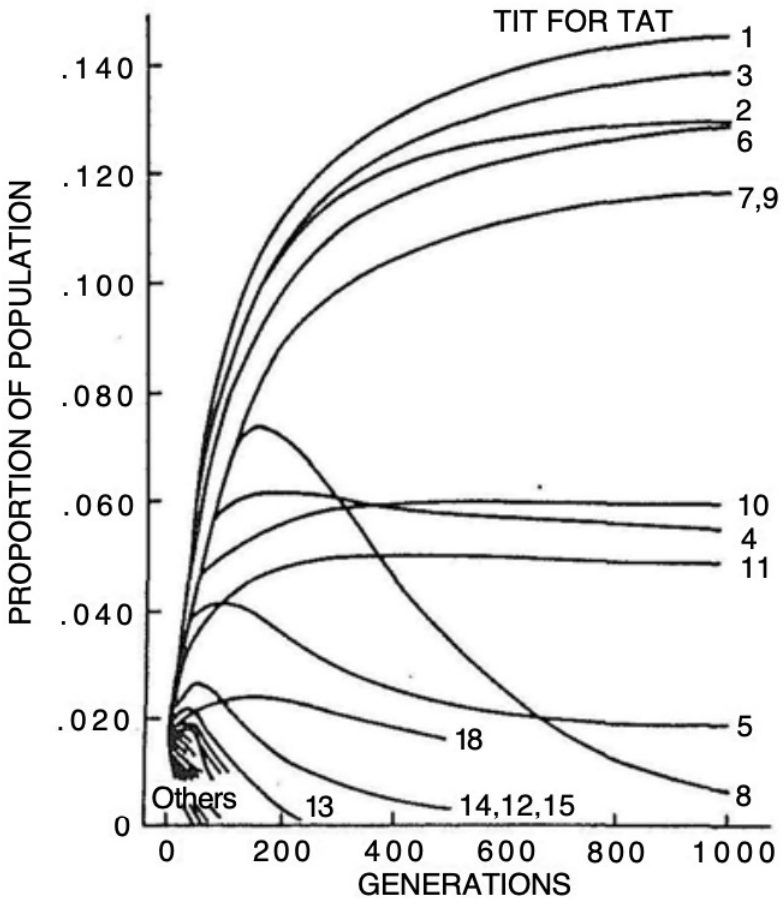
Let's make things more exciting: sum up the score at the end, and make the strategy "reproduce" according to how much they win.



Cooperation can begin with small clusters and thrive in neighborhoods that are "nice," protecting themselves from invasion. **But they can also go extinct with bad neighbors!**

FIGURE 2

Simulated Ecological Success of the Decision Rules



Size of society: 1 – 50,000

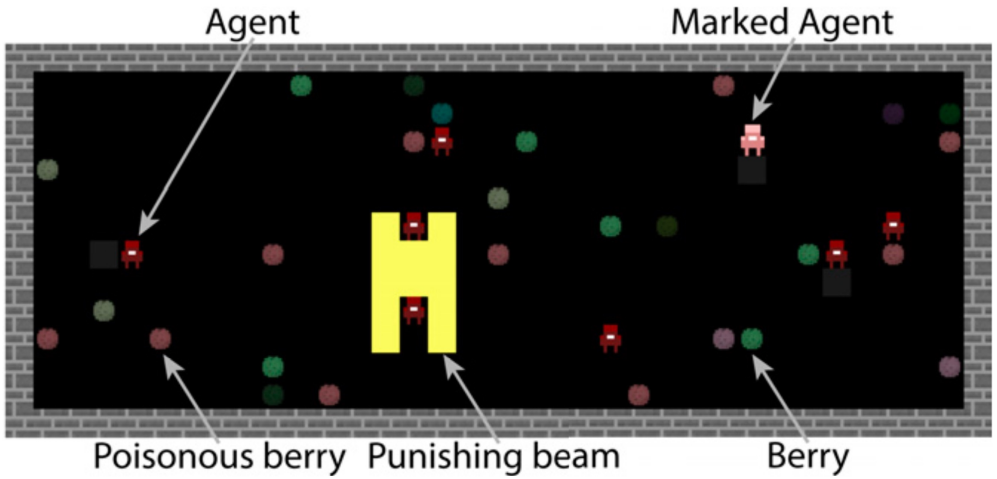
Size of society: 7,000,000

What differentiates us from other animal societies, and what accounts for the enormous gains of human ultrasociality?

