Introduction to Mechanism Design

Mehdi Dastani
BBL-521
M.M.Dastani@cs.uu.nl

Thanks to Paul Harrenstein and Mathijs de Weerdt who provided me some of these slides.

What is Mechanism Design Trying to Accomplish?

- Takahashi wishes to sell a precious Ming vase to either Ann or Bob, depending on who values it most.
- Yet, Takahashi does not know the valuations of either Ann or Bob.
- Hence, the goal he sets himself is to sell the vase to Ann if she values it most, and to Bob otherwise. (Social choice function).
- Then, he tries to construct a protocol (mechanism) that guarantees the vase to be sold to the one who values it most, *for all possible valuations Ann and Bob may have.* (Implement the social choice function).

What is Mechanism Design Trying to Accomplish?

- Fix a set of possible outcomes O.
- Given a preference profile $(\succsim_1, \ldots, \succsim_n)$ certain outcomes $O^* \in O$ are more desirable than others from an outsider's point of view (e.g., assigning an object to the agent that values it most).
- However, preferences of the players are unknown to the designer, or hard to obtain.
- Issue: *Design a game (mechanism) such that the outcome given a particular (fixed) solution concept of this game generates one of the desired outcomes in O^* for all (relevant) preference profiles* $(\succsim_1, \ldots, \succsim_n)$.

Strategic Behavior and the Importance of Truthfulness

Principles of voting make an election more of a game of skill than a real test of the wishes of the electors. My own opinion is that it is better for elections to be decided according to the wish of the majority than of those who happen to have most skill at the game.

(C.L. Dodgson)

Strategic Behavior and the Importance of Truthfulness

Consider the Borda rule.

$\succsim_1$	$\succsim_2$	$\succsim_3$
a	a	d
b	b	c
c	d	b
d	c	a

$\succsim_1$	$\succsim_2$	$\succsim'_3$
a	a	b
b	b	c
c	d	d
d	c	a

- Borda winner given $(\succsim_1, \succsim_2, \succsim_3)$ is a with 6 points
- Borda winner given $(\succsim_1, \succsim_2, \succsim'_3)$ is b with 7 points

Conclusion: *If $\succsim_3$ are player 3's true preferences, he had better lie about them!*

Social Choice Functions and Strategic Game Forms

- Preferences as strategies (lie or tell the truth)
- Combination of such strategies is a preference profile
- Outcomes are defined by a social choice function f
- Social choice functions reflect how to determine the desired (by the designer) outcomes relative to the agents' preferences. E.g., assigning the object to the agent that values it most
- The true preferences of the agents are assumed to be unknown to the designer
- Each possible preference profile $\succsim$ yields such a game

Social Choice Functions and Strategic Game Forms

	$\succsim_2$	$\succsim_2'$
$\succsim_1$	$f(\succsim_1, \succsim_2)$	$f(\succsim_1, \succsim_2')$
$\succsim_1'$	$f(\succsim_1', \succsim_2)$	$f(\succsim_1', \succsim_2')$

Social Choice Functions and Strategic Game Forms

$(\succsim_1, \succsim_2)$	$\succsim_2$	$\succsim_2'$
$\succsim_1$	$f(\succsim_1, \succsim_2)$	$f(\succsim_1, \succsim_2')$
$\succsim_1'$	$f(\succsim_1', \succsim_2)$	$f(\succsim_1', \succsim_2')$

$(\succsim_1, \succsim_2')$	$\succsim_2$	$\succsim_2'$
$\succsim_1$	$f(\succsim_1, \succsim_2)$	$f(\succsim_1, \succsim_2')$
$\succsim_1'$	$f(\succsim_1', \succsim_2)$	$f(\succsim_1', \succsim_2')$

$(\succsim_1', \succsim_2)$	$\succsim_2$	$\succsim_2'$
$\succsim_1$	$f(\succsim_1, \succsim_2)$	$f(\succsim_1, \succsim_2')$
$\succsim_1'$	$f(\succsim_1', \succsim_2)$	$f(\succsim_1', \succsim_2')$

