Weak Strategyproofness in Randomized Social Choice

Felix Brandt and Patrick Lederer

AAAI' 25

Presented by David Kühnemann and Paul Weston

Outline

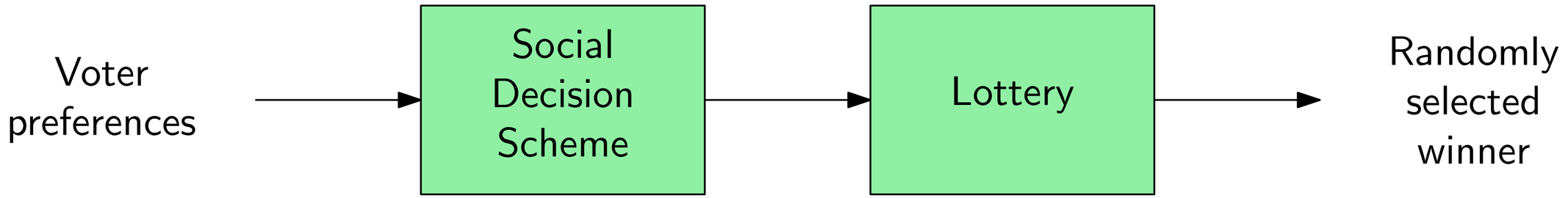
- What are Social Decision Schemes (SDSs)?
- (Weak) strategyproofness in the context of SDSs.
- Weakly strategyproof Social Decision Schemes
- Impossibility results
- Conclusion

Social Decision Schemes

Recall: For n voters and alternatives A, a resolute *Social Choice Function* (SCF) $F\colon \mathcal{L}(A)^n \to A$ picks an alternative given a profile of linear preference orders.

In this talk: A *lottery* is a probability distribution over A. Let $\Delta(A)$ be the set of all lotteries. A *Social Decision Scheme* $f\colon \mathcal{L}(A)^n \to \Delta(A)$ outputs a lottery given a preference profile R.

We let $f(R, x)$ denote the probability that x wins the lottery under the SDS f given the profile R.



Examples of Social Decision Schemes

Suppose there are $n = 100$ voters and $m = 2$ alternatives $M = \{a, b\}$.

A SDS f might result in the following:

- a is guaranteed to win the lottery if $a \succ b$ for > 50 voters,
- b is guaranteed to win the lottery if $b \succ a$ for > 50 voters,
- a and b both have a 50% chance of winning if $a \succ b$ for exactly 50 voters.

This is an example of an *even-chance* lottery.

Another SDS might go as follows: If exactly x voters rank $a \succ b$, a has a $x\%$ chance in the lottery.

Now: Even if a has a 90% majority, b still wins the lottery 10% of the time.

Recall: Strategyproofness

A voter may report a ballot that differs from her true preference order.

Consider a resolute SCF $F : \mathcal{L}(A)^n \to A$.

We say F is *strategyproof* if for each voter i and all possible ballots R_{-i} of the other voters, i submitting her true preferences R_i is optimal *among all possible ballots*.

Suppose i ranks $a \succ b \succ c$ and the others submit ballots R_{-i} such that b wins under F.

Then there is no ballot i can submit to make a win under F and R_{-i}.

Gibbard-Satterthwaite Theorem: Let F be a surjective and strategyproof SCF for $m \geq 3$ alternatives. Then F is a dictatorship.

Recall: Strategyproofness

A voter may report a ballot that differs from her true preference order.

Consider a resolute SCF $F : \mathcal{L}(A)^n \to A$.

We say F is *strategyproof* if for each voter i and all possible ballots R_{-i} of the other voters, i submitting her true preferences R_i is optimal *among all possible ballots*.

Suppose i ranks $a \succ b \succ c$ and the others submit ballots R_{-i} such that b wins under F.

Then there is no ballot i can submit to make a win under F and R_{-i}.

Gibbard-Satterthwaite Theorem: Let F be a surjective and strategyproof SCF for $m \geq 3$ alternatives. Then F is a dictatorship.

Can Social Decision Schemes help?

How to compare different SDS outcomes?

Instead of a single winner, a SDS outputs a lottery over possible winners.

What does it mean to achieve a better outcome under a SDS?

