Theories of justice

- Section 6.4 in JR.
- The choice of one swf over another is ultimately a choice between alternative sets of ethical values.
- Two main approaches:
  - Harsanyi's approach,
  - Rawls' approach,
  - Unification
- All approaches assume that individuals do not yet know which position and circumstance they will occupy, i.e., both use the "veil of ignorance."
- However, these approaches differ in how individuals assign probabilities to each of the possible positions they could occupy.

Theories of justice

- **Harsanyi's approach:**
- He claimed that individuals assign an *equal probability* to the prospect of being in any possible position.
  - This is often referred to as the "principle of insufficient reason".
  - If there are N people in society, there is a probability $\frac{1}{N}$ that he will end up in the circumstances of person j, yielding a utility $u^j(x)$.
  - Hence, every individual's expected utility is

$$\sum_{i=1}^{I} \frac{1}{N} u^i(x)$$

Theories of justice

- **Harsanyi's approach:**
  - Therefore, when society chooses between two alternatives x and y, alternative x is socially preferred if

$$\sum_{i=1}^{I} \frac{1}{N} u^i(x) > \sum_{i=1}^{I} \frac{1}{N} u^i(y) \Longleftrightarrow \sum_{i=1}^{I} u^i(x) > \sum_{i=1}^{I} u^i(x)$$

    which exactly coincides with the utilitarian criterion.
  - Hence, the Harsanyi's approach is often used to support the utilitarian swf.

Theories of justice

- **Rawl's approach:**
  - In contrast, Rawls claimed that individuals have no empirical basis for assigning probabilities to each circumstance, whether equal or unequal probabilities.
  - That is, he viewed the original position as a setting of *complete ignorance*.
  - Assuming people are risk averse, he argues that in total ignorance individuals would order alternatives according to which one provides the highest utility in case he ended up as society's worst-off member.
  - Thus, x is socially preferred to y if and only if

$$\min\left\{u^1(x), ..., u^I(x)\right\} \geq \min\left\{u^1(y), ..., u^I(y)\right\}$$

  i.e., a purely maximin criterion.

Theories of justice

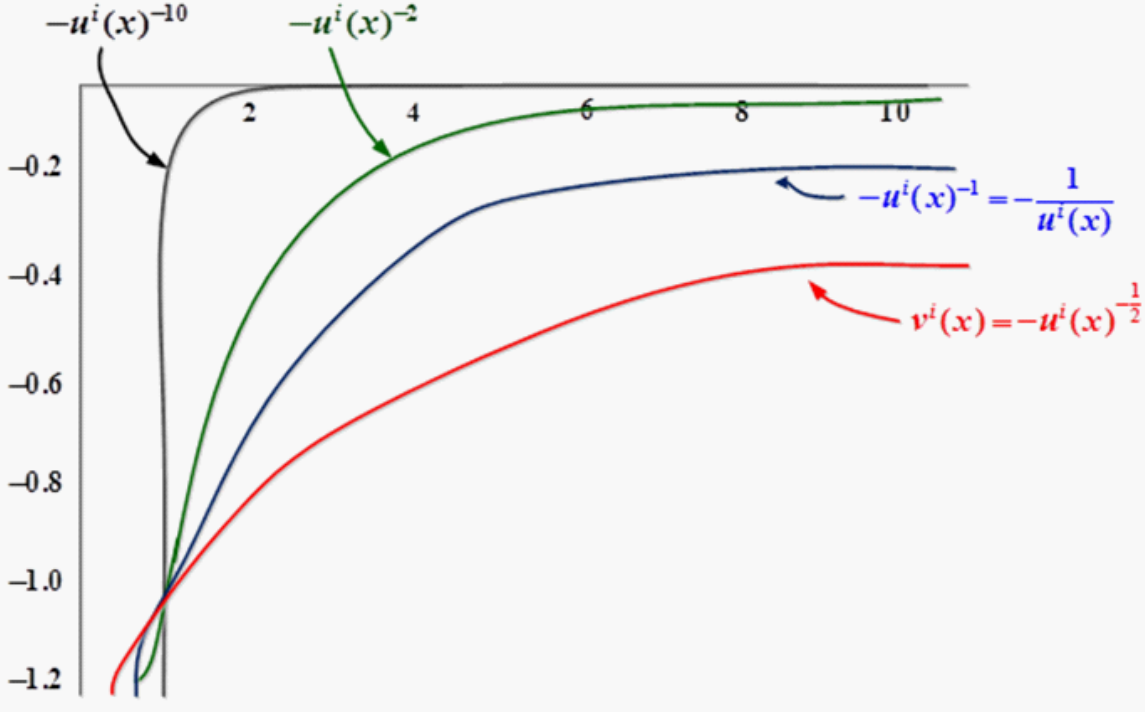
- **Unification of both approaches:**
  - Take an utility function $u^i(x)$.
  - The underlying preferences of this individual can also be represented by monotonic transformations of $u^i(x)$, such as