$(\succsim_1', \succsim_2')$	$\succsim_2$	$\succsim_2'$
$\succsim_1$	$f(\succsim_1, \succsim_2)$	$f(\succsim_1, \succsim_2')$
$\succsim_1'$	$f(\succsim_1', \succsim_2)$	$f(\succsim_1', \succsim_2')$

Social Choice Functions and Strategic Game Forms

$(\succsim_1, \succsim_2)$	$\succsim_2$	$\succsim_2'$
$\succsim_1$	$f(\succsim_1, \succsim_2)$	$f(\succsim_1, \succsim_2')$
$\succsim_1'$	$f(\succsim_1', \succsim_2)$	$f(\succsim_1', \succsim_2')$

$(\succsim_1, \succsim_2')$	$\succsim_2$	$\succsim_2'$
$\succsim_1$	$f(\succsim_1, \succsim_2)$	$f(\succsim_1, \succsim_2')$
$\succsim_1'$	$f(\succsim_1', \succsim_2)$	$f(\succsim_1', \succsim_2')$

$(\succsim_1', \succsim_2)$	$\succsim_2$	$\succsim_2'$
$\succsim_1$	$f(\succsim_1, \succsim_2)$	$f(\succsim_1, \succsim_2')$
$\succsim_1'$	$f(\succsim_1', \succsim_2)$	$f(\succsim_1', \succsim_2')$

$(\succsim_1', \succsim_2')$	$\succsim_2$	$\succsim_2'$
$\succsim_1$	$f(\succsim_1, \succsim_2)$	$f(\succsim_1, \succsim_2')$
$\succsim_1'$	$f(\succsim_1', \succsim_2)$	$f(\succsim_1', \succsim_2')$

Social Choice Functions and Strategic Game Forms

$(\succsim_1, \succsim_2)$	$\succsim_2$	$\succsim_2'$
$\succsim_1$	$f(\succsim_1, \succsim_2)$	$f(\succsim_1, \succsim_2')$
$\succsim_1'$	$f(\succsim_1', \succsim_2)$	$f(\succsim_1', \succsim_2')$

$(\succsim_1, \succsim_2')$	$\succsim_2$	$\succsim_2'$
$\succsim_1$	$f(\succsim_1, \succsim_2)$	$f(\succsim_1, \succsim_2')$
$\succsim_1'$	$f(\succsim_1', \succsim_2)$	$f(\succsim_1', \succsim_2')$

$(\succsim_1', \succsim_2)$	$\succsim_2$	$\succsim_2'$
$\succsim_1$	$f(\succsim_1, \succsim_2)$	$f(\succsim_1, \succsim_2')$
$\succsim_1'$	$f(\succsim_1', \succsim_2)$	$f(\succsim_1', \succsim_2')$

$(\succsim_1', \succsim_2')$	$\succsim_2$	$\succsim_2'$
$\succsim_1$	$f(\succsim_1, \succsim_2)$	$f(\succsim_1, \succsim_2')$
$\succsim_1'$	$f(\succsim_1', \succsim_2)$	$f(\succsim_1', \succsim_2')$

Social Choice Functions and Strategic Game Forms

$(\succsim_1, \succsim_2)$	$\succsim_2$	$\succsim_2'$
$\succsim_1$	$f(\succsim_1, \succsim_2)$	$f(\succsim_1, \succsim_2')$
$\succsim_1'$	$f(\succsim_1', \succsim_2)$	$f(\succsim_1', \succsim_2')$

$(\succsim_1, \succsim_2')$	$\succsim_2$	$\succsim_2'$
$\succsim_1$	$f(\succsim_1, \succsim_2)$	$f(\succsim_1, \succsim_2')$
$\succsim_1'$	$f(\succsim_1', \succsim_2)$	$f(\succsim_1', \succsim_2')$

$(\succsim_1', \succsim_2)$	$\succsim_2$	$\succsim_2'$
$\succsim_1$	$f(\succsim_1, \succsim_2)$	$f(\succsim_1, \succsim_2')$
$\succsim_1'$	$f(\succsim_1', \succsim_2)$	$f(\succsim_1', \succsim_2')$

