

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

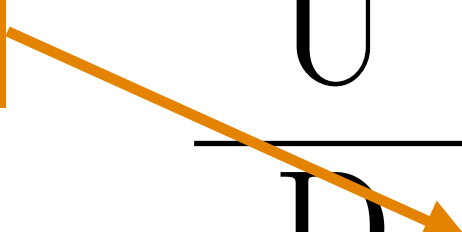
	L	R
U	0, 5	1, 0
D	1, ϵ	0, 0
S	1, 0	0, ϵ

Example

True Game

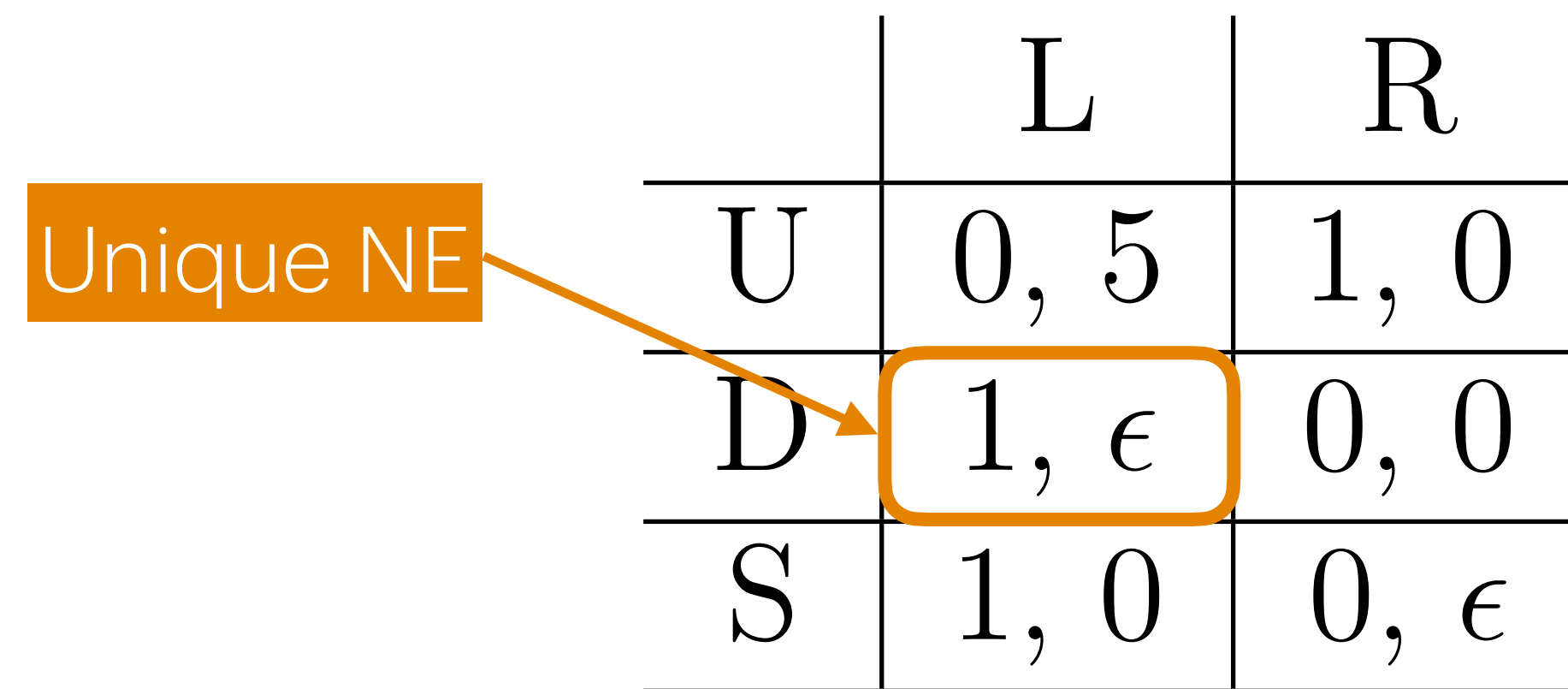
	L	R
U	0, 5	1, 0
D	1, ϵ	0, 0
S	1, 0	0, ϵ

Unique NE



Example

True Game



	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
P2 wants	U	0, 5	1, 0
Unique NE	D	1, ϵ	0, 0
	S	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+ ϵ
D	1, ϵ	0, 2 ϵ
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



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

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

P2 wins! →

Unique NE →

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

Example

Faked Game

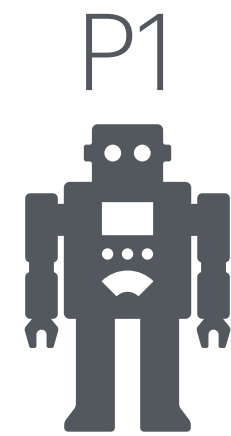
“Inception Attack”



