

Safe Pareto Improvements for Delegated Game Playing

Caspar Oesterheld and Vincent Conitzer

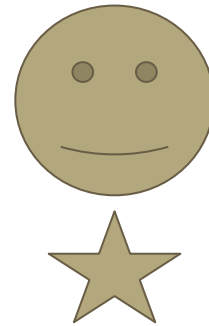


Paper published at AAMAS 2021 and JAAMAS 36 (2022).

Blackmail and Surrogate Goals

Keerti Anand
Caspar Oesterheld

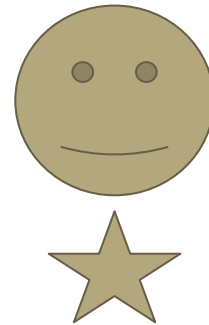
Background: Blackmail



Background: Blackmail

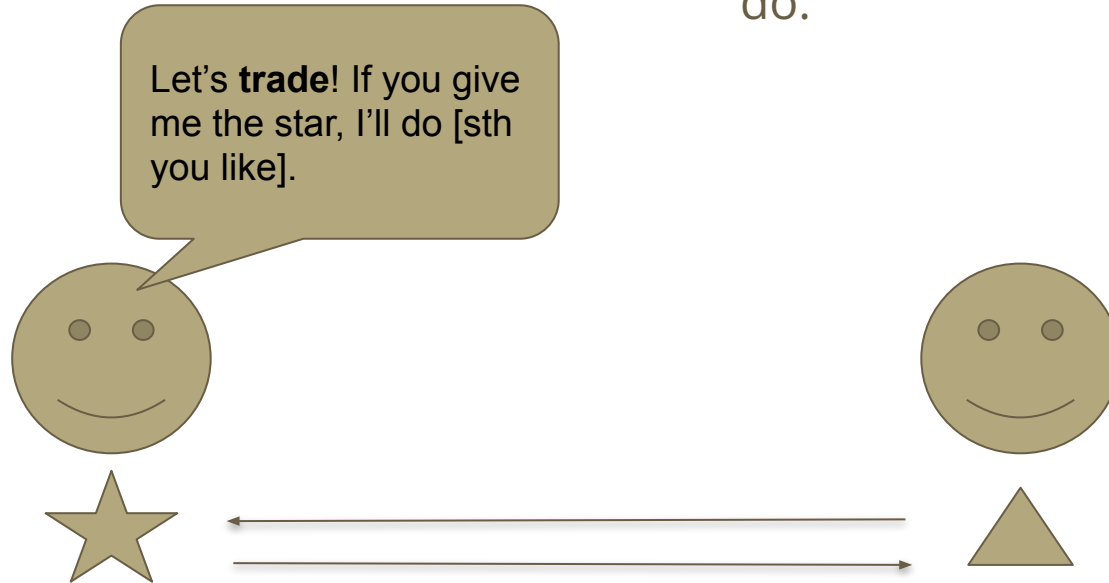


- Has to be credible.
- Offer must be something Player 1 wouldn't otherwise do.



Background: Blackmail

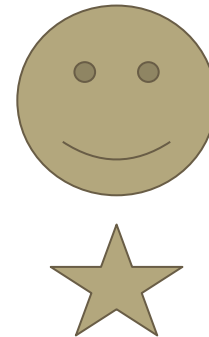
- Has to be credible.
- Offer must be something Player 1 wouldn't otherwise do.



