

Inception: Efficiently Computable Misinformation Attacks on Markov Games

Jeremy McMahan, Young Wu, Yudong Chen, Xiaojin Zhu, and Qiaomin Xie

Motivation



Motivation

- Lying about rewards can improve outcomes.



Motivation

- Lying about rewards can improve outcomes.
- Lies can come from **misinformation** online.



What happens when
assumptions are violated?

Example

True Game

Example

True Game

	L	R
U	0, 5	1, 0
D	1, ϵ	0, 0
S	1, 0	0, ϵ

Example

True Game

Unique NE

	L	R
U	0, 5	1, 0
D	1, ϵ	0, 0
S	1, 0	0, ϵ

Example

True Game

Unique NE



	L	R
U	0, 5	1, 0
D	1, ϵ	0, 0
S	1, 0	0, ϵ

If P1 is rational, it must play D, so P2 gets 0!

Example

True Game

	L	R
U	0, 5	1, 0
D	1, ϵ	0, 0
S	1, 0	0, ϵ

P2 wants
Unique NE

0, 5
1, 0
1, ϵ
0, 0
1, 0
0, ϵ

If P1 is rational, it must play D, so P2 gets 0!

Example

Faked Game

Example

Faked Game

	L	R
U	0, 5	1, $5+\epsilon$
D	1, ϵ	0, 2ϵ
S	1, 0	0, ϵ

Example

Faked Game

	L	R
U	0, 5	1, $5+\epsilon$
D	1, ϵ	0, 2ϵ
S	1, 0	0, ϵ

Increased

Example

Faked Game

	L	R
U	0, 5	1, $5+\epsilon$
D	1, ϵ	0, 2ϵ
S	1, 0	0, ϵ

Unique NE

Example

Faked Game

	L	R
U	0, 5	1, $5+\epsilon$
D	1, ϵ	0, 2ϵ
S	1, 0	0, ϵ

Unique NE

P1 must play U, so P2 can get 5 in true game!

Example

Faked Game

The diagram shows an extensive form game tree. Player 1 (P1) moves first, choosing between U, D, and S. If P1 chooses U, Player 2 (P2) moves second, choosing between L and R. The payoffs are listed as (P1 payoff, P2 payoff). The payoffs for the (U, L) node are 0, 5. The payoffs for the (U, R) node are 1, $5 + \epsilon$. The payoffs for the (D, L) node are 1, ϵ . The payoffs for the (D, R) node are 0, 2ϵ . The payoffs for the (S, L) node are 1, 0. The payoffs for the (S, R) node are 0, ϵ . A purple box labeled "P2 wins!" with an arrow points to the (U, L) node. An orange box labeled "Unique NE" with an arrow points to the (U, R) node.

		L	R
U		0, 5	1, $5 + \epsilon$
D		1, ϵ	0, 2ϵ
S		1, 0	0, ϵ

P1 must play U, so P2 can get 5 in true game!

Example

Faked Game

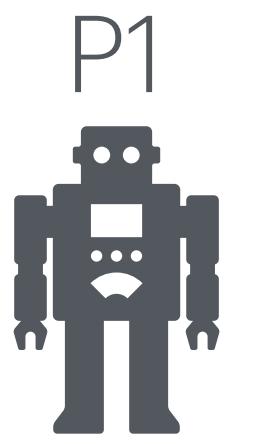
“Inception Attack”



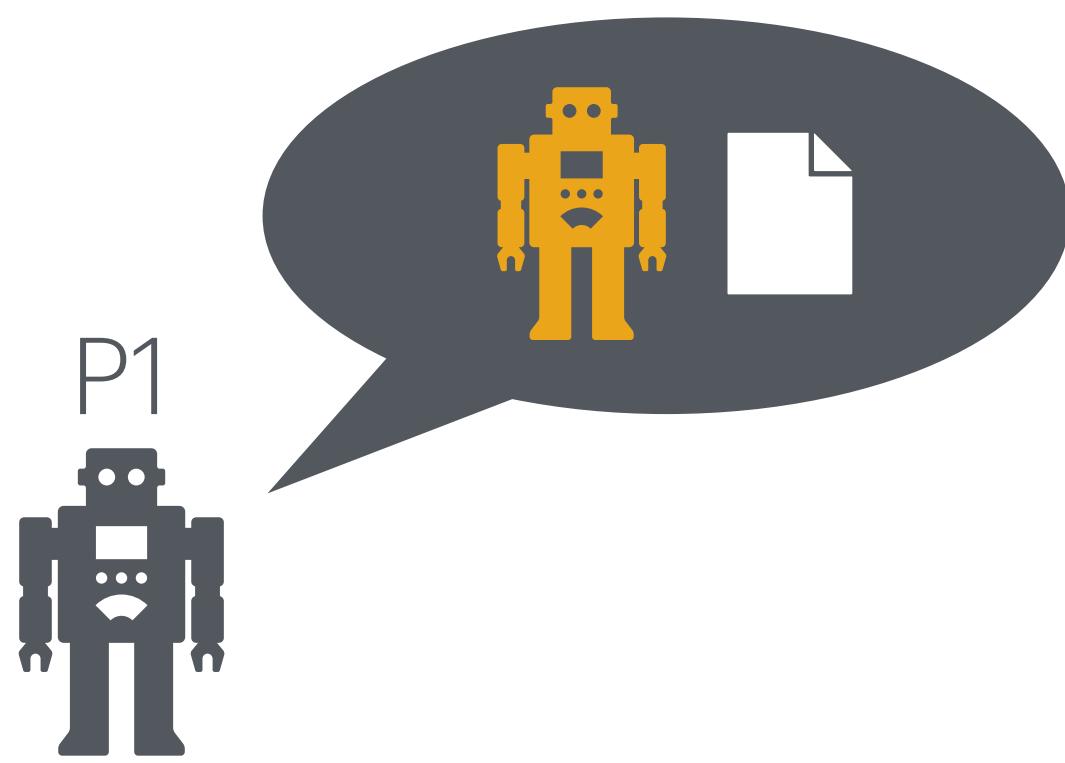
Inception attacks are powerful,
but can they be **computed**?