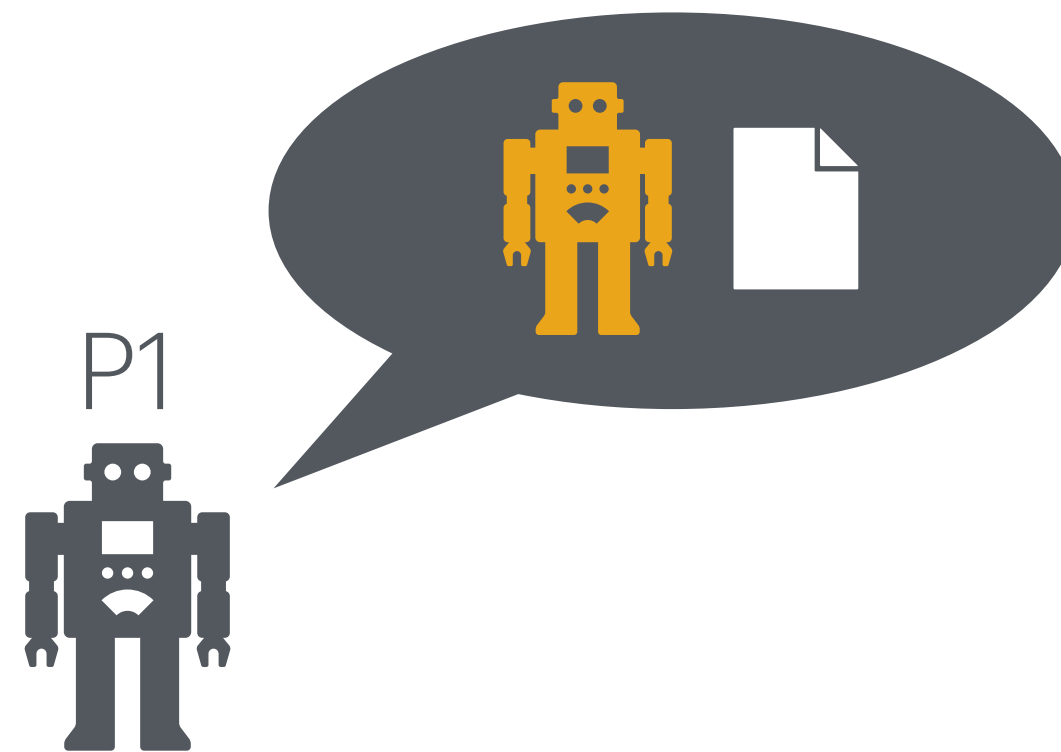
Inception attacks are powerful,
but can they be computed?