$$v^i(x) \equiv -u^i(x)^{-a}, \text{ where } a > 0$$

    i.e., a concave transformation of $u^i(x)$.
  - We can understand $v^i(x)$ as the vNM utility function of this individual, with parameter a capturing his degree of risk aversion.

Theories of justice

Monotonic transformation $v^i(x) = -u^i(x)^{-a}$, where a>0, evaluated at different values of parameter a.



- Note that $-u^i(x)^{-a}$ is only linear if a= -1, which is not allowed by definition i.e., a>0.

Theories of justice

- **Unification of both approaches:**
  - Using the Harsanyi's approach on this monotonic transformation (which captures risk aversion), yields a social welfare function

$$W = \sum_{i=1}^{I} v^i(x) \equiv -\sum_{i=1}^{I} -u^i(x)^{-a}$$

  - Importantly, the social ranking of alternatives provided by swf W must coincide with that of its monotonic tranformation W^* :

$$W^* = (-W)^{-\frac{1}{a}} \equiv \left(-\sum_{i=1}^{I} -u^i(x)^{-a}\right)^{-\frac{1}{a}} = \left(\sum_{i=1}^{I} -u^i(x)^{-a}\right)^{-\frac{1}{a}}$$

Theories of justice

- **Unification of both approaches:**
  - Hence,

$$W^* = \left(\sum_{i=1}^{I} -u^i(x)^{-a} \right)^{-\frac{1}{a}}$$

where, if parameter $-a = \rho$, then the swf takes the form

$$\left(\sum_{i=1}^{I} -u^i(x)^{\rho} \right)^{\frac{1}{\rho}}$$

i.e., that of a CES swf we described in previous sections.

Theories of justice

- **Unification of both approaches:**
- Therefore, when $\rho \longrightarrow -\infty$, the parameter of risk aversion $a \longrightarrow \infty$.
  - The above swf approaches the maximin criterion by Rawls as a limiting case.
  - Hence, the Rawlsian criterion is not incompatible with Harsanyi's utilitarian approach!
  - Instead, it becomes a special case of Harsanyi's approach when individuals become infinitely risk averse.
- When $-\infty < \rho < 1$, the parameter of risk aversion a satisfies $a \in [0, +\infty)$. (Note that we don't claim $a > -1$ since $a > 0$ by definition.)
  - Then, individuals are risk averse (but not infinitely), and
  - Social indifference curves are curvy.

Revelation of individual preferences

- Until this point, we analyzed whether individual preference relations could be aggregated into a social preference relation satisfying a set of desirable properties.
- However, we assumed individual preferences were truthfully reported by each individual.
- But, do individuals have incentives to do that?
- If they do, what are the consequences in terms of social preferences?
    - Reading reference: Section 6.5 in JR.

Revelation of individual preferences

- We will start this section defining what we mean by a social choice function.
  - A **social choice function** $c\left(\succsim^1, \succsim^2, ..., \succsim^I\right) \in X$ maps the profile of individual preferences $\left(\succsim^1, \succsim^2, ..., \succsim^I\right)$ into an alternative $x \in X$
    - That is, society uses the social choice function (scf) to "select" an alternative $x \in X$, using the information in the profile of individual preferences $\left(\succsim^1, \succsim^2, ..., \succsim^I\right)$.

Revelation of individual preferences

- Similarly as in previous contexts, we still seek to avoid aggregations of individual preferences that are dictatorial.
- Let's then define what we mean by a dictorial scf:
  - A scf $c(\cdot)$ is **dictatorial** if there is an individual h such that, if $x \succsim^h y$ for every two alternatives $x, y \in X$, then the scf selects x, i.e., $c\left(\succsim^1, \succsim^2, ..., \succsim^I\right) = x$.
    - That is, a scf is dictatorial if there is an individual h such that $c(\cdot)$ always chooses h's top choices.