Language, cultural values, economic systems, and third-party enforced norms.

Spurious normativity enhances learning of compliance and enforcement behavior in artificial agents

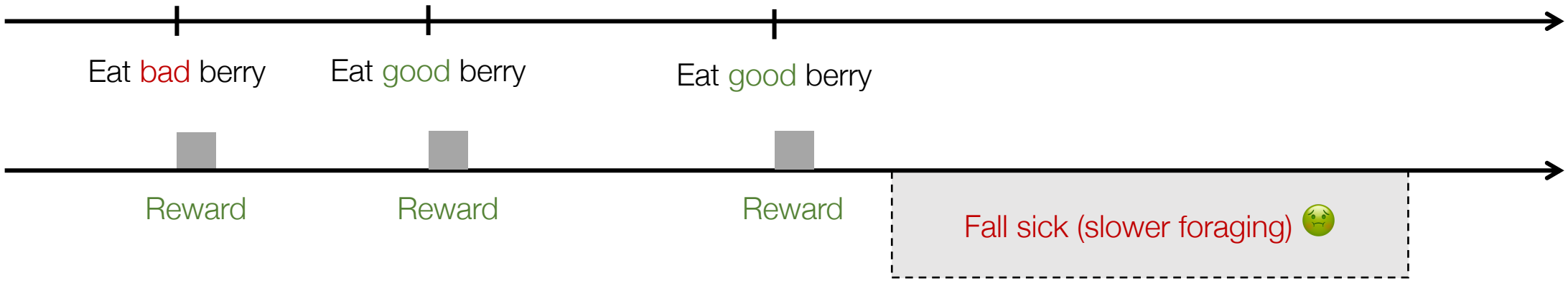
Raphael Köster[a,1], Dylan Hadfield-Menell[b,c], Richard Everett[a], Laura Weidinger[a], Gillian K. Hadfield[c,d,e,f,g,h], and Joel Z. Leibo[a,1]



Agents try to forage for food, but there are two berries:

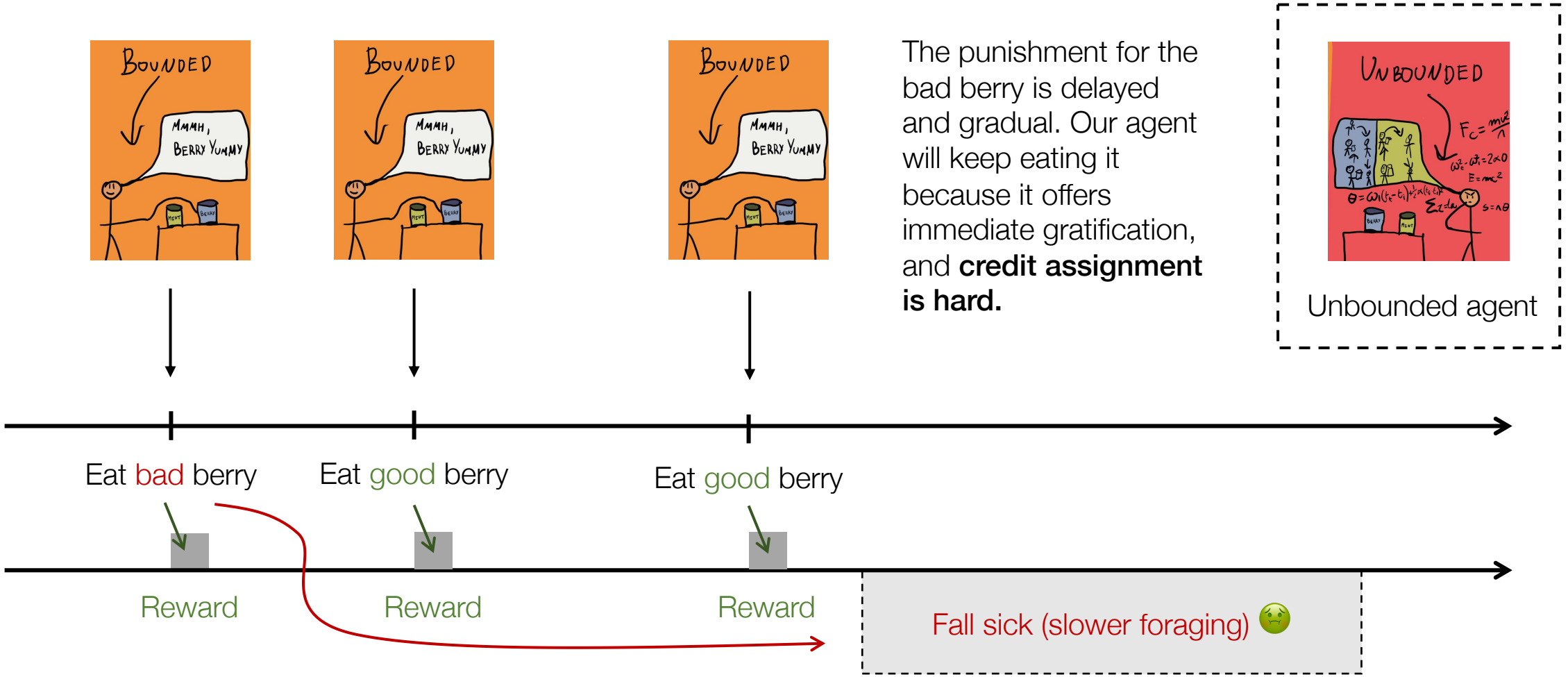
Good berry: $r = +1$

Poisonous berry: $r = +1$, but you get food poisoning 10 hours later.



Spurious normativity enhances learning of compliance and enforcement behavior in artificial agents

Raphael Köster[a,1], Dylan Hadfield-Menell[b,c], Richard Everett[a], Laura Weidinger[a], Gillian K. Hadfield[c,d,e,f,g,h], and Joel Z. Leibo[a,1]



The punishment for the bad berry is delayed and gradual. Our agent will keep eating it because it offers immediate gratification, and **credit assignment is hard.**