$(\succsim_1', \succsim_2')$	$\succsim_2$	$\succsim_2'$
$\succsim_1$	$f(\succsim_1, \succsim_2)$	$f(\succsim_1, \succsim_2')$
$\succsim_1'$	$f(\succsim_1', \succsim_2)$	$f(\succsim_1', \succsim_2')$

Social Choice Functions and Strategic Game Forms

$(\succsim_1, \succsim_2)$	$\succsim_2$	$\succsim_2'$
$\succsim_1$	$f(\succsim_1, \succsim_2)$	$f(\succsim_1, \succsim_2')$
$\succsim_1'$	$f(\succsim_1', \succsim_2)$	$f(\succsim_1', \succsim_2')$

$(\succsim_1, \succsim_2')$	$\succsim_2$	$\succsim_2'$
$\succsim_1$	$f(\succsim_1, \succsim_2)$	$f(\succsim_1, \succsim_2')$
$\succsim_1'$	$f(\succsim_1', \succsim_2)$	$f(\succsim_1', \succsim_2')$

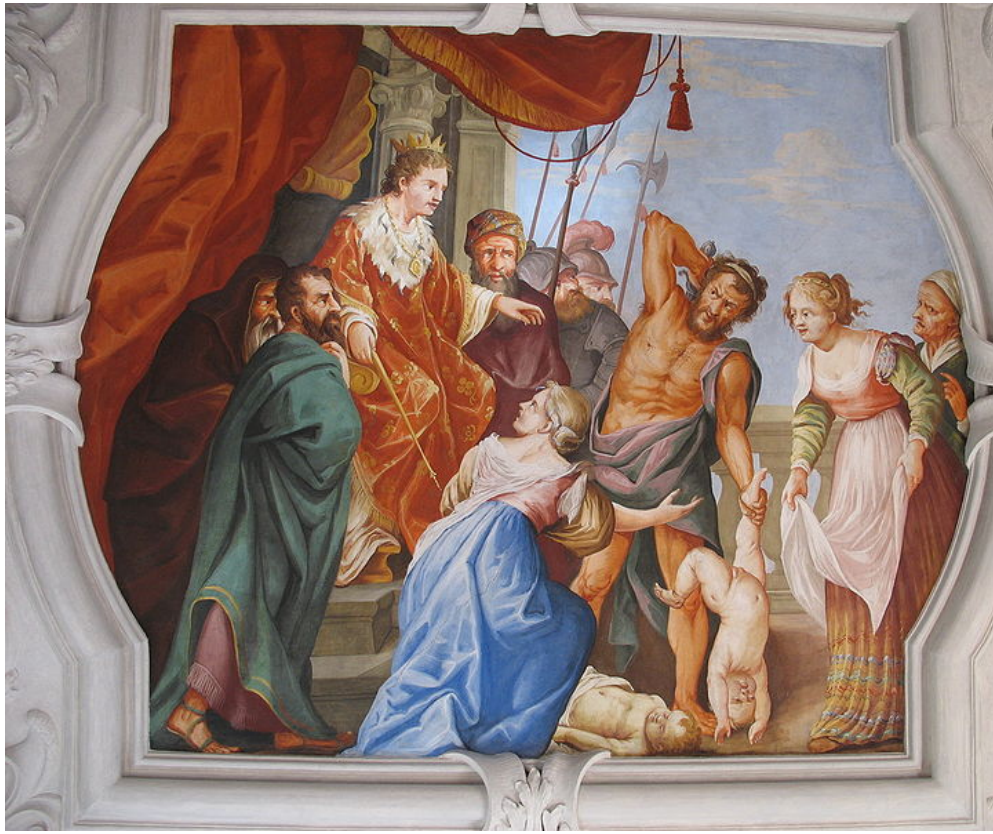
$(\succsim_1', \succsim_2)$	$\succsim_2$	$\succsim_2'$
$\succsim_1$	$f(\succsim_1, \succsim_2)$	$f(\succsim_1, \succsim_2')$
$\succsim_1'$	$f(\succsim_1', \succsim_2)$	$f(\succsim_1', \succsim_2')$

