

## THE PRISONERS' DILEMMA

### I. The Grading Game

#### A. What the game looks like:

		The Other Person	
		Write 'B'	Write 'A'
You	Write 'B'	B E	A E
	Write 'A'	E A	C C

#### B. The Rational Strategy: Dominance

1. What is a dominant act? A dominant act is one that you are better off doing whatever others do.

#### C. The Unfortunate Outcome:

1. The result of everyone acting rationally (writing an 'A') is that everyone gets a 'C'. This is not only unfortunate for you but, for everyone. Everyone prefers the outcome that results from everyone writing down a 'B' to the outcome that results from everyone writing down an 'A'.

#### D. An Analysis:

1. Everyone acting in a way that is individually rational results in an outcome that is collectively irrational. (We would all do better if we would all write 'B's.)

#### E. A Serious Application:

1. Nuclear Disarmament: If you replace 'Write a "B"' with 'Dispose of your nuclear arms' and 'Write an "A"' with 'Keep your nuclear arms,' then you begin to see why disarmament is so difficult.

#### F. A More Formal Account of the Prisoners' Dilemma:

1. A two-person, two-option prisoners' dilemma is a choice situation involving two agents, each with two alternatives, one of which is dominant, such that the outcome resulting from each agent selecting her dominant strategy is sub-optimal.

##### A. Dominance

- (1) *Strong Dominance*:  $a_1$  is strongly dominant with respect to  $a_2$  if, and only if for every possible state of the world,  $S_i$ , the agent prefers the outcome of  $a_1$  in  $S_i$  to  $a_2$  in  $S_i$ .

- (2) *Weak Dominance*:  $a_1$  is weakly dominant with respect to  $a_2$  if, and only if for some possible state of the world,  $S_n$ , the agent prefers the outcome of  $a_1$  in  $S_n$  to  $a_2$  in  $S_n$  and for no possible state of the world,  $S_i$ , does the agent prefer the outcome of  $a_2$  in  $S_i$  to  $a_1$  in  $S_i$ . (Weak dominance is sufficient for generating the prisoners' dilemma, but our examples will all involve strong dominance.)
- (3) *Sub-optimality*: An outcome is sub-optimal if, and only if, there is some other outcome that is preferred by all parties.

2. A Generalized Representation of the Prisoners' Dilemma:

		The Other Person	
		c	d
You	c	r	t
	d	s	p

Where:  $t > r > p > s$ .

**II. Some Possible Solutions to the Problem:**

- A. *Agreements and Promises*: We could all promise to write down a 'B'. Then we would each write down a 'B' and we would each get a 'B'.
  - 1. *Problem*: This will work only if each of us could count on everyone else to keep promises. But there are reasons for thinking that we could not always count on this.
    - a) Motivations for breaking your promise in Grading Game situations:
      - (1) *Greed*: Perhaps you are greedy and break your promise just to get an 'A' (even though you are confident that the other people won't).
      - (2) *Fear of Greed*: Perhaps you are willing to keep your promise if you are confident that others will, but you fear that others will be greedy and write 'A's'. So, in self-defense (to avoid getting an 'E') you write an 'A'.
      - (3) *Fear of Fear of Greed*: Perhaps you are confident that others will not be greedy but you are afraid that they might think that you (or someone else) will be greedy. If so, you reason, they will write 'A's to protect themselves. But, of course, you then need to write an 'A' to protect yourself from their (misguided) protection of themselves
      - (4) *And So On, And So On, And So On . . .*

- B. *Force and Fear*: A better solution is to force each individual to act contrary to her own or his own interest in order to produce the best outcome for all.

1. The Grading Game Revised:

		The Other Person	
		Write 'B'	Write 'A'
You	Write 'B'	<div style="display: flex; justify-content: space-between; align-items: center;"> <span style="font-size: 0.8em;">B</span> <span style="font-size: 0.8em;">B</span> </div>	<div style="display: flex; justify-content: space-between; align-items: center;"> <span style="font-size: 0.8em;">A</span> <span style="font-size: 0.8em;">+2BK</span> </div>
	Write 'A'	<div style="display: flex; justify-content: space-between; align-items: center;"> <span style="font-size: 0.8em;">E</span> <span style="font-size: 0.8em;">A +2BK</span> </div>	<div style="display: flex; justify-content: space-between; align-items: center;"> <span style="font-size: 0.8em;">C</span> <span style="font-size: 0.8em;">+2BK</span> </div>

(Where '2BK' means 'Two Broken Kneecaps'.)

2. How does this solve the problem? The problem of the Grading Game (like the serious real life cases of prisoners' dilemmas) is caused by too much liberty. Liberty, in these cases, undermines security. Since everyone is free to write an 'A' without any threat of penalty, we cannot always trust others not to do so. Nor can they always trust us not to do so. And so, we all get 'C's. But if we can attach a penalty to writing an 'A' that is sufficient to assure us that others will write a 'B' and to assure them that we will do so, then we can bring it about that we all get 'B's. We solve the problem by being willing to restrict our liberty provided others do so as well.

**III. The Moral of the Problem**

- A. Rational pursuit of one's goals sometimes requires the willingness to accept restrictions on one's own freedom to pursue those goals *provided others do so as well*.