Unbounded agent

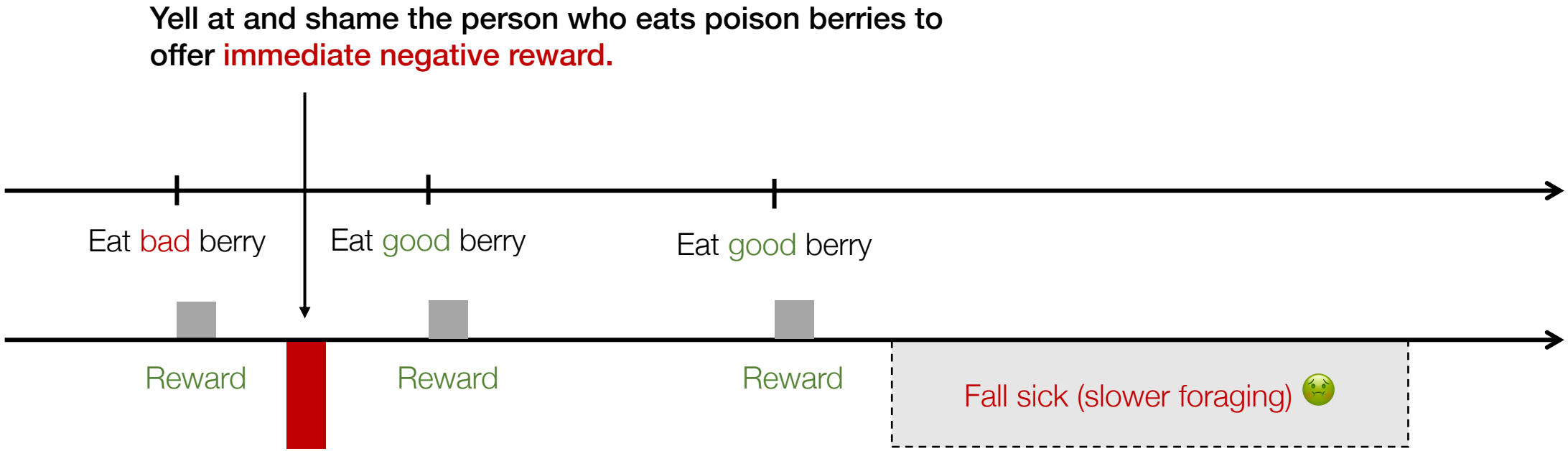
Picture credits: The Decision Lab

Spurious normativity enhances learning of compliance and enforcement behavior in artificial agents

Raphael Köster[a,1], Dylan Hadfield-Menell[b,c], Richard Everett[a], Laura Weidinger[a], Gillian K. Hadfield[c,d,e,f,g,h], and Joel Z. Leibo[a,1]

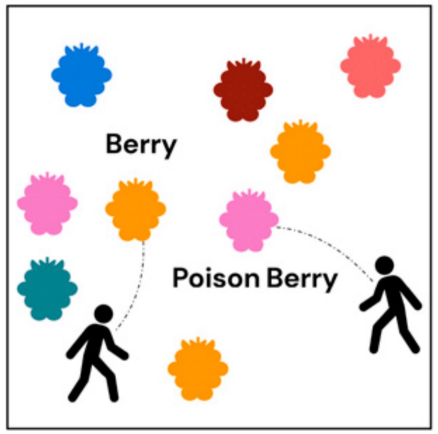
The punishment for the bad berry is delayed and gradual. Our agent will keep eating it because it offers immediate gratification, and **credit assignment is hard.**

Solution: social/cultural punishment.



Taboos and spurious normativity

No Rules

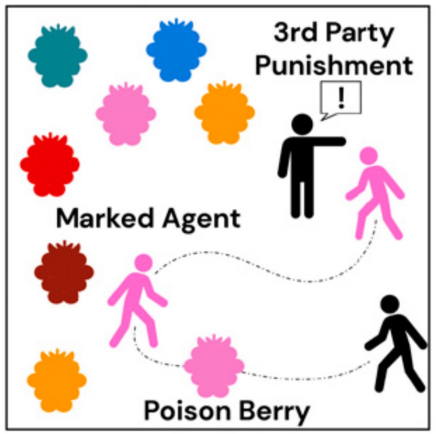


Good berry: $r = +1$

Poisonous berry: $r = +1$, food poisoning 10 hours later.

Delayed credit assignment

Important Rule



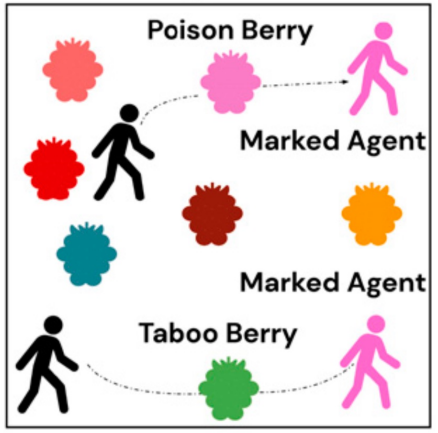
Good berry: $r = +1$

Poisonous berry: $r = +1$, labeled, food poisoning 10 hours later.

Punishment: $r = +0.1$ for punisher, $r = -1$ for violator

Hard to learn punishing behavior

Silly Rule



Good berry: $r = +1$

Poisonous berry: $r = +1$, labeled, food poisoning 10 hours later.