$(\succsim_1', \succsim_2')$	$\succsim_2$	$\succsim_2'$
$\succsim_1$	$f(\succsim_1, \succsim_2)$	$f(\succsim_1, \succsim_2')$
$\succsim_1'$	$f(\succsim_1', \succsim_2)$	$f(\succsim_1', \succsim_2')$

Example: Solomon's Verdict

He sent for a sword, and when it was brought, he said, "Cut the living child in two and give each woman half of it". The real mother, her heart full of love for her son, said to the king, "Please, Your Majesty, don't kill the child! Give it to her!" But the other woman said, "Don't give it to either of us; go on and cut it in two". Then Solomon said, "Don't kill the child! Give it to the first woman, she is its real mother."

*(*1 Kings 3: 16-28*)*



Example: Solomon's Verdict

- Three outcomes:

  a: First woman gets the baby.
  b: Second woman gets the baby.
  c: The baby is bisected.

- For each woman, two possible types (good mother, bad mother):

$\succsim_1$: $a \succ b \succ c$ $\quad$ $\succsim_2$: $b \succ a \succ c$
$\succsim_1'$: $a \succ c \succ b$ $\quad$ $\succsim_2'$: $b \succ c \succ a$

	$\succsim_2$	$\succsim_2'$
$\succsim_1$	c	**a**
$\succsim_1'$	b	c

	$\succsim_2$	$\succsim_2'$
$\succsim_1$	c	a
$\succsim_1'$	**b**	c

Example: Solomon's Verdict

	$\succsim_2$	$\succsim_2'$
$\succsim_1$	c	a
$\succsim_1'$	b	c

$\succsim_1$	$\succsim_2'$
a	b
b	c
c	a

Observation: *Be the true types given by* $(\succsim_1, \succsim_2')$*, revealing her true preferences is* ***not*** *a dominant strategy for the second woman (the bad mother)!*

Example: Solomon's Verdict

	$\succsim_2$	$\succsim_2'$
$\succsim_1$	c	a
$\succsim_1'$	b	c

$\succsim_1$	$\succsim_2'$
a	b
b	c
c	a

Observation: *Be the true types given by* $(\succsim_1, \succsim_2')$*, revealing her true preferences is* ***not*** *a dominant strategy for the second woman (the bad mother)!*

Example: Solomon's Verdict

	$\succsim_2$	$\succsim_2'$
$\succsim_1$	c	a
$\succsim_1'$	b	c

$\succsim_1$	$\succsim_2'$
a	b
b	c
c	a

Observation: *Be the true types given by* $(\succsim_1, \succsim_2')$*, revealing her true preferences is* ***not*** *a dominant strategy for the second woman (the bad mother)!*

Incentive Compatibility of Social Choice Functions

$(\succsim_1, \succsim_2)$	$\succsim_2$	$\succsim_2'$
$\succsim_1$	$f(\succsim_1, \succsim_2)$	$f(\succsim_1, \succsim_2')$
$\succsim_1'$	$f(\succsim_1', \succsim_2)$	$f(\succsim_1', \succsim_2')$

$(\succsim_1, \succsim_2')$	$\succsim_2$	$\succsim_2'$
$\succsim_1$	$f(\succsim_1, \succsim_2)$	$f(\succsim_1, \succsim_2')$
$\succsim_1'$	$f(\succsim_1', \succsim_2)$	$f(\succsim_1', \succsim_2')$

$(\succsim_1', \succsim_2)$	$\succsim_2$	$\succsim_2'$
$\succsim_1$	$f(\succsim_1, \succsim_2)$	$f(\succsim_1, \succsim_2')$
$\succsim_1'$	$f(\succsim_1', \succsim_2)$	$f(\succsim_1', \succsim_2')$