Inception Approach

Inception Approach



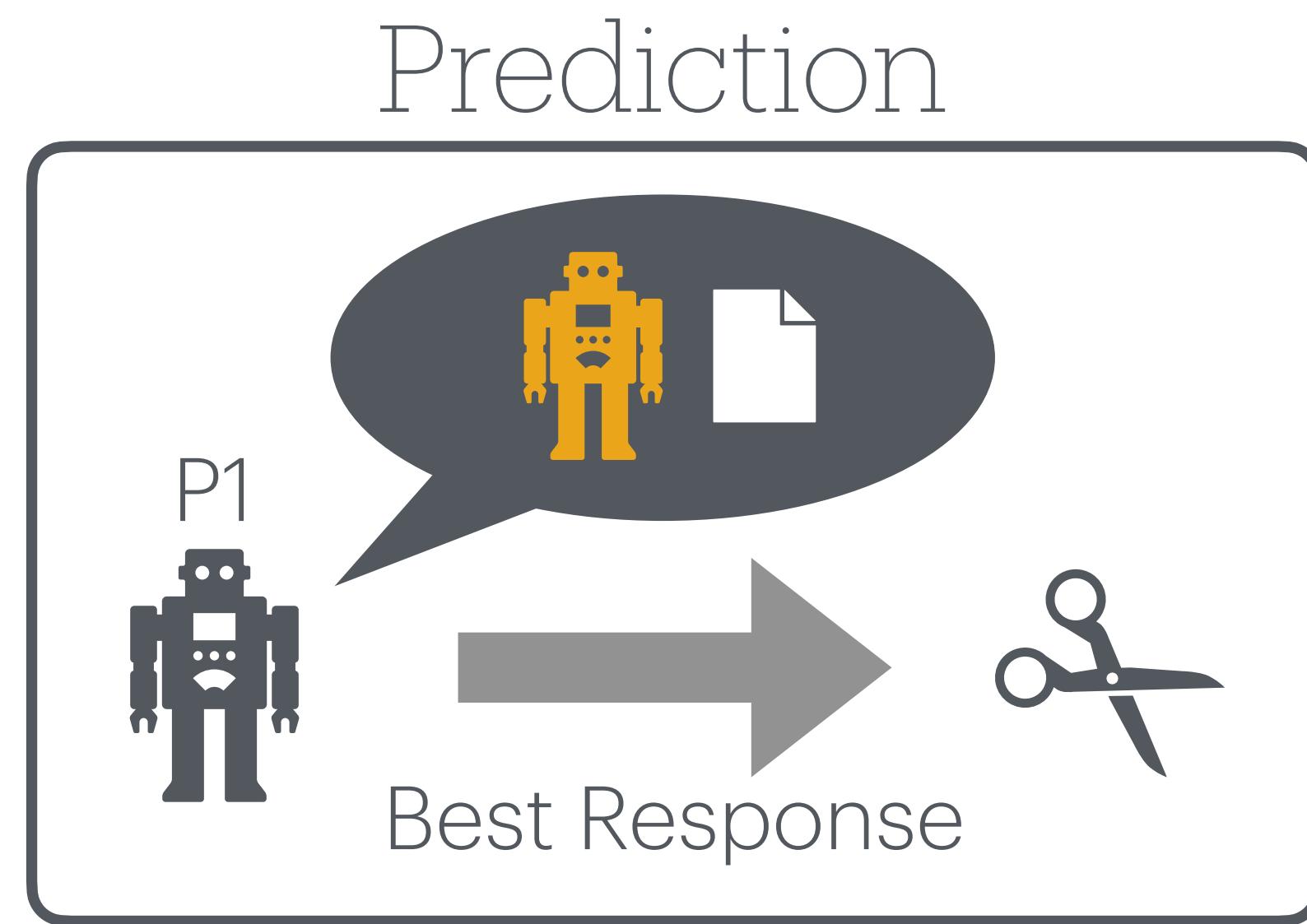
Inception Approach



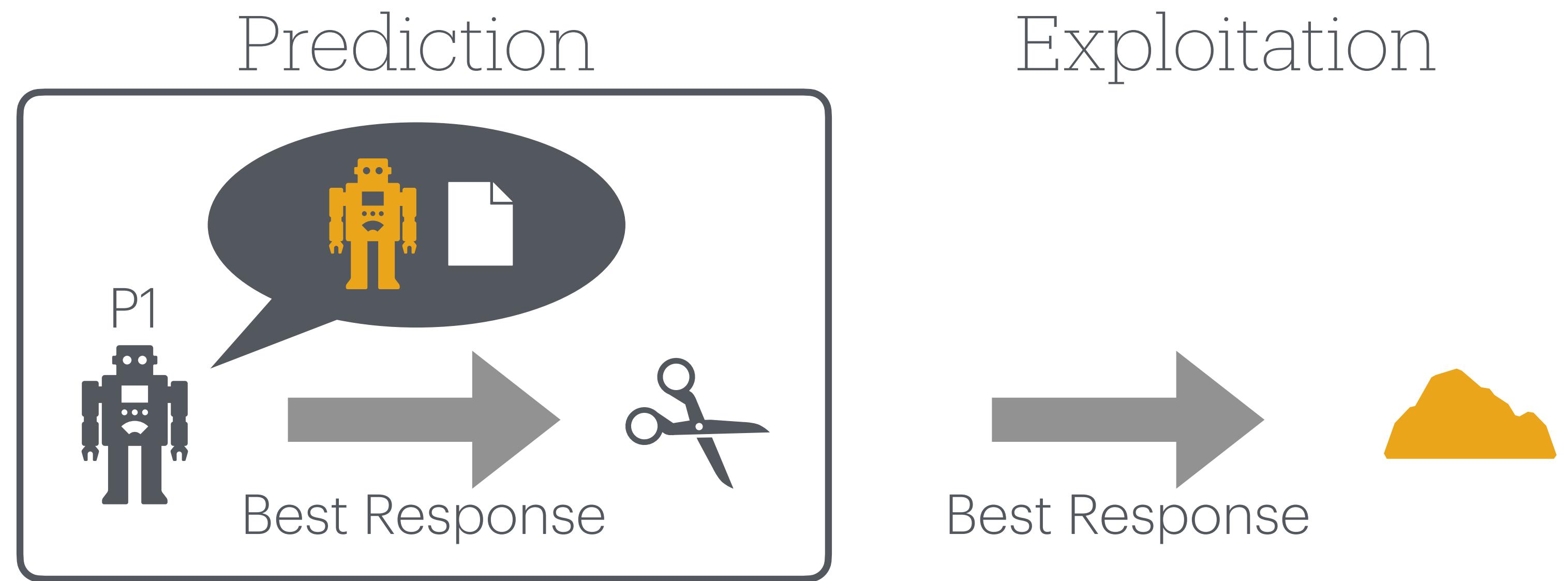
Inception Approach



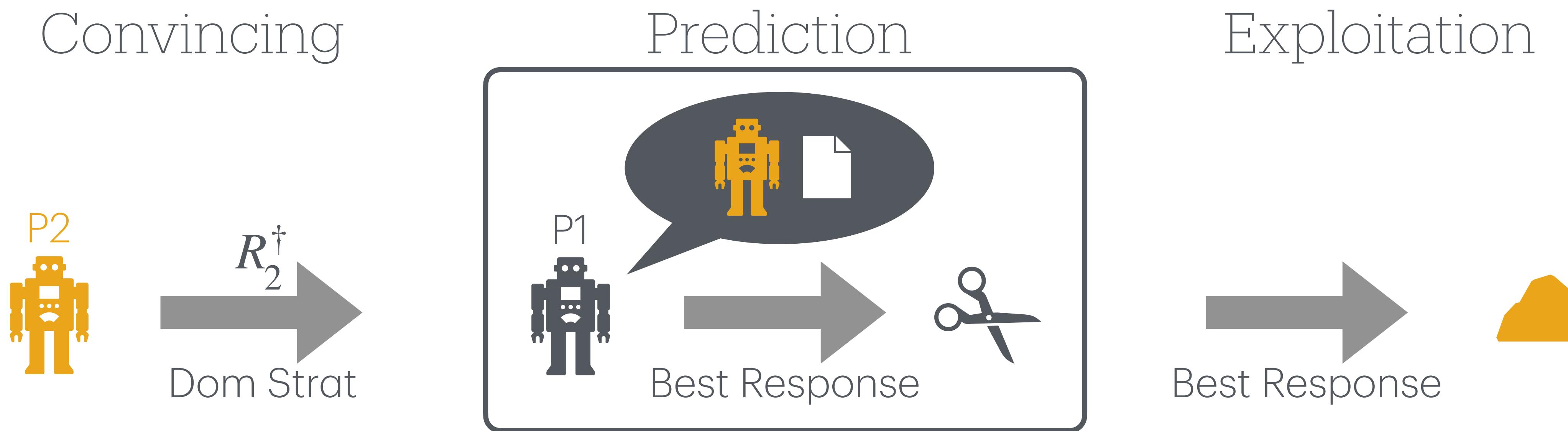
Inception Approach



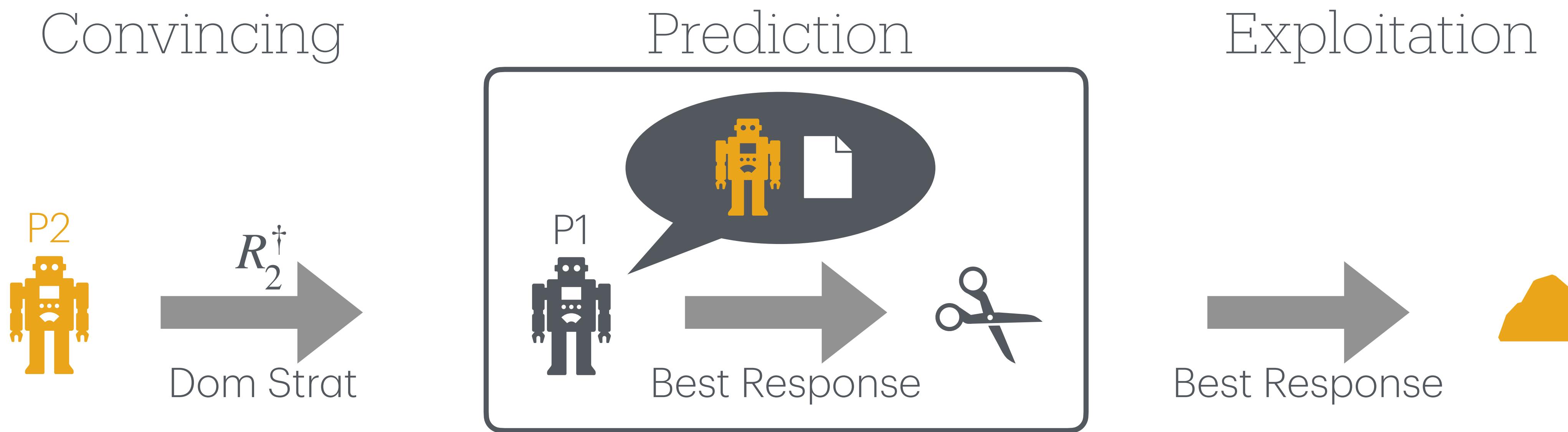
Inception Approach



Inception Approach



Inception Approach



Repeat to find the best pure strategy inception!

Algorithm

P2 fakes R

Algorithm

P2 fakes R

	L	R
U	0, 5	1, 5+ ϵ
D	1, ϵ	0, 2 ϵ
S	1, 0	0, ϵ

Algorithm

P2 fakes R

The diagram shows an extensive form game tree. Player 1 (P1) moves first, choosing between U, D, and S. If P1 chooses U, Player 2 (P2) moves, choosing between L and R. The payoffs are (Player 1, Player 2): (U, L) = (0, 5), (U, R) = (1, 5+ ϵ), (D, L) = (1, ϵ), (D, R) = (0, 2 ϵ), and (S, L) = (1, 0), (S, R) = (0, ϵ). The cell (U, R) is highlighted with a thick purple border.