Revelation of individual preferences

- What property needs to hold for individuals to not have incentives to misreport their individual preferences?
  - Consider an individual i with true preferences given by $\succsim^i$.
  - A scf $c(\cdot)$ is **strategy-proof** if every individual i prefers the alternative that the scf selects when he reports his true preferences, $c\left(\succsim^i, \succsim^{-i}\right) = x$, than that arising when he misreports them, $c\left(\succsim'^i, \succsim^{-i}\right) = y$, i.e., $x \succsim^i y$, where $\succsim^{-i}$ denotes the profile of individual preferences by all other individuals $\left(\succsim^1, ..., \succsim^{i-1}, \succsim^{i+1}, ..., \succsim^I\right)$.
    - Hence, if a scf is strategy proof, individuals have no strict incentives to misreport their preferences, regardless of the preferences other individuals report, $\succsim^{-i}$.
    - This is true even if the other individuals are misreporting their preferences.

Revelation of individual preferences

- Unfortunately, strategy-proofness has deep consequences:
  - The negative results in Arrow's impossibility theorem are hunting us again!
  - **Gibbard-Satterthwaite theorem.** If $\#X \geq 3$, then every strategy-proof scf is dictatorial.
  - Note that we cannot apply Arrow's results here, since his proof was based on the swf (showing it is dictatorial), while we now seek to show that the scf is dictatorial.

Revelation of individual preferences

- **Gibbard-Satterthwaite theorem.**
  - *Proof* :
  - What are we planning to do:

$$\text{Strategy-proof scf} \Longrightarrow \text{Dictatorial scf}$$

  - *1st part.* Show that a strategy-proof scf must exhibit two properties: Pareto efficiency and monotonicity.
  - *2nd part*. Every Pareto efficient and monotonic scf is dictatorial.

Revelation of individual preferences

- **Gibbard-Satterthwaite theorem.**
  - Let's start defining what we mean by Pareto efficient scf:
    - A scf is **Pareto efficient** when every individuals' strictly preference for x over y, $x \succ^i y$, where $x, y \in X$, yields the scf to select x, i.e., $c\left(\succsim^1, \succsim^2, ..., \succsim^I\right) = x$.

Revelation of individual preferences

- **Gibbard-Satterthwaite theorem.**
  - Let us now define a **monotonic scf**:
    - Consider a initial profile of individual preferences, $\left(\succsim^1, \succsim^2, ..., \succsim^I\right)$, yielding that alternative x is chosen by the scf, i.e., $c\left(\succsim^1, \succsim^2, ..., \succsim^I\right) = x$.
    - Assume that the preferences of at least individual i change from $x \succsim^i y$ to $x \succ'^i y$, for every $y \in X$, i.e., alternative x rises to the only spot at the top of his ranking of alternatives, and the preference for x is not lowered for any individual, i.e., $x \not\prec y$.
    - We then say that a scf is **monotonic** if the scf still selects x under the new profile of individual preferences, $c\left(\succ'^1, \succ'^2, ..., \succ'^I\right) = x$.
  - Loosely speaking, a scf is monotonic if it keeps selecting x as socially preferred when x becomes the top alternative for at least individual.

Revelation of individual preferences

- **Gibbard-Satterthwaite theorem.**
  - **Example of monotonic scf:**

Morning			*Afternoon*		
$\succsim^1$	$\succsim^2$	Social choice	$\succsim^1$	$\succsim^2$	Social choice
x a	x b	x	x	x b	x
b	c		a b	c	
c	a		c	a	
.	.		.	.	
.	.		.	.	

  - In the morning, individual 1 was indifferent between x and a, while individual 2 was indifferent between x and b.
  - In the afternoon, the preference of individual 2 for x has not changed, but that of individual 1 become more intense (i.e., x is not his only top alternative).
  - Monotonicity says that x should still be selected by the social choice function.

Revelation of individual preferences

- **Gibbard-Satterthwaite theorem.**
- First part:
  - In this part, we NTS that strategy-proofness implies Pareto efficiency and monotonicity on the scf.
  - **Monotonicity:**
  - Consider an arbitrary profile of individual preferences, $\left(\succsim^i, \succsim^{-i}\right)$, where $c\left(\succsim^i, \succsim^{-i}\right) = x$.
  - Consider an arbitrary indidividual i, whose preferences change from $x \succsim^i y$ to $x \succ'^i y$ for every $y \in X$.
  - We then NTS that the scf still selects x, i.e., $c\left(\succsim'^i, \succsim^{-i}\right) = x$.