$(\succsim_1', \succsim_2')$	$\succsim_2$	$\succsim_2'$
$\succsim_1$	$f(\succsim_1, \succsim_2)$	$f(\succsim_1, \succsim_2')$
$\succsim_1'$	$f(\succsim_1', \succsim_2)$	$f(\succsim_1', \succsim_2')$

Incentive Compatibility of Social Choice Functions

$(\succsim_1, \succsim_2)$	$\succsim_2$	$\succsim_2'$
$\succsim_1$	$f(\succsim_1, \succsim_2)$	$f(\succsim_1, \succsim_2')$
$\succsim_1'$	$f(\succsim_1', \succsim_2)$	$f(\succsim_1', \succsim_2')$

$(\succsim_1, \succsim_2')$	$\succsim_2$	$\succsim_2'$
$\succsim_1$	$f(\succsim_1, \succsim_2)$	$f(\succsim_1, \succsim_2')$
$\succsim_1'$	$f(\succsim_1', \succsim_2)$	$f(\succsim_1', \succsim_2')$

$(\succsim_1', \succsim_2)$	$\succsim_2$	$\succsim_2'$
$\succsim_1$	$f(\succsim_1, \succsim_2)$	$f(\succsim_1, \succsim_2')$
$\succsim_1'$	$f(\succsim_1', \succsim_2)$	$f(\succsim_1', \succsim_2')$

$(\succsim_1', \succsim_2')$	$\succsim_2$	$\succsim_2'$
$\succsim_1$	$f(\succsim_1, \succsim_2)$	$f(\succsim_1, \succsim_2')$
$\succsim_1'$	$f(\succsim_1', \succsim_2)$	$f(\succsim_1', \succsim_2')$

Incentive Compatibility of Social Choice Functions

$(\succsim_1, \succsim_2)$	$\succsim_2$	$\succsim_2'$
$\succsim_1$	$f(\succsim_1, \succsim_2)$	$f(\succsim_1, \succsim_2')$
$\succsim_1'$	$f(\succsim_1', \succsim_2)$	$f(\succsim_1', \succsim_2')$

$(\succsim_1, \succsim_2')$	$\succsim_2$	$\succsim_2'$
$\succsim_1$	$f(\succsim_1, \succsim_2)$	$f(\succsim_1, \succsim_2')$
$\succsim_1'$	$f(\succsim_1', \succsim_2)$	$f(\succsim_1', \succsim_2')$

$(\succsim_1', \succsim_2)$	$\succsim_2$	$\succsim_2'$
$\succsim_1$	$f(\succsim_1, \succsim_2)$	$f(\succsim_1, \succsim_2')$
$\succsim_1'$	$f(\succsim_1', \succsim_2)$	$f(\succsim_1', \succsim_2')$

$(\succsim_1', \succsim_2')$	$\succsim_2$	$\succsim_2'$
$\succsim_1$	$f(\succsim_1, \succsim_2)$	$f(\succsim_1, \succsim_2')$
$\succsim_1'$	$f(\succsim_1', \succsim_2)$	$f(\succsim_1', \succsim_2')$

Incentive Compatibility of Social Choice Functions

$(\succsim_1, \succsim_2)$	$\succsim_2$	$\succsim_2'$
$\succsim_1$	$f(\succsim_1, \succsim_2)$	$f(\succsim_1, \succsim_2')$
$\succsim_1'$	$f(\succsim_1', \succsim_2)$	$f(\succsim_1', \succsim_2')$

$(\succsim_1, \succsim_2')$	$\succsim_2$	$\succsim_2'$
$\succsim_1$	$f(\succsim_1, \succsim_2)$	$f(\succsim_1, \succsim_2')$
$\succsim_1'$	$f(\succsim_1', \succsim_2)$	$f(\succsim_1', \succsim_2')$

$(\succsim_1', \succsim_2)$	$\succsim_2$	$\succsim_2'$
$\succsim_1$	$f(\succsim_1, \succsim_2)$	$f(\succsim_1, \succsim_2')$
$\succsim_1'$	$f(\succsim_1', \succsim_2)$	$f(\succsim_1', \succsim_2')$