Inception Approach

Inception Approach



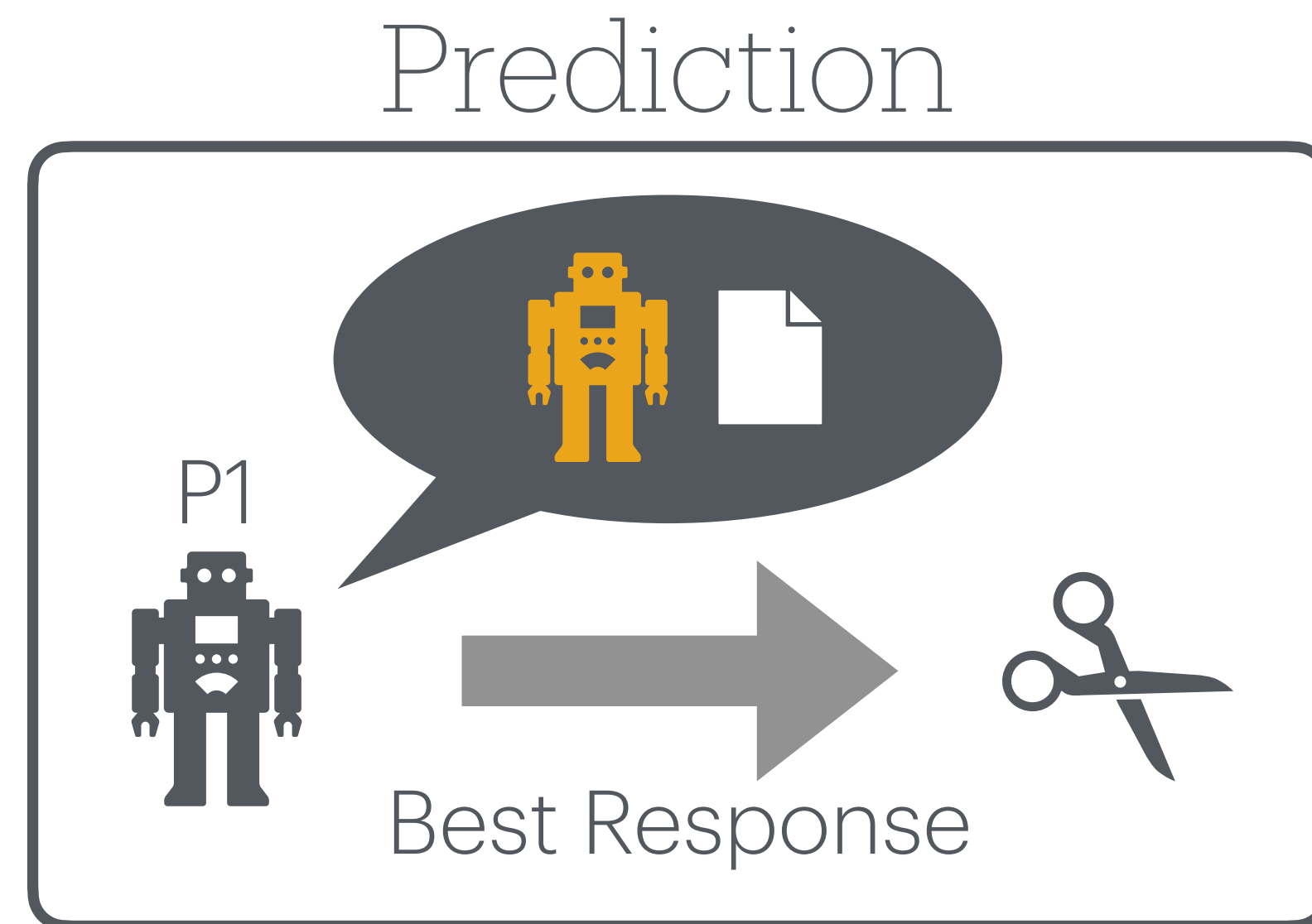
Inception Approach



Inception Approach

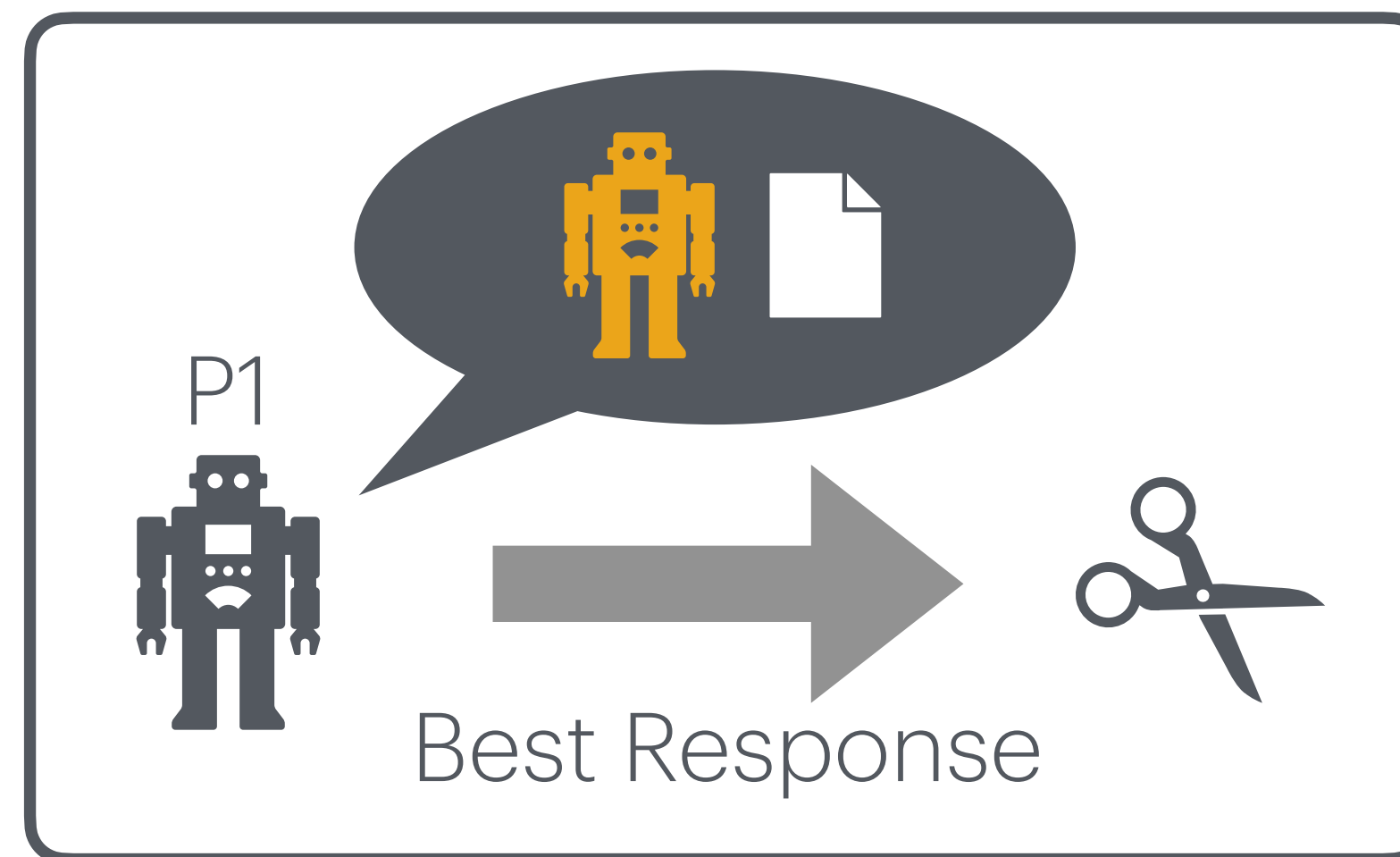