Background: Blackmail

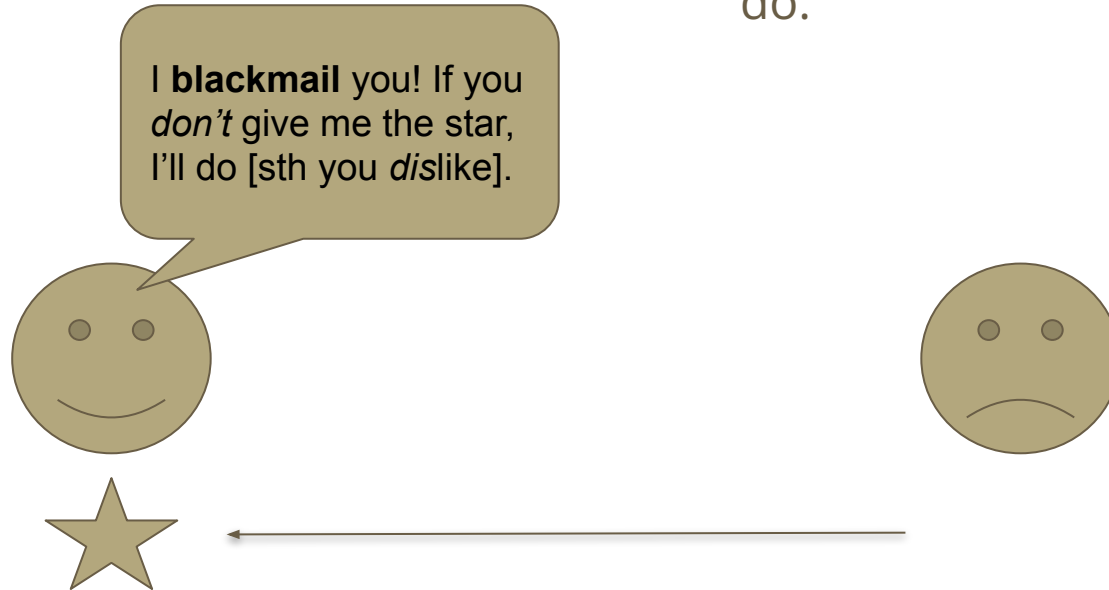


- Has to be credible
- Threat must be something the threatener wouldn't otherwise do.



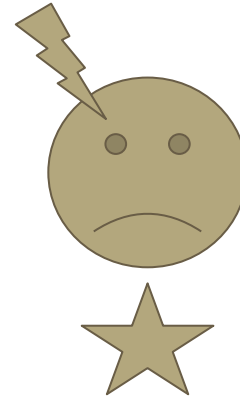
Background: Blackmail

- Has to be credible
- Threat must be something the threatener wouldn't otherwise do.



Background: Blackmail

- Has to be credible
- Threat must be something I wouldn't otherwise do.



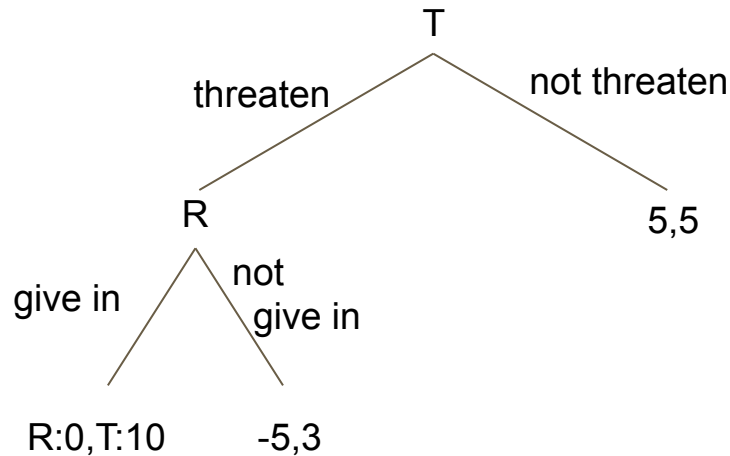
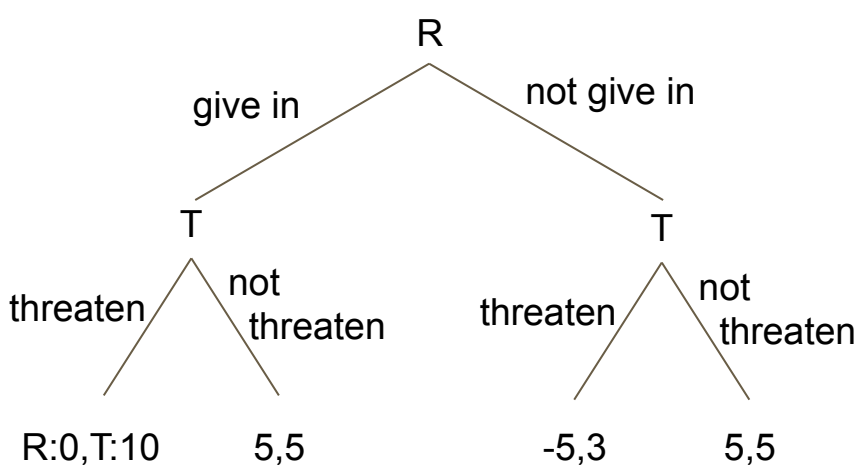
Blackmail as a normal-form game

| | Give in | Not give in |
|--------------|---------|-------------|
| Threaten | 10,0 | 3,-5 |
| Not threaten | 5,5 | 5,5 |

Blackmail as a normal-form game

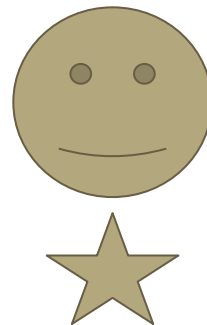
| | Give in | Not give in |
|--------------|---------|-------------|
| Threaten | 10,0 | 3,-5 |
| Not threaten | 5,5 | 5,5 |

Blackmail as an extensive-form game: Winning by being the first to commit?