		L	R
		0, 5	1, 5+ ϵ
		1, ϵ	0, 2 ϵ
S	U	1, 0	0, ϵ

Algorithm

P2 fakes R

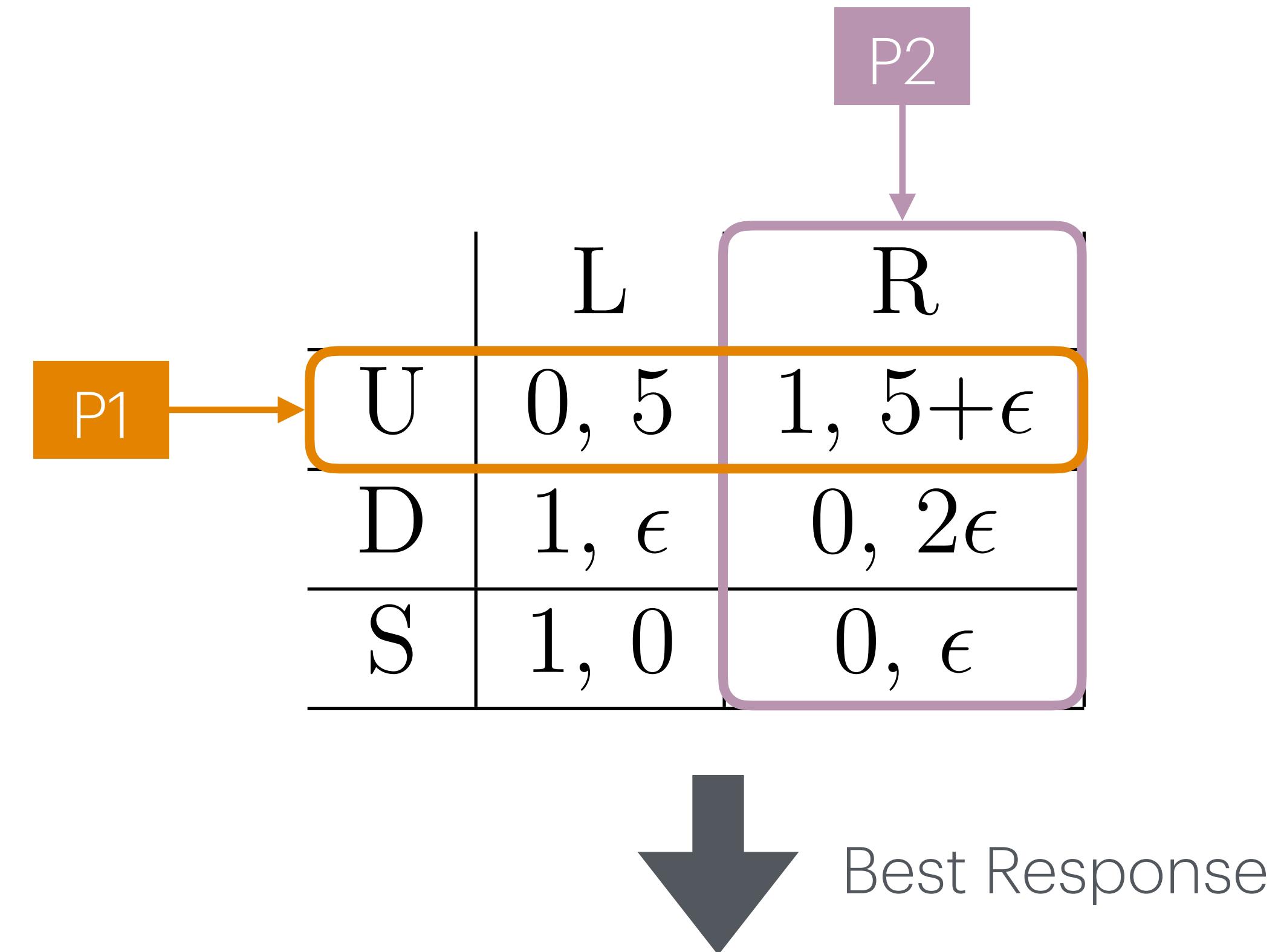
The diagram illustrates an extensive form game tree between two players, P1 and P2. The game starts with P1 choosing between U, D, and S. If P1 chooses U, P2 chooses between L and R. The payoffs are as follows:

P1\ P2	L	R
U	0, 5	1, 5+ ϵ
D	1, ϵ	0, 2 ϵ
S	1, 0	0, ϵ

P1's strategy U is highlighted with an orange border, and P2's strategy R is highlighted with a purple border. Arrows indicate the flow of the game: an orange arrow from P1 to the U strategy, and a purple arrow from P2 to the R strategy.