Inception Approach

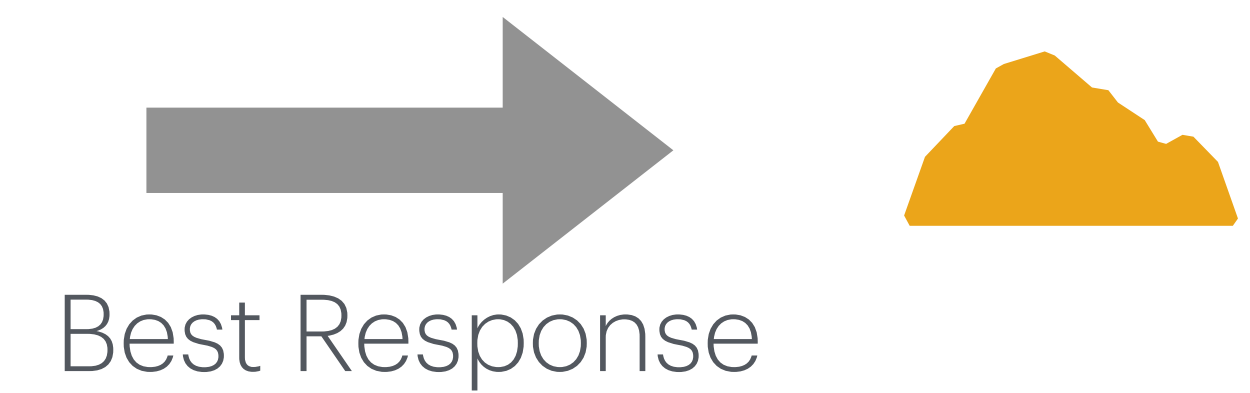


Inception Approach

Prediction

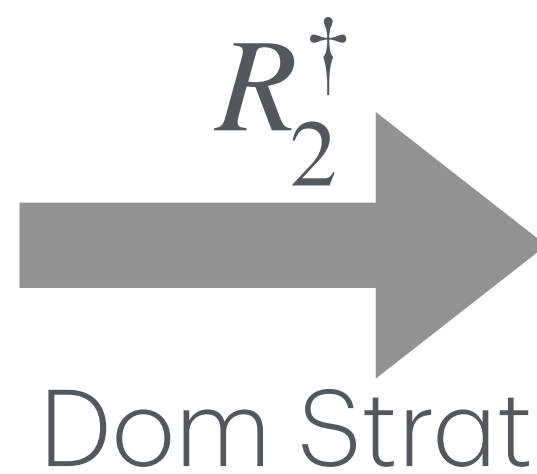
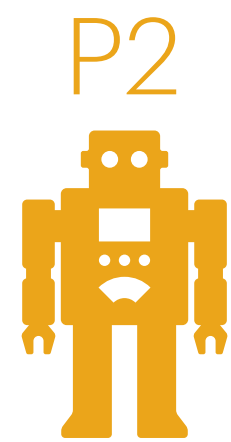