Lottery P:

Candidate	a	b	c
Chance of winning	0	1	0

Lottery Q:

Candidate	a	b	c
Chance of winning	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$

Voter i ranks $a \succ b \succ c$, which does she prefer?

How to compare different SDS outcomes?

Instead of a single winner, a SDS outputs a lottery over possible winners.

What does it mean to achieve a better outcome under a SDS?

Lottery P:

Candidate	a	b	c
Chance of winning	0	1	0

Lottery Q:

Candidate	a	b	c
Chance of winning	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$

Assume a *consistent* utility function $u_i : A \to \mathbb{R}$ for each voter: $x \succ_i y$ implies $u_i(x) > u_i(y)$.

Compare the *expected utilities* under different lotteries: $\mathbb{E}[P]$ vs $\mathbb{E}[Q]$.

Possible utility functions of voter i:

Candidate	a	b	c
Candidate utilitiy	6	5	1

$$\Rightarrow \quad \mathbb{E}[P] = 5 > 4 = \tfrac{6+5+1}{3} = \mathbb{E}[Q]$$

Candidate	a	b	c
Candidate utilitiy	6	2	1

$$\Rightarrow \quad \mathbb{E}[P] = 2 < 3 = \tfrac{6+2+1}{3} = \mathbb{E}[Q]$$

Takeaway: Two lotteries may be incomparable!

Strong strategyproofness for SDSs

A SDS is *strongly strategyproof* if for each voter i and every profile R including i's true preference, there exists no ballot R_i' s.t. $\mathbb{E}[f(R_i', R_{-i})] > \mathbb{E}[f(R)]$ for *some* consistent utility function.

A SDS is *ex post efficient* if $x \succ_i y$ for all i implies $f(R, y) = 0$.

Theorem (Gibbard '77): Let f be a strategyproof and ex post efficient SDS. Then f is a *random dictatorship*, i.e. f adopts the preferences of each voter with some fixed probability.

Weak strategyproofness for SDSs

A SDS is *strongly strategyproof* if for each voter i and every profile R including i's true preference, there exists no ballot R_i' s.t. $\mathbb{E}[f(R_i', R_{-i})] > \mathbb{E}[f(R)]$ for *some* consistent utility function.

A SDS is *weakly strategyproof* if for each voter i and every profile R including i's true preference, there exists no ballot R_i' s.t.

- $\mathbb{E}[f(R_i', R_{-i})] \geq \mathbb{E}[f(R)]$ for *every* consistent utility function, and
- $\mathbb{E}[f(R_i', R_{-i})] > \mathbb{E}[f(R)]$ for *some* consistent utility function.

Nuance: There may exist a profile where for every consistent utility function, there exists a ballot that i can deviate to to increase her expected utility. Weak strategyproofness merely guarantees that no *single* ballot achieves this for *every* consistent utility function.

Score-based SDSs

A *score function* $s\colon \mathcal{L}^n \times A \to \mathbb{R}_{\geq 0}$ assigns each candidate in each profile a score such that for all profiles R and R' that differ only in changing i's preference from $y \succ_i z$ to $z \succ_i y$:

- *localizedness:* $s(R, x) = s(R', x)$ for $x \notin \{y, z\}$,
- *monotonicity:* $s(R, z) \leq s(R', z)$,
- *balancedness:* $s(R, z) = s(R', z)$ implies $s(R, y) = s(R', y)$, and
- *positivity:* $\sum_{x \in A} s(R, x) > 0$.

Every score function s induces a *score-based* SDS f where $f(R, x) = \frac{s(R,x)}{\sum_y s(R,y)}$.

Example: The *plurality* score function $s_P(R, x)$ counts how many voters rank x first.
If a is the first choice of 90% of voters, she has a 90% chance of winning the lottery.

Aside: We can even allow one alternative x to receive score $s(R, x) = \infty$.

Score-based SDSs are weakly strategyproof

Theorem (Brandt-Lederer '25): Every score-based SDS is weakly strategyproof.

