

CSC 163: Computing, Ethics, and Society

Chapter 7

Collective Action Problems Exercises Solution

Department of Computer Science,
Hofstra University

Exercise: Corporate AI race game

Two companies each choose D (delay release for safety testing) or R (release now). Payoffs (A, B): (D,D) = (5,5); (D,R) = (0,8); (R,D) = (8,0); (R,R) = (2,2).

(a) Does A have a dominant strategy?

(b) Fill in the table: is the outcome a Nash equilibrium? Pareto efficient?

(c) The game is repeated 5 rounds. A plays Tit-for-Tat (starts with D, then copies B's last move); B always plays R. Total payoff of each?

Outcome	Payoffs	Nash eq.?	Pareto eff.?
(D,D)	5, 5		
(D,R)	0, 8		
(R,D)	8, 0		
(R,R)	2, 2		

Exercise Solution: Corporate AI race game

- (a) Yes: **R**. If B plays D: $8 > 5$; if B plays R: $2 > 0$. Same for B.
- (b) The only Nash equilibrium is **(R,R) = (2,2)**, and it is the only outcome that is not Pareto efficient: (D,D) = (5,5) is a Pareto improvement.
- (c) Round 1: (D,R) $\rightarrow$ A 0, B 8. Rounds 2–5: (R,R) $\rightarrow$ 2 each. $A = 0 + 4 \times 2 = \mathbf{8}$; $B = 8 + 4 \times 2 = \mathbf{16}$.
- Racing dynamics: everyone prefers (D,D), yet defecting is individually rational.

Outcome	Payoffs	Nash eq.?	Pareto eff.?
(D,D)	5, 5	No	Yes
(D,R)	0, 8	No	Yes
(R,D)	8, 0	No	Yes
(R,R)	2, 2	Yes	No

Exercise: Many-player free riding

10 companies each choose to cooperate (keep safety standards) or defect (cut corners). A defecting company gains +4 for itself and imposes a cost of 1 on every company, including itself. A cooperating company gains and costs nothing.

- (a) Exactly one company defects. What is its payoff? What does each other company get?
- (b) All 10 defect. What is each company's payoff? Compare with all cooperating.
- (c) Is defection individually rational? Is it collectively harmful? Name the type of problem.
- (d) Which collective action problems in the deck have this structure? Give two.

Exercise Solution: Many-player free riding

- (a) Defector: $+4 - 1 = +3$. Each of the other 9: -1 .
- (b) Each defector: $4 - 10 \times 1 = -6$. If all cooperate: 0 each. Everyone is worse off.
- (c) Individually rational: private gain $4 >$ own share of the cost 1 , whatever the others do (dominant strategy). Collectively harmful: total cost 10 per defector $>$ gain 4 . A **collective action problem** (free riding, negative externalities).
- (d) Examples: military AI arms race, autonomous economy (corporate automation), public health measures during COVID-19, climate change.

Exercise: Cooperation mechanisms

- (a) Name the cooperation mechanism: direct reciprocity, indirect reciprocity, group selection, kin selection, institutions.
- (b) Which two of these does the deck argue would fail to make AIs cooperate with humans, and why?

Scenario	Mechanism
Two shopkeepers trade repeatedly, each rewarding the other's past cooperation	
A seller with a good reputation receives help from strangers	
A parent shares food with its offspring	
A team whose members help each other beats rival teams	
An international treaty with an inspection agency mediates cooperation	

Exercise Solution: Cooperation mechanisms

(b) **Kin selection** fails: AIs have very little information similarity to humans, so the cost outweighs the benefit. **Group selection** fails: AIs would have in-group bias towards other AIs, and against humans by exclusion.

Scenario	Mechanism
Two shopkeepers trade repeatedly, each rewarding the other's past cooperation	Direct reciprocity
A seller with a good reputation receives help from strangers	Indirect reciprocity
A parent shares food with its offspring	Kin selection
A team whose members help each other beats rival teams	Group selection
An international treaty with an inspection agency mediates cooperation	Institutions

Exercise: Conflict and evolution

Match each statement with a conflict factor (power shifts, first strike advantage, issue indivisibility, information problems, inequality) or a Lewontin condition (variation, retention, differential fitness).

Statement	Factor / condition
Rapid AI progress shifts the balance of military capability	
Whoever strikes the other's datacenters first reduces its retaliation	
Both sides want the same GPU server; it cannot be split	
Each side misjudges the other's capabilities and intentions	
Income and educational inequality predict crime	
AI models differ in size, data and architecture	
A new model is fine-tuned from an earlier model's architecture	
Users adopt some models much more than others	

Exercise Solution: Conflict and evolution

Lewontin's three conditions (variation, retention, differential fitness) are necessary and sufficient for evolution by natural selection: advanced AI satisfies them.

Statement	Factor / condition
Rapid AI progress shifts the balance of military capability	Power shifts
Whoever strikes the other's datacenters first reduces its retaliation	First strike advantage
Both sides want the same GPU server; it cannot be split	Issue indivisibility
Each side misjudges the other's capabilities and intentions	Information problems
Income and educational inequality predict crime	Inequality
AI models differ in size, data and architecture	Variation
A new model is fine-tuned from an earlier model's architecture	Retention
Users adopt some models much more than others	Differential fitness