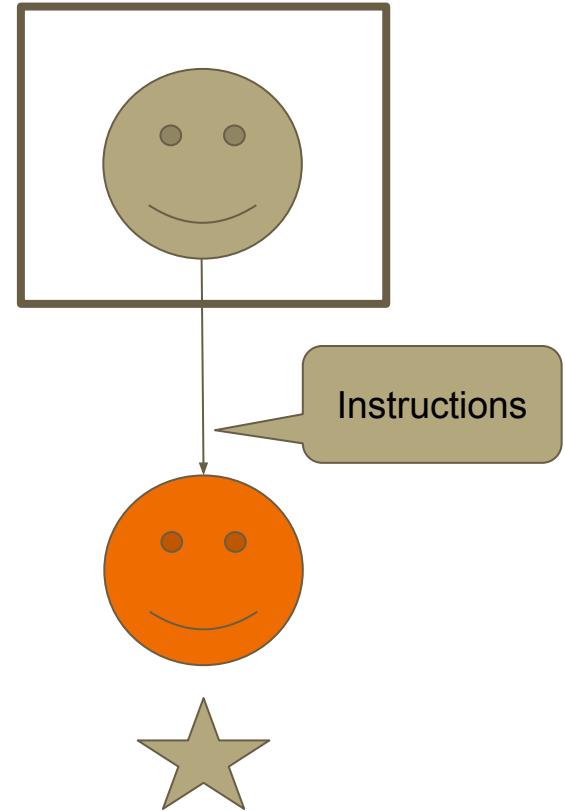


- Why care about subgame-perfection?
- "Teaching", reputation
- Threats are carried out in practice