Proof: Let R, R' be profiles that only differ in the ballot of i.
Assume all scores are > 0 and that in R, voter i reports preferences $x_1 \succ x_2 \succ \cdots \succ x_m$.
We distinguish three cases by comparing the score totals $T := \sum_x s(R, x)$ and $T' := \sum_x s(R', x)$:

Goal: $\mathbb{E}[f(R)] > \mathbb{E}[f(R')]$ for *some* utility function u_i consistent with i's preferences

Score-based SDSs are weakly strategyproof

Theorem (Brandt-Lederer '25): Every score-based SDS is weakly strategyproof.

Proof: Let R, R' be profiles that only differ in the ballot of i.
Assume all scores are > 0 and that in R, voter i reports preferences $x_1 \succ x_2 \succ \cdots \succ x_m$.
We distinguish three cases by comparing the score totals $T := \sum_x s(R, x)$ and $T' := \sum_x s(R', x)$:

Case (i): $T < T'$. There exists a finite sequence of pairwise swaps in the preferences of i that transforms R to R' such that x_1 never moves "up" in the ranking.

$$\begin{aligned} R\colon\ & x_1 \succ x_2 \succ x_3 \succ x_4 \\ & x_1 \succ x_3 \succ x_2 \succ x_4 \\ & x_3 \succ x_1 \succ x_2 \succ x_4 \\ R'\colon\ & x_3 \succ x_1 \succ x_4 \succ x_2 \end{aligned}$$

$$\Rightarrow s(R, x_1) \geq s(R', x_1) \quad \Rightarrow \quad f(R, x_1) = \frac{s(R, x_1)}{T} > \frac{s(R', x_1)}{T'} = f(R', x_1)$$

Reminder: Switching $y \succ z$ to $z \succ y$:
Localizedness: Score of $x \notin \{y, z\}$ unaffected.
Monotonicity: Score of y cannot increase.

Goal: $\mathbb{E}[f(R)] > \mathbb{E}[f(R')]$ for *some* utility function u_i consistent with i's preferences

Score-based SDSs are weakly strategyproof

Theorem (Brandt-Lederer '25): Every score-based SDS is weakly strategyproof.

Proof: Let R, R' be profiles that only differ in the ballot of i.
Assume all scores are > 0 and that in R, voter i reports preferences $x_1 \succ x_2 \succ \cdots \succ x_m$.
We distinguish three cases by comparing the score totals $T := \sum_x s(R, x)$ and $T' := \sum_x s(R', x)$:

Case (i): $T < T'$. There exists a finite sequence of pairwise swaps in the preferences of i that transforms R to R' such that x_1 never moves "up" in the ranking.

$$\begin{aligned} R\colon\ & x_1 \succ x_2 \succ x_3 \succ x_4 \\ & x_1 \succ x_3 \succ x_2 \succ x_4 \\ & x_3 \succ x_1 \succ x_2 \succ x_4 \\ R'\colon\ & x_3 \succ x_1 \succ x_4 \succ x_2 \end{aligned}$$

$$\Rightarrow s(R, x_1) \geq s(R', x_1) \quad \Rightarrow \quad f(R, x_1) = \frac{s(R, x_1)}{T} > \frac{s(R', x_1)}{T'} = f(R', x_1)$$

Reminder: Switching $y \succ z$ to $z \succ y$:
Localizedness: Score of $x \notin \{y, z\}$ unaffected.
Monotonicity: Score of y cannot increase.

Goal: $\mathbb{E}[f(R)] > \mathbb{E}[f(R')]$ for any u_i that assigns a large enough utility to x_1. ✓

Score-based SDSs are weakly strategyproof

Theorem (Brandt-Lederer '25): Every score-based SDS is weakly strategyproof.

Proof: Let R, R' be profiles that only differ in the ballot of i.
Assume all scores are > 0 and that in R, voter i reports preferences $x_1 \succ x_2 \succ \cdots \succ x_m$.
We distinguish three cases by comparing the score totals $T := \sum_x s(R, x)$ and $T' := \sum_x s(R', x)$:

Case (ii): $T > T'$. There exists a finite sequence of pairwise swaps in the preferences of i that transforms R to R' such that x_m never moves "down" in the ranking.