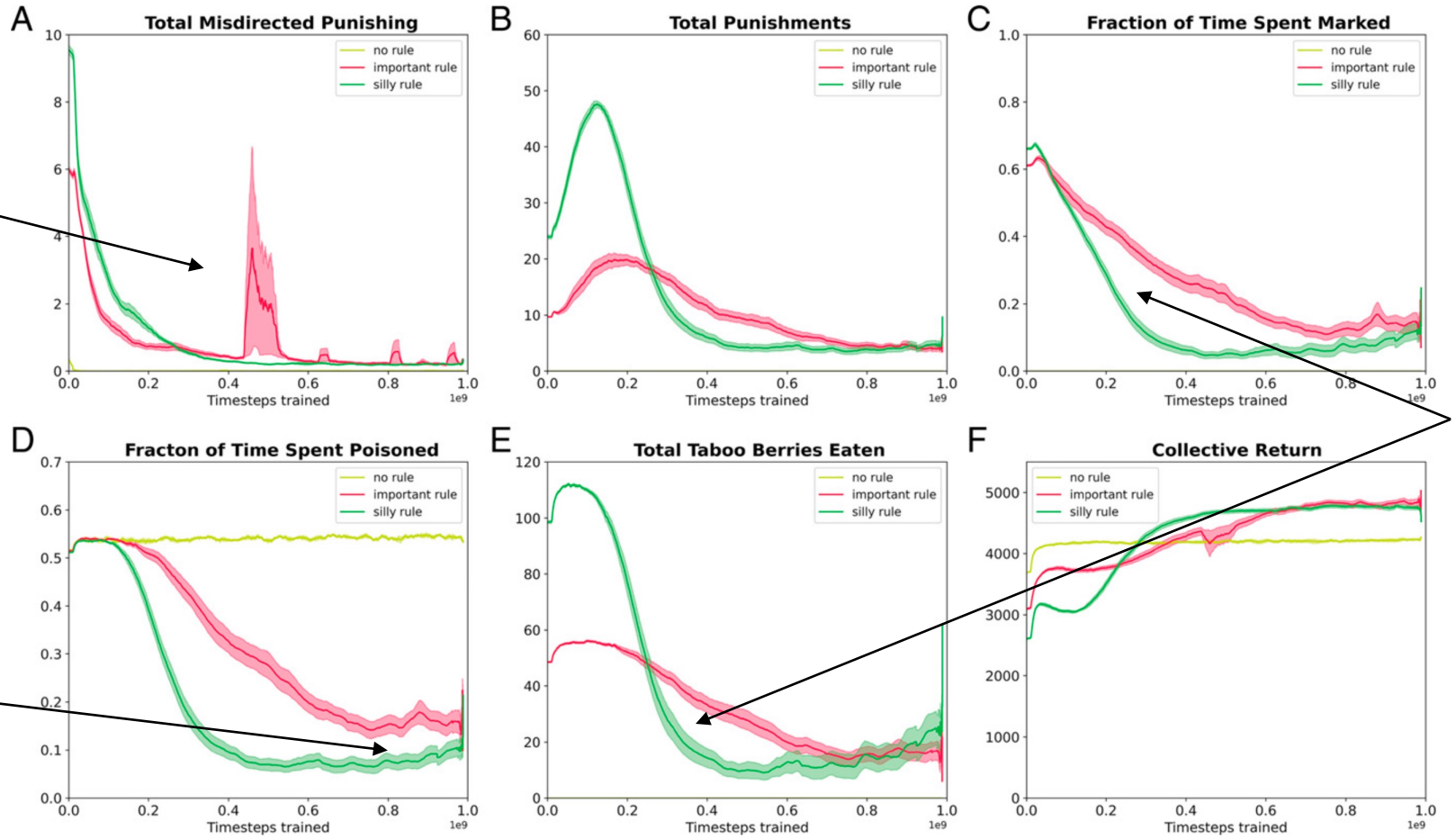
Punishment: $r = +0.1$ for punisher, $r = -1$ for violator

Taboo berry: $r = +1$, labeled for violation.