Revelation of individual preferences

- **Gibbard-Satterthwaite theorem.**
- First part:
  - **Monotonicity:**
  - By contradiction, suppose that the scf doesn't select x under the new preferences, i.e.,

$$c\left(\succsim'^{i}, \succsim^{-i}\right) = y \neq x$$

  - Then, the social choice that arises when individual i truthfully reports his preferences, y, is less preferred than the alternative that would emerge when he misreports his preferences, x, i.e., $x \succ'^{i} y$.
  - Therefore, under the new preference relation $\succsim'^{i}$, individual i has incentives to misreport his preferences, thus violating strategy-proofness.

Revelation of individual preferences

- **Gibbard-Satterthwaite theorem.**
- First part:
  - **Pareto efficiency:**
  - We will be using monotonicity (just proved) to show Pareto efficiency.
  - We NTS that in a profile of individual preference relations $\succsim$ where $x \succ^i y$ for every $y \neq x$ and for every individual i, the scf selects $c(\succsim) = x$.

Revelation of individual preferences

- **Gibbard-Satterthwaite theorem.**
- First part:
  - **Pareto efficiency:**
  - *Proof:*
  - Consider a profile of individual preferences $\succsim'$ which yields $c(\succsim') = x$, where $x \succ'^{i} y$ does not necessarily hold for all individuals.
  - Construct now a new profile of individual preferences $\succsim''$ where $x \succ''^{i} y$ holds for one individual i,
    - i.e., alternative x has been moved to the top of i's ranking, but leaving the remaining ranking unaffected.
  - By monotonicity, $c(\succsim'') = x$.
  - We can now repeat the process for all individuals (moving x to the top of their rankings).
  - Applying monotonicity again yields $c(\succsim) = x$, as required.

Revelation of individual preferences

- **Gibbard-Satterthwaite theorem.**
- Second part:
  - After demonstrating that strategy-proofness implies monotonicity and Pareto efficiency, we are ready to show that:
  - $\#X \geq 3 + Monotonicity + ParetoEfficiency \Longrightarrow Dictatorship$
  - We will demonstrate that using five steps.

Revelation of individual preferences

- **Gibbard-Satterthwaite theorem.**
- Second part:
  - *Step 1.*
  - Consider a profile of strict rankings in which alternative x is ranked highest and y lowest for every individual i.
  - Pareto efficiency implies that the scf must select x.

Revelation of individual preferences

- **Gibbard-Satterthwaite theorem.**
- Step 1

$\succsim^1$	...	$\succsim^{n-1}$	$\succsim^n$	$\succsim^{n+1}$	...	$\succsim^l$	Social choice
x	...	x	x	x	...	x	x
.		.	.	.		.	
.		.	.	.		.	
.		.	.	.		.	
y	...	y	y	y	...	y	

Revelation of individual preferences

- **Gibbard-Satterthwaite theorem.**
- Second part:
  - *Step 1.*
  - Consider now that we change individual 1's ranking by raising y in it one position at a time.
  - By monotonicity, the social choice must remain to be x as long as $x \succ^1 y$.
  - But when y is raised above x, the social choice can change to y, or remain at x.
  - If the social choice is still x, we then begin raising y for individual 2 one position at a time.
  - Eventually, the social ranking will change when y is raised above x in individual n's ranking.

Revelation of individual preferences

- **Gibbard-Satterthwaite theorem.**
- Second part:
  - *Step 2.*
  - Consider now a different profile of individual preferences in which:
    - x is moved to the bottom of individual i's ranking, for all $i < n$, and
    - x is moved to the second last position in individual i's ranking, for all $i > n$.
  - See figures.
  - We wish to show that this change in individual preferences does not change the selection of the scf.

Revelation of individual preferences

- **Gibbard-Satterthwaite theorem.**
- Figure 6.11: After raising y to the top position for $n-1$ individuals.

$\succsim^1$	...	$\succsim^{n-1}$	$\succsim^n$	$\succsim^{n+1}$	...	$\succsim^l$	Social choice
y	...	y	x	x	...	x	x
x	...	x	y	.		.	
.		.	.	.		.	
.		.	.	.		.	
.		.	.	y	...	y	

Revelation of individual preferences

- **Gibbard-Satterthwaite theorem.**
- Figure 6.12: After raising y to the top position for n individuals (which changes the social choice).

$\succsim^1$	...	$\succsim^{n-1}$	$\succsim^n$	$\succsim^{n+1}$	...	$\succsim^I$	Social choice
y	...	y	y	x	...	x	y
x	...	x	x	.		.	
.		.	.	.		.	
.		.	.	.		.	
.		.	.	y	...	y	

Revelation of individual preferences

- **Gibbard-Satterthwaite theorem.**
- Figure 6.13.