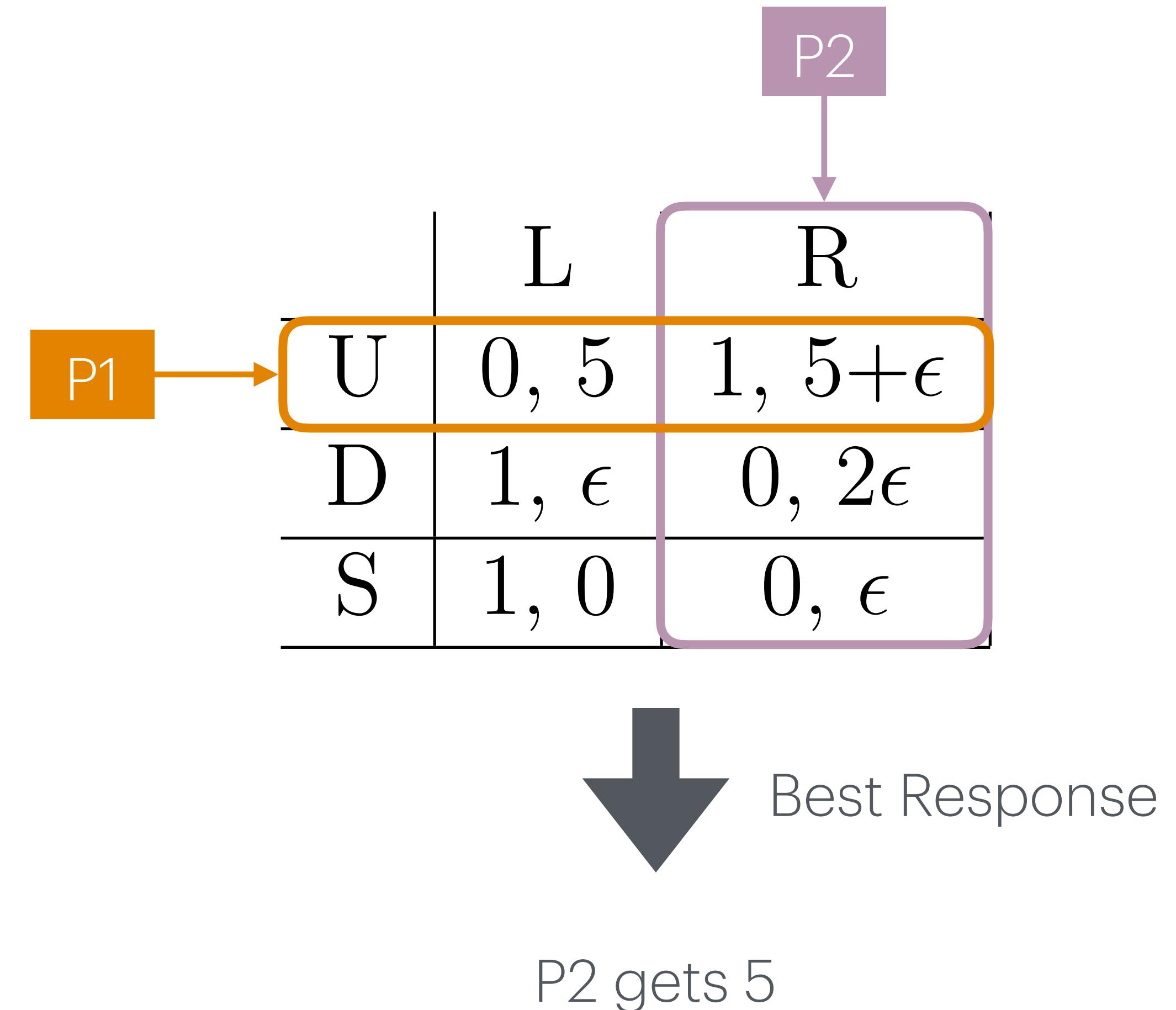
Algorithm

P2 fakes R



Algorithm

P2 fakes R



Algorithm

P2 fakes L

Algorithm

P2 fakes L

	L	R
U	0, 5	1, 0
D	1, ϵ	0, 0
S	1, 2ϵ	0, ϵ

Algorithm

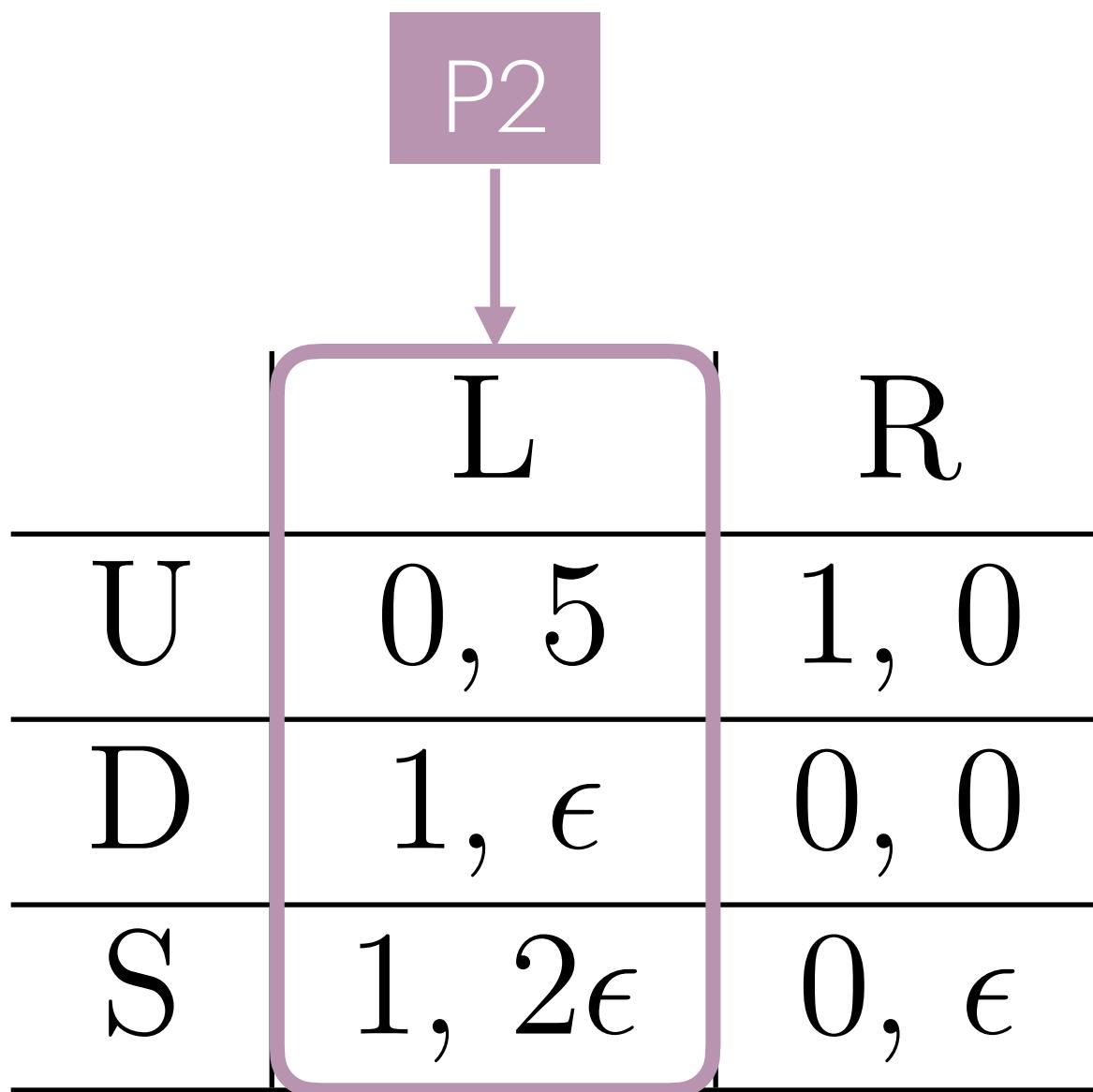
P2 fakes L

	L	R