Experimenting on a multi-agent RL system

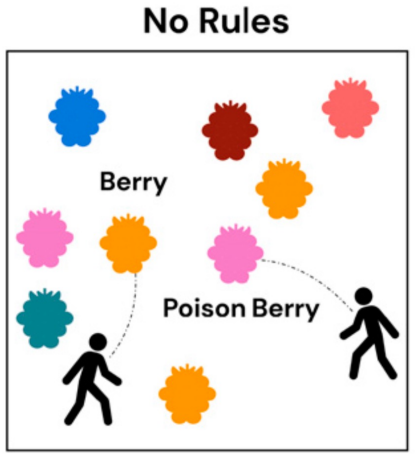
Without taboos, agents misdirect punishment very often

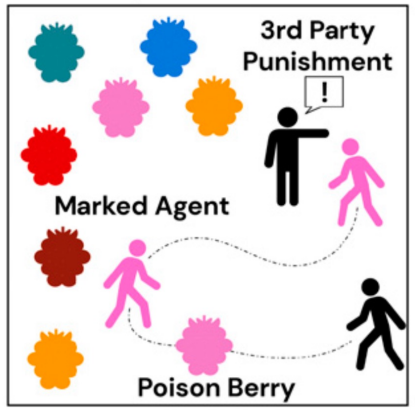
People in the taboo world are poisoned much less

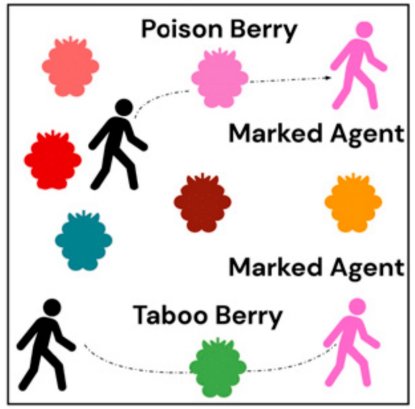


Learning occurs much faster with the taboo berry!

In summary







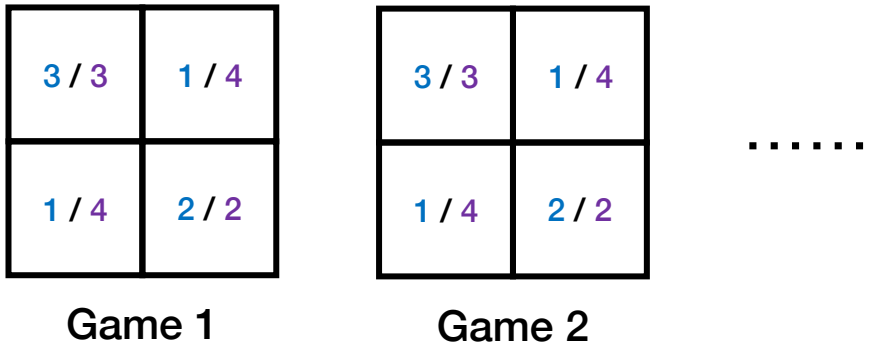
"The skills involved in third-party norm enforcement readily transfer from norm to norm, while the skills involved in compliance are norm specific.

Thus, adding a silly rule to a normative system that already contains some number of important rules can be beneficial because the silly rule may provide greater opportunity to practice third-party norm enforcement"

BUT: even though this is a cool demonstration, it doesn't show why humans have arbitrary social norms. Rather, it is an "existence proof" that arbitrary social norms can arise under simple rules.

Recap

- Welfare: multi-agent equivalent of utility functions
- Prisoner's dilemma, hide and seek, and Nash equilibria
- Problems with naïve rational choice theory: institutions, norms, and coupling between issues
- The evolution of cooperation and the enforcement of social values
- Individuals > Games > Multiple Games > Societies



Lecture 1	Lecture 2	Lecture 3	Lecture 4	**Lecture 5**	Lecture 6
Introduction and the RL problem	How computers learn	How people learn	Multi-agent systems	Interactions on graphs	Complex systems science

References and additional resources

- Governing the Commons, Elinor Ostrom (1990)
- A General Framework for Analyzing Sustainability of Socio-Ecological Systems, Elinor Ostrom (2009)
- A summary of The Evolution of Cooperation by Robert Axelrod
- The Iterated Prisoner's Dilemma and the Evolution of Cooperation: a great Youtube video
- Spurious Normativity Enhances Learning of Compliance and Enforcement Behavior in Artificial Agents