$(\succsim_1', \succsim_2')$	$\succsim_2$	$\succsim_2'$
$\succsim_1$	$f(\succsim_1, \succsim_2)$	$f(\succsim_1, \succsim_2')$
$\succsim_1'$	$f(\succsim_1', \succsim_2)$	$f(\succsim_1', \succsim_2')$

Incentive Compatibility of Social Choice Functions

$(\succsim_1, \succsim_2)$	$\succsim_2$	$\succsim_2'$
$\succsim_1$	$f(\succsim_1, \succsim_2)$	$f(\succsim_1, \succsim_2')$
$\succsim_1'$	$f(\succsim_1', \succsim_2)$	$f(\succsim_1', \succsim_2')$

$(\succsim_1, \succsim_2')$	$\succsim_2$	$\succsim_2'$
$\succsim_1$	$f(\succsim_1, \succsim_2)$	$f(\succsim_1, \succsim_2')$
$\succsim_1'$	$f(\succsim_1', \succsim_2)$	$f(\succsim_1', \succsim_2')$

$(\succsim_1', \succsim_2)$	$\succsim_2$	$\succsim_2'$
$\succsim_1$	$f(\succsim_1, \succsim_2)$	$f(\succsim_1, \succsim_2')$
$\succsim_1'$	$f(\succsim_1', \succsim_2)$	$f(\succsim_1', \succsim_2')$

$(\succsim_1', \succsim_2')$	$\succsim_2$	$\succsim_2'$
$\succsim_1$	$f(\succsim_1, \succsim_2)$	$f(\succsim_1, \succsim_2')$
$\succsim_1'$	$f(\succsim_1', \succsim_2)$	$f(\succsim_1', \succsim_2')$

Incentive Compatibility of Social Choice Functions

Definition: A social choice function f is *dominant strategy incentive compatible* w.r.t. a set A of type profiles, if for all $\succsim^* \in A$, all $\succsim, \succsim' \in A$ and all $i \in N$:

$$f(\succsim_1, \ldots, \succsim_i^*, \ldots, \succsim_n) \succcurlyeq_i^* f(\succsim_1, \ldots, \succsim_i', \ldots, \succsim_n)$$

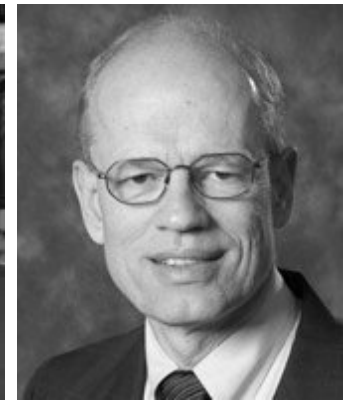
Intuition: *A social choice function f is truthfully implementable if for no player there are situations in which telling the truth can hurt.*

Remark: *Truthful implementation is important, for otherwise the mechanism does not guarantee that the goals of the designer, which are defined relative to the preferences of individuals, are reached.*

Gibbard-Satterthwaite Theorem

Definition: An scf f is *dictatorial* if there is a player i with $f(\succsim) \succsim_i o$ for all $\succsim$ and all o in O.

Intuition: For every preference profile $\succsim$, the outcome of the social choice function, i.e. $f(\succsim)$, is among the dictator's most preferred outcomes.

Gibbard-Satterthwaite Theorem

Definition: An scf f is *dictatorial* if there is a player i with $f(\succsim) \succsim_i o$ for all $\succsim$ and all o in O.

Intuition: For every preference profile $\succsim$, the outcome of the social choice function, i.e. $f(\succsim)$, is among the dictator's most preferred outcomes.

Theorem *(Gibbard, 1973 and Satterthwaite, 1975)***:** *If $|O| = 3$, every incentive compatible social choice function onto O is dictatorial.*

Intuition: For all non-trivial (i.e., non-dictatorial) social choice functions there are circumstances (i.e., type profiles) in which it is profitable for some player to lie about his true preferences.