Exploitation



Inception Approach

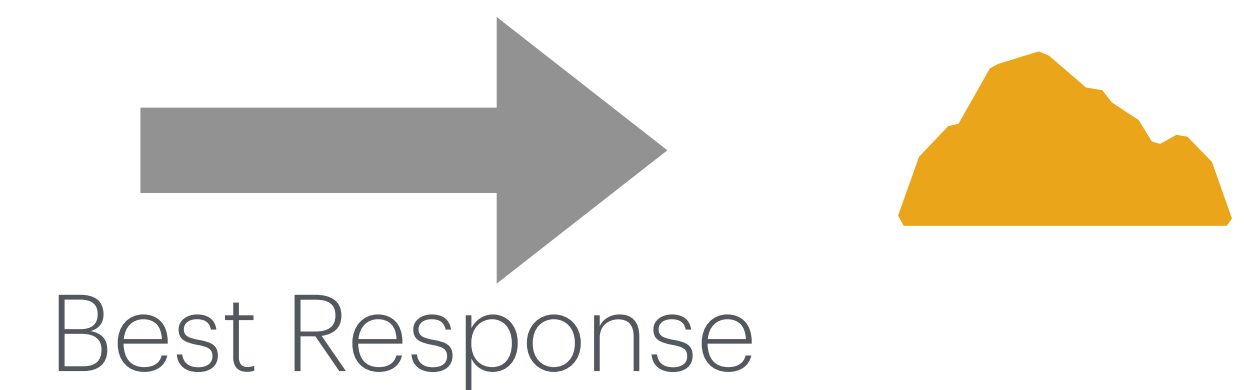
Convincing



Prediction

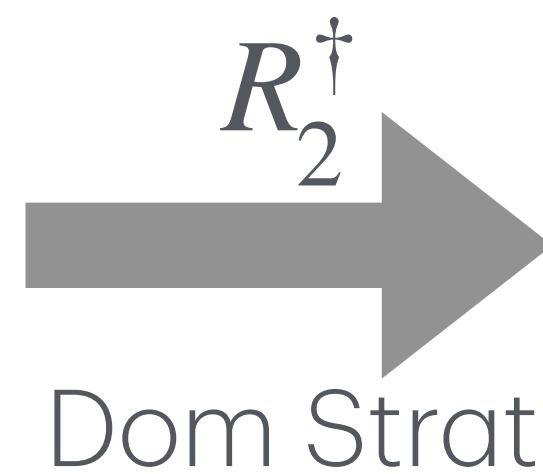
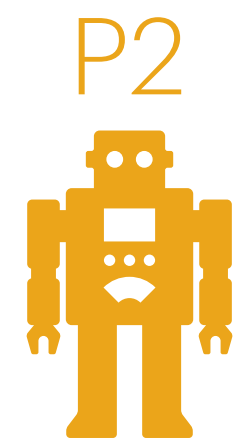