U	0, 5	1, 0
D	1, ϵ	0, 0
S	1, 2ϵ	0, ϵ

Increased

Algorithm

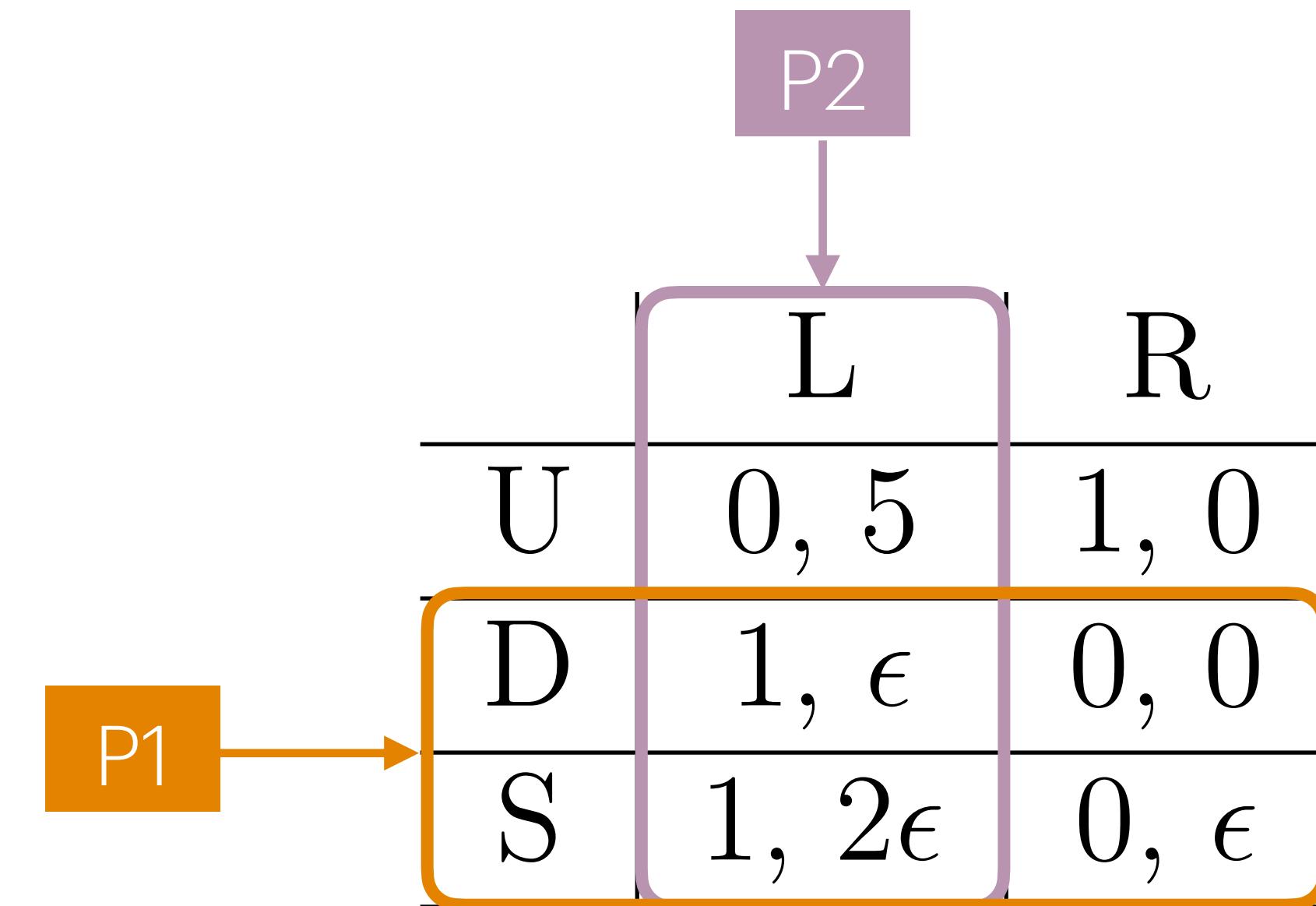
P2 fakes L



		L	R
		0, 5	1, 0
		1, ϵ	0, 0
U	0, 5	1, 0	
D	1, ϵ	0, 0	
S	1, 2ϵ	0, ϵ	

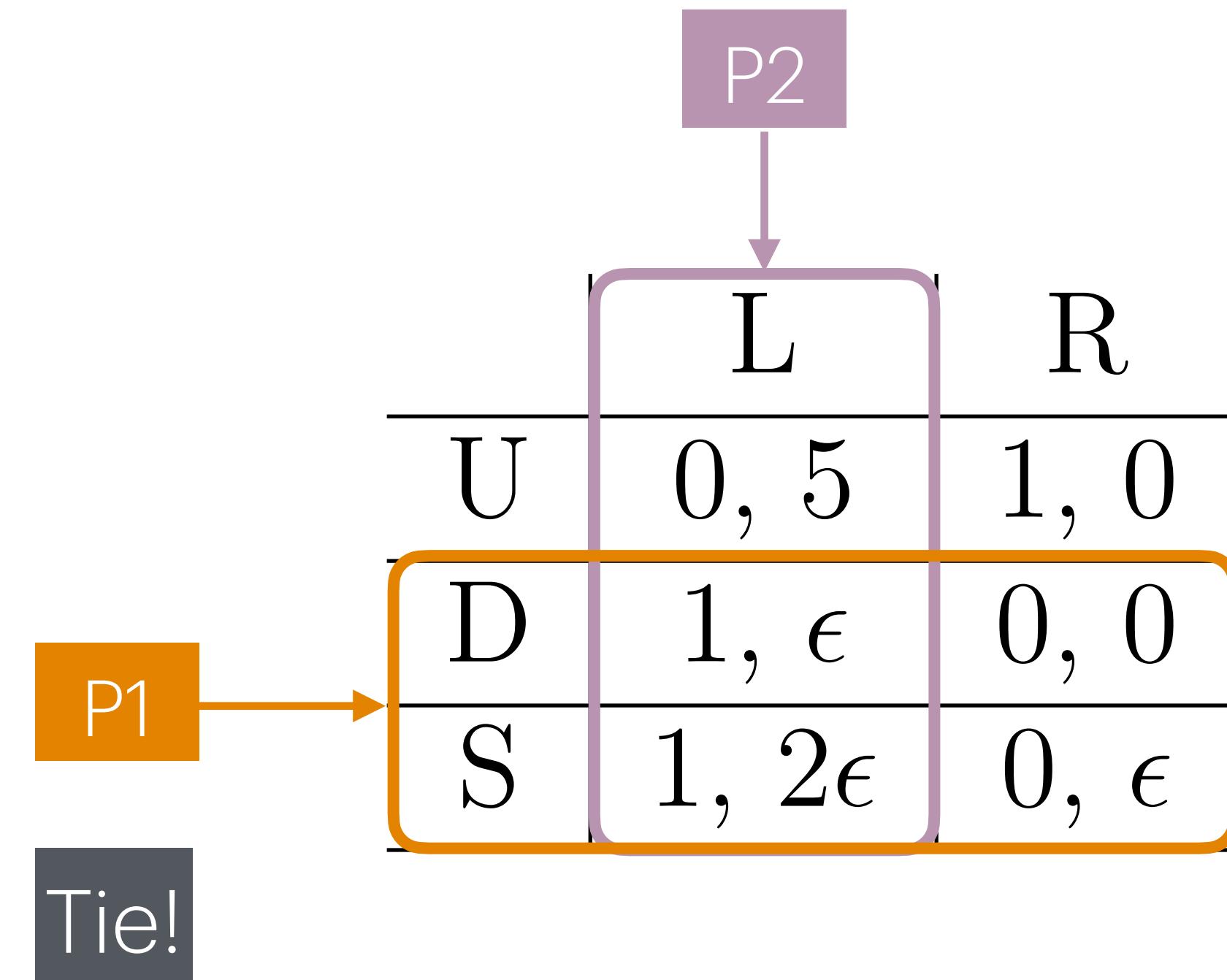
Algorithm

P2 fakes L