Implementation and Solution Concepts

Let S be a solution concept and $\phi\colon A \to 2^O$ a social choice correspondence.

- A mechanism is a game form $G = (N, A, O, g)$
- For each $\succsim \in A$, $(G, \succsim)$ is a game
- For each $\succsim \in A$, S selects a subset of strategy profiles of the game $(G, \succsim)$
- Via g, for each $\succsim \in A$, S selects a subset of *outcomes* in O
- For each $\succsim \in A$, ϕ also selects a subset of *outcomes* in O

Implementation and Solution Concepts

Let S be a solution concept and $\phi\colon A \to 2^O$ a social choice correspondence.

- A mechanism is a game form $G = (N, A, O, g)$
- For each $\succsim \in A$, $(G, \succsim)$ is a game
- For each $\succsim \in A$, S selects a subset of strategy profiles of the game $(G, \succsim)$
- Via g, for each $\succsim \in A$, S selects a subset of *outcomes* in O
- For each $\succsim \in A$, ϕ also selects a subset of *outcomes* in O

Definition An scc $\phi\colon A \to 2^O$ is *S-implementable* if there is some *game form* $G = (N, A, O, g)$ such that for all preference profiles $\succsim$:

$$\phi(\succsim) \;=\; \{g(s) \in A : s \in S(G, \succsim)\}$$

Intuition: The *game form G implements ϕ in S* if for each $\succsim$ the sets of outcomes selected by S and ϕ coincide.

Implementation and Solution Concepts

$(\succsim_1, \succsim_2)$	L	R
T	a	b
B	c	d

$(\succsim_1, \succsim_2')$	L	R
T	a	b
B	c	d

$(\succsim_1', \succsim_2)$	L	R
T	a	b
B	c	d

$(\succsim_1', \succsim_2')$	L	R
T	a	b
B	c	d

	$\phi(\succsim)$	$S(G(\succsim))$
$(\succsim_1, \succsim_2)$	a	a
$(\succsim_1, \succsim_2')$	c	c
$(\succsim_1', \succsim_2)$	d	d
$(\succsim_1', \succsim_2')$	c	c

Implementation in Dominant Strategies

Definition: A strategy profile s^* is a *dominant strategy equilibrium* in a game G whenever for all $i \in N$ and all $s, s' \in S$:

$$g(s_1, \ldots, s_i^*, \ldots, s_n) \succsim_i g(s_1, \ldots, s_i', \ldots, s_n)$$

- Very robust in that it assumes very little about the agents
- Dominant strategy equilibrium does not allow for much flexibility

Dominant Strategy Implementation with two Alternatives

	okay	*veto*
okay	*a*	*b*
veto	*b*	*b*

$\succsim^a$	$\succsim^b$
a	*b*
b	*a*

$$f^{veto}(\succsim_1, \succsim_2) = \begin{cases} a & \text{if } \succsim_1 = \succsim^a \text{ and } \succsim_2 = \succsim^a \\ b & \text{otherwise} \end{cases}$$

Dominant Strategy Implementation with two Alternatives

$(\succsim^a, \succsim^a)$	*okay*	*veto*
okay	*a*	*b*
veto	*b*	*b*

$(\succsim^a, \succsim^b)$	*okay*	*veto*
okay	*a*	*b*
veto	*b*	*b*

$\succsim^a$	$\succsim^b$
a	*b*
b	*a*

$(\succsim^b, \succsim^a)$	*okay*	*veto*
okay	*a*	*b*
veto	*b*	*b*

$(\succsim^b, \succsim^b)$	*okay*	*veto*
okay	*a*	*b*
veto	*b*	*b*

The Revelation Principle

Lemma (*Revelation Principle*): If there exists any mechanism that implements an scf f in dominant strategies then there exists a direct mechanism that implements f in dominant strategies, i.e., an scf f is implementable in dominant strategies if and only if f is incentive compatible.

Intuition: For implementation in dominant strategy equilibria we can restrict ourselves to direct mechanisms.