Exploitation

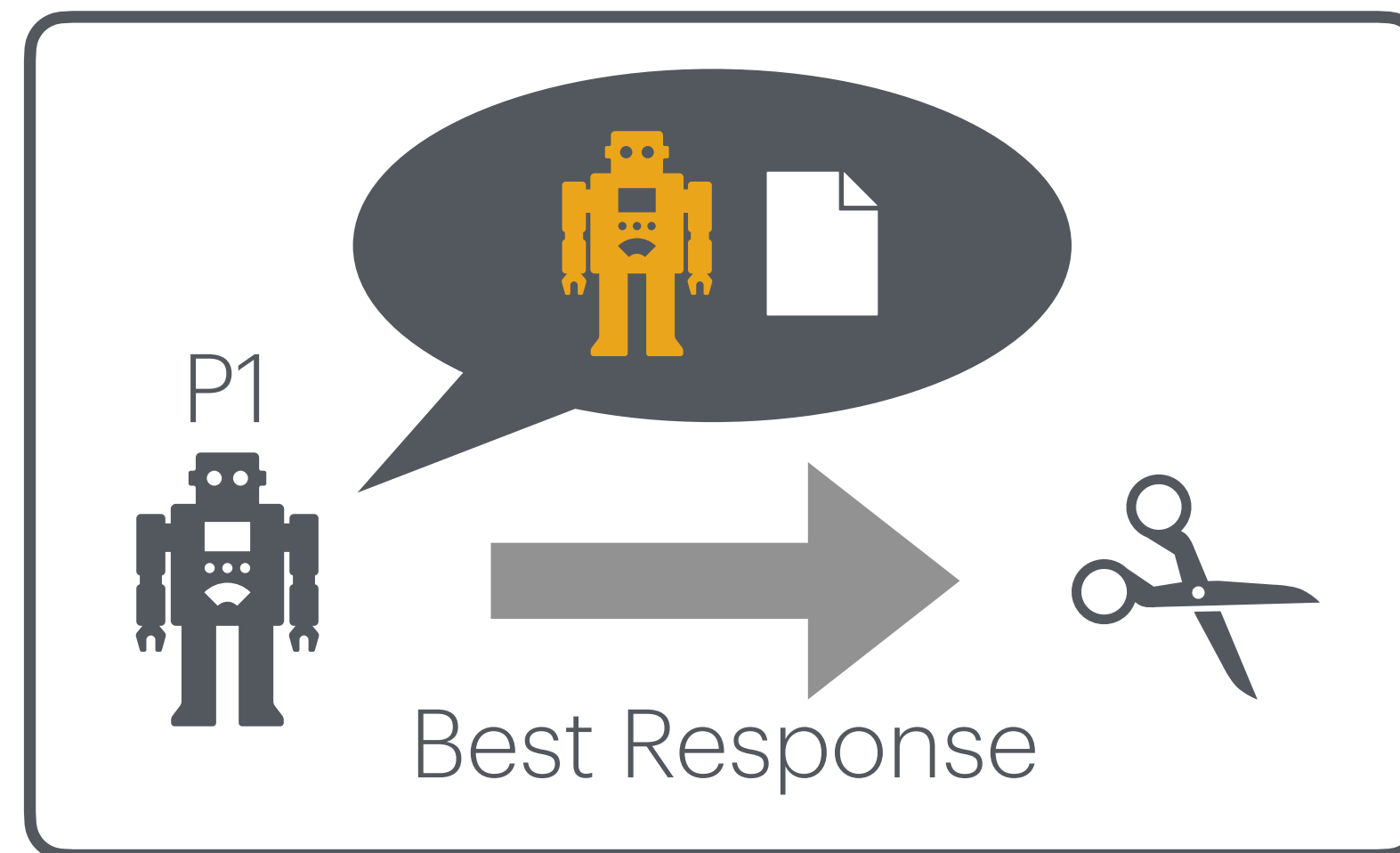


Inception Approach

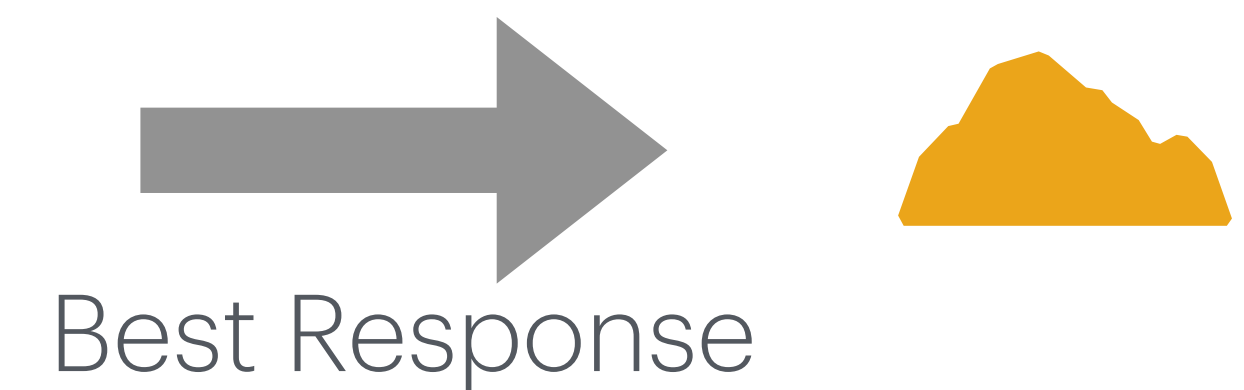
Convincing



Prediction



Exploitation



Repeat to find the best pure strategy inception!

Algorithm

P2 fakes R

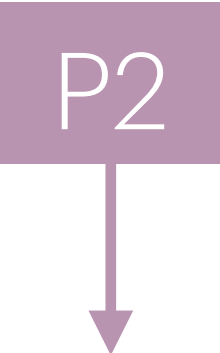
Algorithm

P2 fakes R

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

Algorithm

P2 fakes R



A purple square labeled 'P2' has a purple arrow pointing down to the 'R' column of the game matrix.