Reminder: Switching $y \succ z$ to $z \succ y$:
Localizedness: Score of $x \notin \{y, z\}$ unaffected
Monotonicity: Score of z cannot decrease

$$\Rightarrow s(R, x_m) \leq s(R', x_m) \quad \Rightarrow \quad f(R, x_m) = \frac{s(R, x_m)}{T} < \frac{s(R', x_1)}{T'} = f(R', x_m)$$

Goal: $\mathbb{E}[f(R)] > \mathbb{E}[f(R')]$ for *some* utility function u_i consistent with i's preferences

Score-based SDSs are weakly strategyproof

Theorem (Brandt-Lederer '25): Every score-based SDS is weakly strategyproof.

Proof: Let R, R' be profiles that only differ in the ballot of i.
Assume all scores are > 0 and that in R, voter i reports preferences $x_1 \succ x_2 \succ \cdots \succ x_m$.
We distinguish three cases by comparing the score totals $T := \sum_x s(R, x)$ and $T' := \sum_x s(R', x)$:

Case (ii): $T > T'$. There exists a finite sequence of pairwise swaps in the preferences of i that transforms R to R' such that x_m never moves "down" in the ranking.

Reminder: Switching $y \succ z$ to $z \succ y$:
Localizedness: Score of $x \notin \{y, z\}$ unaffected
Monotonicity: Score of z cannot decrease

$$\Rightarrow s(R, x_m) \leq s(R', x_m) \quad \Rightarrow \quad f(R, x_m) = \frac{s(R, x_m)}{T} < \frac{s(R', x_1)}{T'} = f(R', x_m)$$

Goal: $\mathbb{E}[f(R)] > \mathbb{E}[f(R')]$ for any u_i that assigns a small enough utility to x_m. ✓

Score-based SDSs are weakly strategyproof

Theorem (Brandt-Lederer '25): Every score-based SDS is weakly strategyproof.

Proof: Let R, R' be profiles that only differ in the ballot of i.
Assume all scores are > 0 and that in R, voter i reports preferences $x_1 \succ x_2 \succ \cdots \succ x_m$.
We distinguish three cases by comparing the score totals $T := \sum_x s(R, x)$ and $T' := \sum_x s(R', x)$:

Case (iii): $T = T'$. Pick the smallest h such that $s(R, x_h) \neq s(R', x_h)$. There exists a sequence of pairwise swaps from R to R' such that the scores of $x_{\ell<h}$ never change and that x_h only moves "up" when swapped with an $x_{\ell<h}$ and "down" when swapped with an $x_{\ell>h}$.

$$\Rightarrow s(R, x_h) > s(R', x_h)$$
$$\Rightarrow f(R, x_\ell) = f(R', x_\ell) \text{ for } \ell < h,$$
$$\text{and } f(R, x_h) > f(R', x_h)$$

Reminder: Switching $y \succ z$ to $z \succ y$:
Monotonicity: Score of y cannot increase.
Balancedness: If the score of y does not change, neither does the score of z.

Goal: $\mathbb{E}[f(R)] > \mathbb{E}[f(R')]$ for *some* utility function u_i consistent with i's preferences

Score-based SDSs are weakly strategyproof

Theorem (Brandt-Lederer '25): Every score-based SDS is weakly strategyproof.

Proof: Let R, R' be profiles that only differ in the ballot of i.
Assume all scores are > 0 and that in R, voter i reports preferences $x_1 \succ x_2 \succ \cdots \succ x_m$.
We distinguish three cases by comparing the score totals $T := \sum_x s(R, x)$ and $T' := \sum_x s(R', x)$:

Case (iii): $T = T'$. Pick the smallest h such that $s(R, x_h) \neq s(R', x_h)$. There exists a sequence of pairwise swaps from R to R' such that the scores of $x_{\ell<h}$ never change and that x_h only moves "up" when swapped with an $x_{\ell<h}$ and "down" when swapped with an $x_{\ell>h}$.

$$\Rightarrow s(R, x_h) > s(R', x_h)$$
$$\Rightarrow f(R, x_\ell) = f(R', x_\ell) \text{ for } \ell < h,$$
$$\text{and } f(R, x_h) > f(R', x_h)$$

Reminder: Switching $y \succ z$ to $z \succ y$:
Monotonicity: Score of y cannot increase.
Balancedness: If the score of y does not change, neither does the score of z.