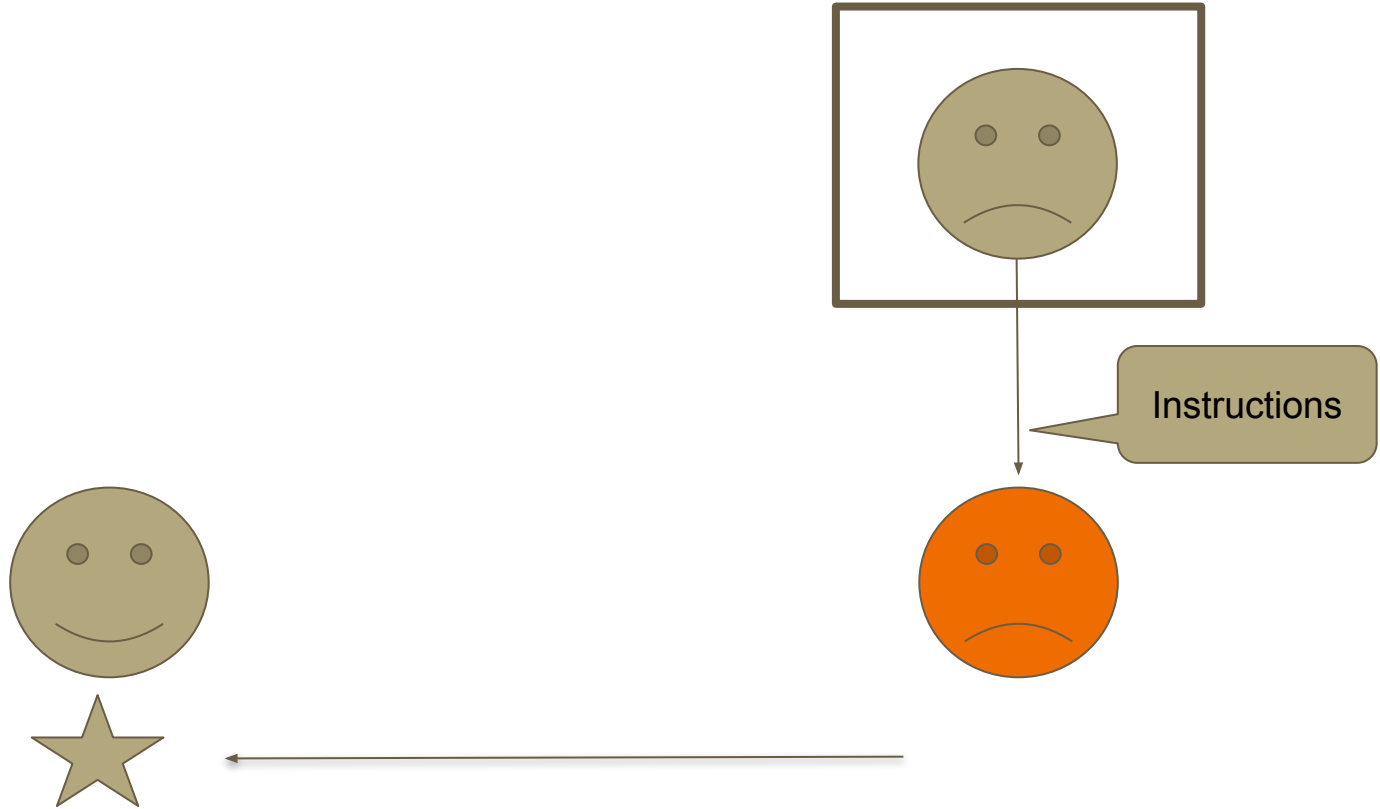
A unilateral safe Pareto improvement



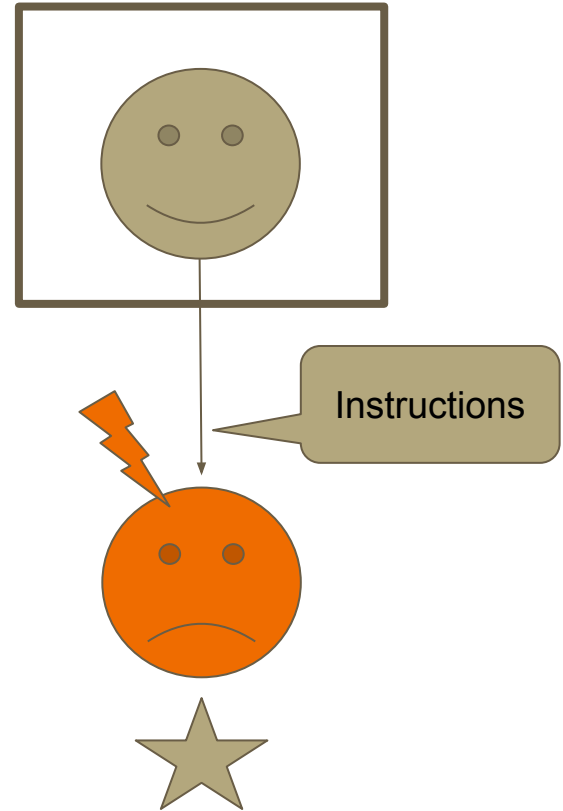
A unilateral safe Pareto improvement







A unilateral safe Pareto improvement



A unilateral safe Pareto improvement



| | Give in to A threats | Give in to B threats | Not give in |
|--|----------------------|----------------------|-------------|
| Threaten A  | 10,0 | 3,-5 | 3,-5 |
| Threaten B  | 3,5 | 10,0 | 3,5 |
| Not threaten | 5,5 | 5,5 | 5,5 |

| | Give in to A threats | Give in to B threats | Not give in |
|--|----------------------|----------------------|---------------------|
| Threaten A  | 10,0 | 3,-5 | 3,-5 |
| Threaten B  | 3,5 | 10,0 10, 0 | 3,5 3,- 5 |
| Not threaten | 5,5 | 5,5 5, 5 | 5,5 5, 5 |

Safe Pareto Improvement (SPI): If Player 1 and Player 2's representative play the bottom rather than top game, this is guaranteed to be (weakly) Pareto-better for the original players!

Results

- Formal justification for surrogate goals
 - Threatener “has no reason” (in some formal sense) to punish use of surrogate goals (under some assumptions)
 - Therefore, the resource holder prefers to employ the surrogate goal method so that the threat against the original goal is never carried out.
- As a future direction, we might be able to show that if both players choose to employ surrogate goals then we can avoid the worst possible scenario (threaten, not give in), which works well for both players in the sense that no resources are wasted.

In the paper...

- Bilateral SPIs
- All safe Pareto improvements are based on a specific notion of outcome correspondence between games.
- Specific assumptions for deriving SPIs (iterated strict dominance, isomorphisms)
- Computational complexity (given a game, find the SPIs using the assumptions)
 - NP-complete by reduction from the subgraph isomorphism problem
 - Solvable by linear programming in alternative setup

Thank You



Keerti, Caspar.
Kindly stop
talking now !!