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

Algorithm

P2 fakes R

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

Algorithm

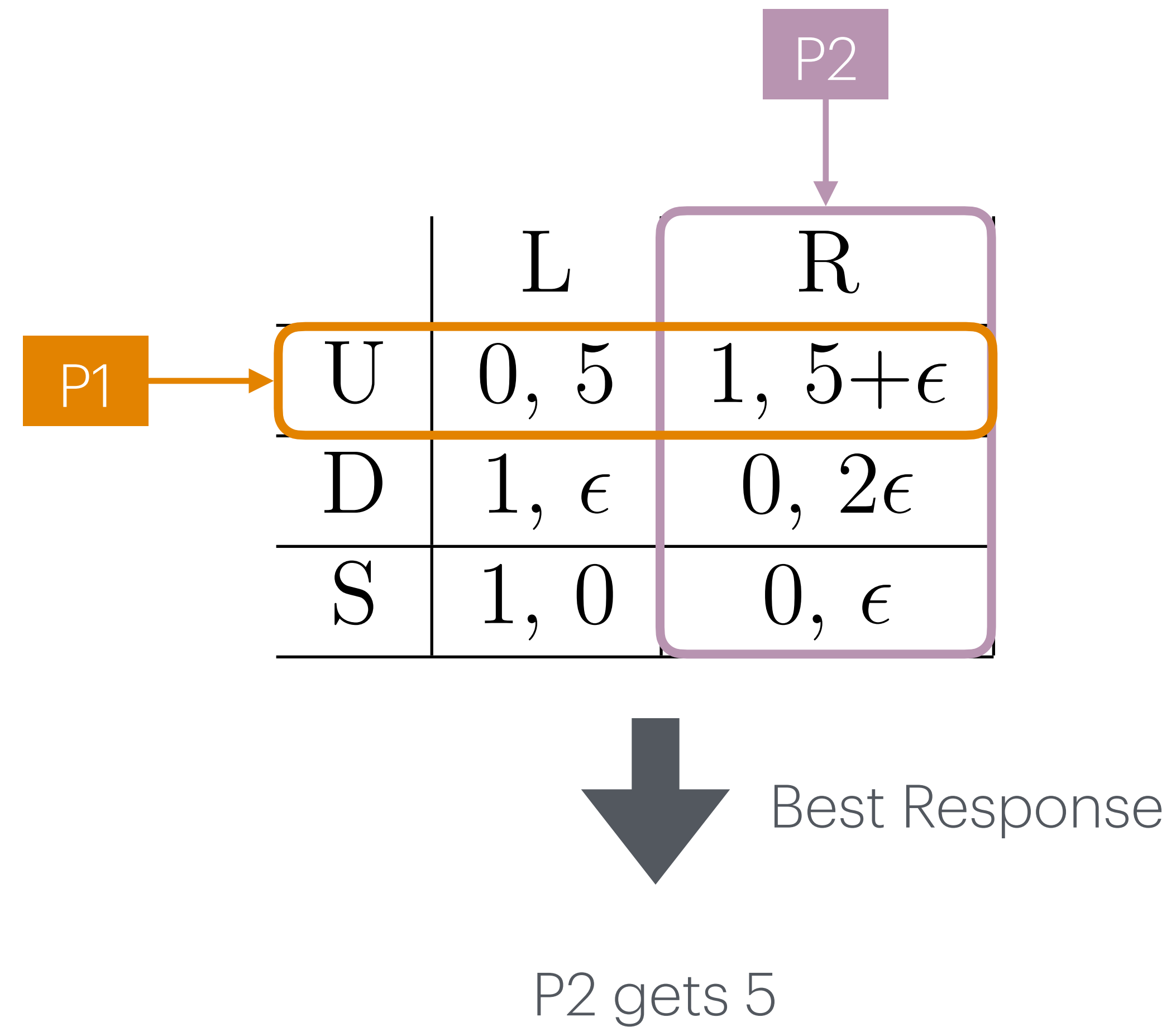
P2 fakes R

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

Best Response

Algorithm

P2 fakes R



Algorithm

P2 fakes L

Algorithm

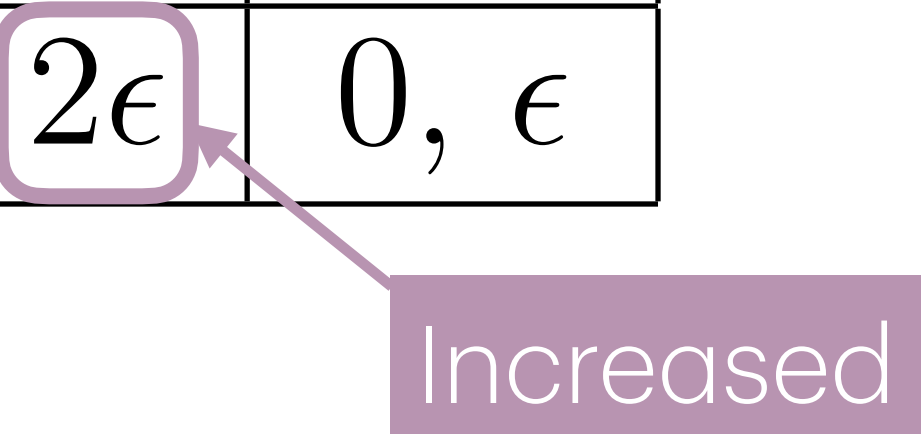
P2 fakes L

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

Algorithm

P2 fakes L

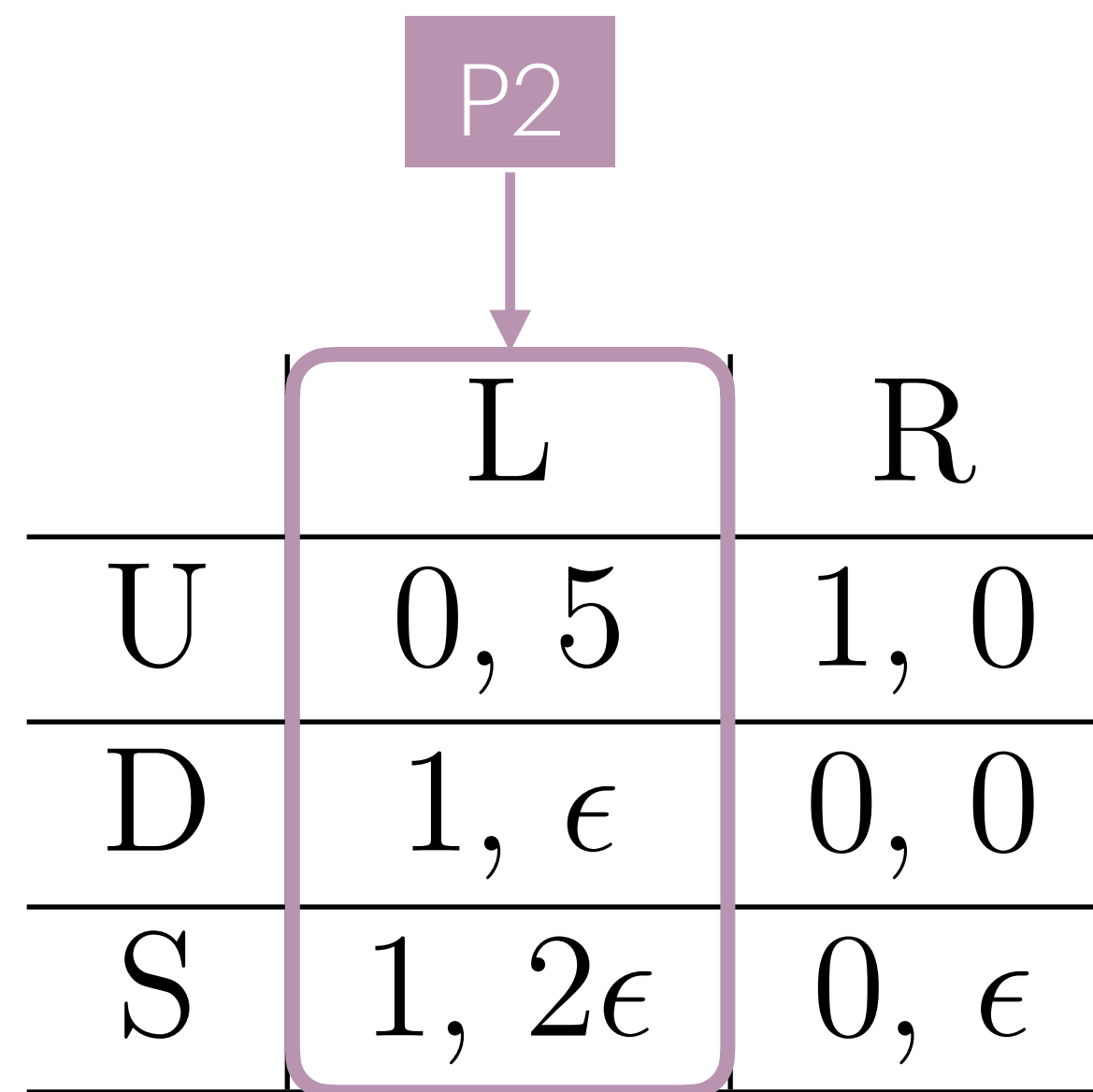
	L	R
U	0, 5	1, 0
D	1, ϵ	0, 0
S	1, 2ϵ	0, ϵ



Increased

Algorithm

P2 fakes L



A diagram illustrating a game matrix. A purple box labeled 'P2' has a downward arrow pointing to the first column of a 3x2 matrix. The first column is highlighted with a purple rounded rectangle. The matrix has rows labeled U, D, and S, and columns labeled L and R. The payoffs are as follows:

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

Algorithm

P2 fakes L