Goal: $\mathbb{E}[f(R)] > \mathbb{E}[f(R')]$ for any u_i that assigns a large enough utility to $x_1, \ldots, x_h$. ✓

Impossibility Results

Strict Preferences

- **Theorem** For $m \geq 5$ and odd $n \geq 5$. No even-chance SDS on $\mathcal{L}^n$ satisfies weak strategyproofness, Condorcet-consistency, and ex post efficiency.

- **Theorem** For $m \geq 5$. No pairwise, neutral, and weakly strategyproof SDS on $\mathcal{L}$ satisfies ex post efficiency.

Weak Preferences

- **Theorem** For $n \geq 4$ and $m \geq 4$. No anonymous and neutral SDS on $\mathcal{R}^N$ satisfies ex ante efficiency and weak strategyproofness.

- **Theorem** Every ex post efficient and weakly strategyproof even-chance SDS on $\mathcal{R}^N$ is dictatorial or bidictatorial.

Impossibility Result

Theorem Assume that $m \geq 5$ and $n \geq 5$ is odd. No even-chance, Condorcet-consistent, and ex post efficient SDS satisfies weak strategyproofness.

Impossibility Result

Theorem Assume that $m \geq 5$ and $n \geq 5$ is odd. No even-chance, Condorcet-consistent, and ex post efficient SDS satisfies weak strategyproofness.

Axioms:

Even-chance: An SDS is even-chance if for every profile R, there exists some $X \subseteq A$, $f(R,x) = \frac{1}{|X|}$ for $x \in X$ and $f(R,x) = 0$ otherwise.

Condorcet-Consistency: $f(R,x) = 1$ whenever x is the Condorcet winner in R.

Ex post efficient: $f(R,x) = 0$ whenever x is the Pareto dominated by another alternative

Impossibility Result

Theorem Assume that $m \geq 5$ and $n \geq 5$ is odd. No even-chance, Condorcet-consistent, and ex post efficient SDS satisfies weak strategyproofness.

Proof Sketch

We will focus on the case when $m = n = 5$. Consider the following profiles R and $\hat{R}$,

$$R: \begin{array}{r} 1: b, e, d, c, a \\ 2: a, b, c, e, d \\ 3: e, d, c, a, b \\ 4: b, c, a, e, d \\ 5: e, d, a, b, c \end{array} \qquad \hat{R}: \begin{array}{r} 1: b, e, d, c, a \\ 2: a, b, c, e, d \\ 3: d, a, e, b, c \\ 4: b, c, a, e, d \\ 5: e, d, a, b, c \end{array}$$

Claim 1: $f(R) = \{a, b, c, e\}$

Claim 2: $f(\hat{R}) = \{a, b, d, e\}$

Since player 3 ranks d above c, player 3 can manipulate by deviating from R to $\hat{R}$, contradicting weak strategyproofness.

Impossibility Result

Theorem Assume that $m \geq 5$ and $n \geq 5$ is odd. No even-chance, Condorcet-consistent, and ex post efficient SDS satisfies weak strategyproofness.

Claim 1: $f(R) = \{a, b, c, e\}$
We will argue by the size of $f(R)$,

- $|f(R)| \neq 1$: Follows from Condorcet-consistency, weak strategyproofness
  - No Condorcet winner $\implies$ No x s.t. $f(R, x) = 1$

- $|f(R)| \neq 2$: Follows from n odd, all assumed axioms
  - $f(R) = \{x, y\}$ iff the number of voters that rank x above y is the same as those who rank y above x

- $|f(R)| \neq 3$: Follows from cases

- $|f(R)| \neq 5$: Follows from ex post efficiency

$$R : \begin{array}{r} 1 : b, e, d, c, a \\ 2 : a, b, c, e, d \\ 3 : e, d, c, a, b \\ 4 : b, c, a, e, d \\ 5 : e, d, a, b, c \end{array}$$

Impossibility Result

Theorem Assume that $m \geq 5$ and $n \geq 5$ is odd. No even-chance, Condorcet-consistent, and ex post efficient SDS satisfies weak strategyproofness.

Goal: Show $|f(R)| \neq 3$. Proof by cases.
Consider $f(R) \neq \{b, c, e\}$.

Suppose, for sake of contradiction, that $f(R) = \{b, c, e\}$.
Consider the profile R^2 to the right.