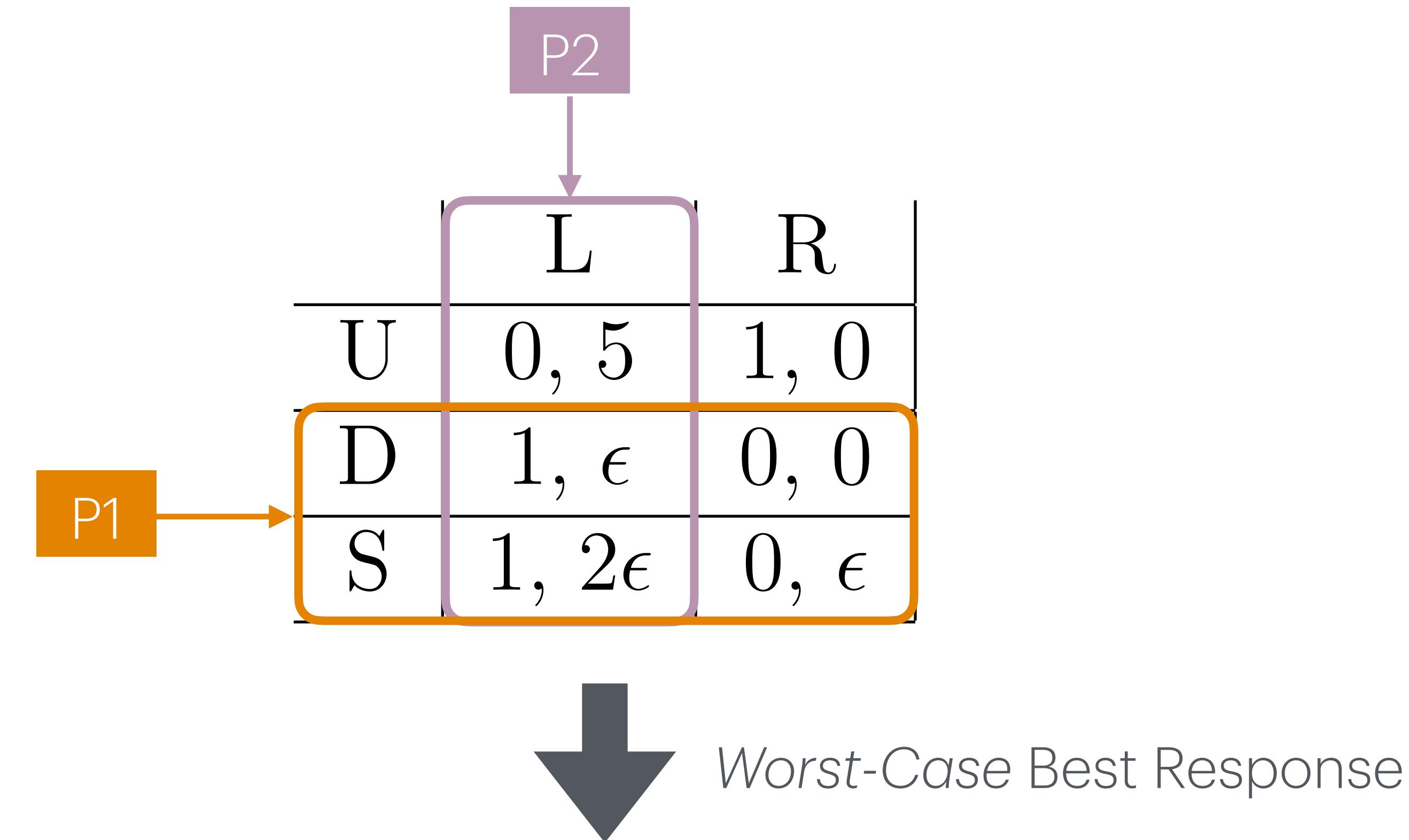
Algorithm

P2 fakes L



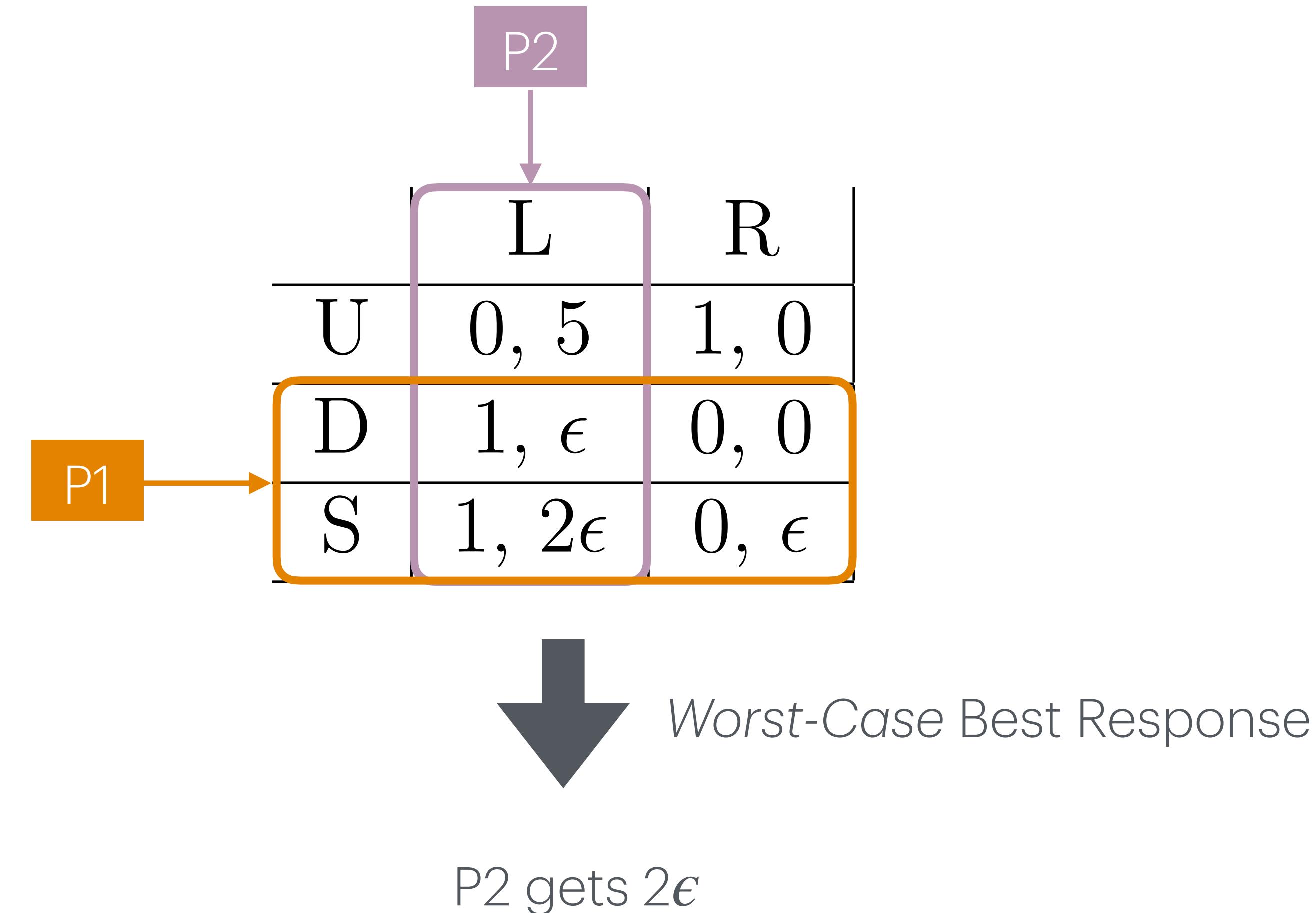
Algorithm

P2 fakes L



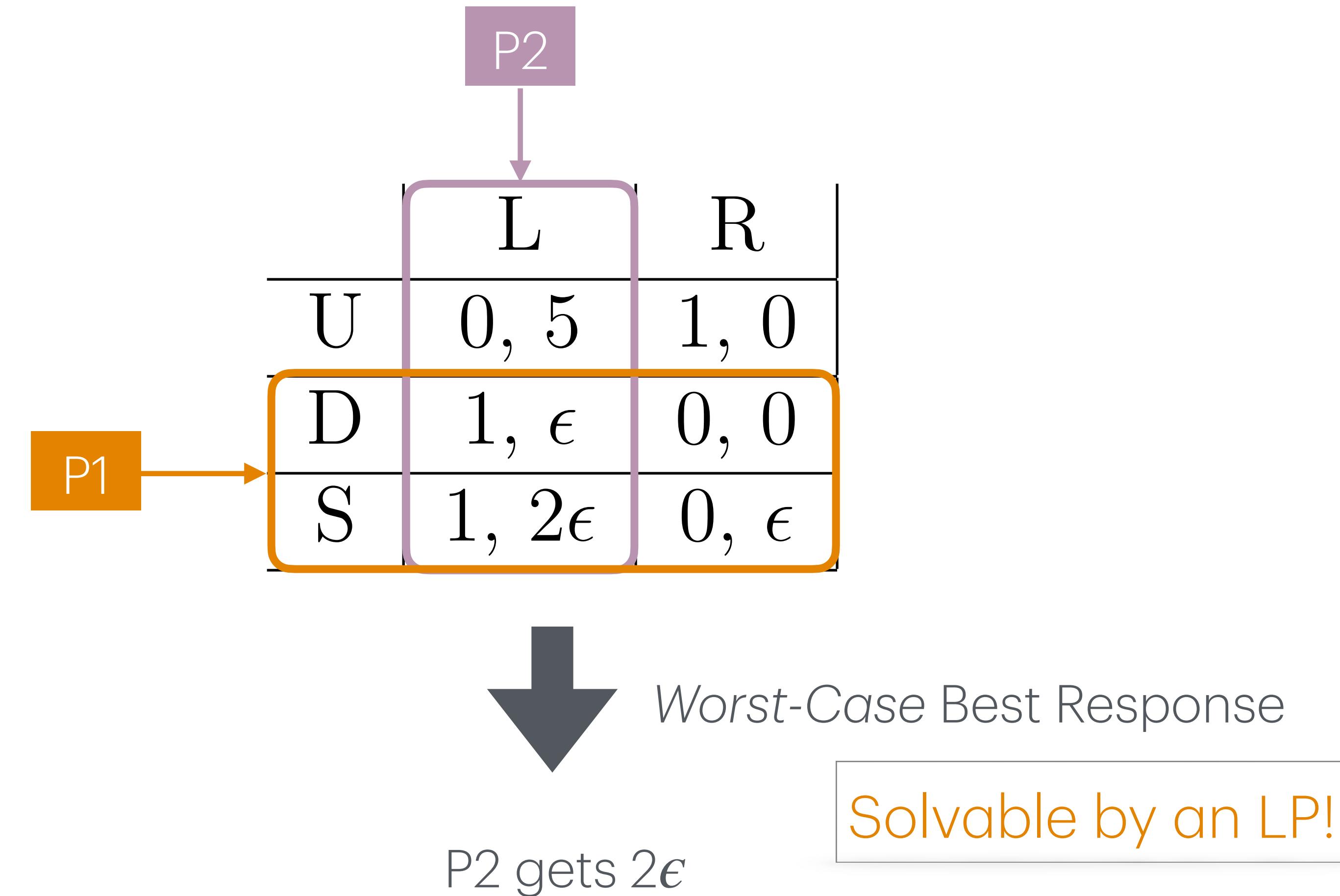
Algorithm

P2 fakes L



Algorithm

P2 fakes L



Conclusion

Conclusion



Misinformation attacks can be computed in polynomial time by exploiting rationality.

Conclusion



Misinformation attacks can be computed in polynomial time by exploiting rationality.

New mechanisms are needed to combat misinformation attacks!