A game tree diagram illustrating a sequential game between Player 1 (P1) and Player 2 (P2). Player 2 moves first, choosing between L and R. Player 1 then moves, choosing between U, D, and S. The payoffs are given as (P1, P2).

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

Algorithm

P2 fakes L

A game tree diagram illustrating a game between Player 1 (P1) and Player 2 (P2). Player 2 moves first, choosing between L and R. Player 1 then moves, choosing between U, D, and S. The payoffs are given as (P1, P2).

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

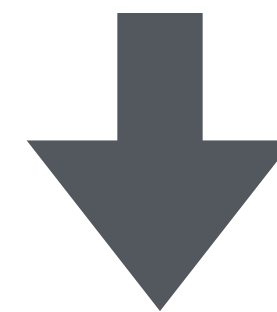
P1

Tie!

Algorithm

P2 fakes L

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

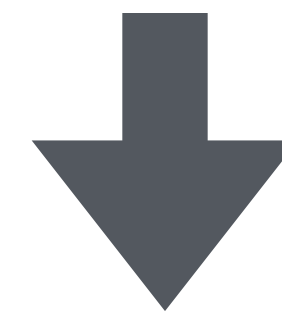


Worst-Case Best Response

Algorithm

P2 fakes L

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



Worst-Case Best Response

P2 gets 2ϵ

Algorithm

P2 fakes L

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



Worst-Case Best Response

P2 gets 2ϵ

Solvable by an LP!

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!