Since b is the Condorcet winner, by Condorcet-consistency $f(R^2) = \{b\}$.

However, since player 2 ranks b above both c and e, player 2 is incentivized to deviate to R^2.

Contradiction with weak strategyproofness!
Therefore, $f(R) \neq \{b, c, e\}$.

$$R: \begin{array}{r} 1: b, e, d, c, a \\ 2: a, b, c, e, d \\ 3: e, d, c, a, b \\ 4: b, c, a, e, d \\ 5: e, d, a, b, c \end{array}$$

$$R^2: \begin{array}{r} 1: b, e, d, c, a \\ 2: b, a, c, e, d \\ 3: e, d, c, a, b \\ 4: b, c, a, e, d \\ 5: e, d, a, b, c \end{array}$$

Impossibility Result

Theorem Assume that $m \geq 5$ and $n \geq 5$ is odd. No even-chance, Condorcet-consistent, and ex post efficient SDS satisfies weak strategyproofness.

Claim 1: $f(R) = \{a, b, c, e\}$
We will argue by the size of $f(R)$,

- $|f(R)| \neq 1$: Follows from Condorcet-consistency, weak strategyproofness
  - No Condorcet winner $\implies$ No x s.t. $f(R, x) = 1$
- $|f(R)| \neq 2$: Follows from n odd,all assumed axioms
  - $f(R) = \{x, y\}$ iff the number of voters that rank x above y is the same as those who rank y above x
- $|f(R)| \neq 3$: Follows from cases
- $|f(R)| \neq 5$: Follows from ex post efficiency
  - e Pareto dominates d, by ex post efficiency $d \notin f(R)$

$$R : \begin{matrix} 1 : b, e, d, c, a \\ 2 : a, b, c, e, d \\ 3 : e, d, c, a, b \\ 4 : b, c, a, e, d \\ 5 : e, d, a, b, c \end{matrix}$$

Therefore, $f(R) = \{a, b, c, e\}$!

Impossibility Result

Theorem Assume that $m \geq 5$ and $n \geq 5$ is odd. No even-chance, Condorcet-consistent, and ex post efficient SDS satisfies weak strategyproofness.

Proof Sketch

We will focus on the case when $m = n = 5$. Consider the following profiles R and $\hat{R}$,

$$R: \begin{array}{r} 1: b, e, d, c, a \\ 2: a, b, c, e, d \\ 3: e, d, c, a, b \\ 4: b, c, a, e, d \\ 5: e, d, a, b, c \end{array} \qquad \hat{R}: \begin{array}{r} 1: b, e, d, c, a \\ 2: a, b, c, e, d \\ 3: d, a, e, b, c \\ 4: b, c, a, e, d \\ 5: e, d, a, b, c \end{array}$$

Claim 1: $f(R) = \{a, b, c, e\}$

Claim 2: $f(\hat{R}) = \{a, b, d, e\}$

Since player 3 ranks d above c, player 3 can manipulate by deviating from R to $\hat{R}$, contradicting weak strategyproofness.

Other Negative Results

Strict Preferences

- **Theorem** For $m \geq 5$ and odd $n \geq 5$. No even-chance SDS on $\mathcal{L}^n$ satisfies weak strategyproofness, Condorcet-consistency, and ex post efficiency.

- **Theorem** For $m \geq 5$. No pairwise, neutral, and weakly strategyproof SDS on $\mathcal{L}$ satisfies ex post efficiency.

Weak Preferences

- **Theorem** For $n \geq 4$ and $m \geq 4$. No anonymous and neutral SDS on $\mathcal{R}^N$ satisfies ex ante efficiency and weak strategyproofness.

- **Theorem** Every ex post efficient and weakly strategyproof even-chance SDS on $\mathcal{R}^N$ is dictatorial or bidictatorial.

Conclusion

- Social Decision Schemes generalize social choice functions to probabilistic outcomes.
- Lotteries can be incomparable and depend on a voter's utility function
- Weak notion of strategyproofness which turns out to be quite subtle.
- Class of *score-based* SDSs satisfies weak strategyproofness.
- As with SCFs, combining strategyproof SDSs with a few more reasonable axioms turns out to be impossible.