$\succsim^1$	...	$\succsim^{n-1}$	$\succsim^n$	$\succsim^{n+1}$	...	$\succsim^l$	Social choice
y	...	y	x	.	...	.	x
.	...	.	y	.		.	
.		.	.	.		.	
.		.	.	x		x	
x	...	x	.	y	...	y	

Revelation of individual preferences

- **Gibbard-Satterthwaite theorem.**
- Figure 6.14.

$\succsim^1$	...	$\succsim^{n-1}$	$\succsim^n$	$\succsim^{n+1}$	...	$\succsim^l$	Social choice
y	...	y	y	.	...	.	y
.	...	.	x	.		.	
.		.	.	.		.	
.		.	.	x		x	
x	...	x	.	y	...	y	

Revelation of individual preferences

- **Gibbard-Satterthwaite theorem.**
- Second part:
  - *Step 2.*
  - Comparing figure 6.14 to 6.12, note that, by monotonicity, the scf must still select y:
    - The social choice in figure 6.12 was y, and no individual's ranking of y versus any other alternative has changed when we moved from figure 6.12 to 6.14.

Revelation of individual preferences

- **Gibbard-Satterthwaite theorem.**
- Second part:
  - *Step 2.*
  - Let us now compare figures 6.13 and 6.14:
    - They only differ in the ranking of individual n: he prefers x to y in figure 6.13, but prefers y to x in figure 6.14.
    - Hence, if the scf selects y in figure 6.14, it must still select either y or x in figure 6.13 (by the same logic as in step 1).
    - But, can the scf select y in figure 6.13?
    - No! By monotonicity, scf should then select y in figure 6.11 as well, which is a contradiction.
    - Therefore, the social choice in figure 6.13 must be x.

Revelation of individual preferences

- **Gibbard-Satterthwaite theorem.**
- Second part:
  - *Step 3.*
  - We need to make sure we are using the assumption of $\#X \geq 3$.
  - For that, we only need to consider an alternative $z \neq x, y$.
  - See figure 6.15.
  - Note that the social choice is still x:
    - Indeed, we have not changed the ranking of x against any other alternative in any individual's ranking.

Revelation of individual preferences

- **Gibbard-Satterthwaite theorem.**
- Figure 6.15.

$\succsim^1$	...	$\succsim^{n-1}$	$\succsim^n$	$\succsim^{n+1}$	...	$\succsim^I$	Social choice
.		.	x	.	...	.	x
.		.	z	.		.	
.		.	y	.		.	
z	...	z		z	...	z	
y	...	y	.	x	...	x	
x	...	x	.	y	...	y	

Revelation of individual preferences

- **Gibbard-Satterthwaite theorem.**
- Second part:
  - *Step 4.*
  - Consider now a profile of individual preferences in which we take figure 6.15 and interchange the ranking of x and y for all individuals $i > n$.
  - This profile of individual preferences is depicted in figure 6.16.

Revelation of individual preferences

- **Gibbard-Satterthwaite theorem.**
- Figure 6.16.

$\succsim^1$	...	$\succsim^{n-1}$	$\succsim^n$	$\succsim^{n+1}$	...	$\succsim^l$	Social choice
.		.	x	.	...	.	x
.		.	z	.		.	
.		.	y	.		.	
z	...	z		z	...	z	
y	...	y	.	y	...	y	
x	...	x	.	x	...	x	

Revelation of individual preferences

- **Gibbard-Satterthwaite theorem.**
- Second part:
  - *Step 4.*
  - When moving from 6.15 to 6.16, we made alternative y preferred for $I - n$ individuals.
  - Hence, either x is still selected by the scf (as under figure 6.15), or y becomes selected.
  - But, can alternative y be selected?
  - No! By Pareto efficiency, $z \succ^i y$ for all i.
  - Thus, x must be socially selected.

Revelation of individual preferences

- **Gibbard-Satterthwaite theorem.**
- Second part:
  - *Step 5.*
  - Hence, individual n's top choice (any arbitrary x) becomes selected by the scf; see figure 6.16.

Revelation of individual preferences

- How to avoid the unfortunate result in the Gibbard-Satterthwaite theorem?
  - That is, under which conditions truthtelling becomes incentive compatible for all individuals and the scf is not dictatorial?
  - When utility functions are quasilinear and we design a mechanism such as the Vickrey-Clarks-Groves (VCG) mechanism.

Revelation of individual preferences

- The VCG mechanism is strategy proof, e.g., inducing truthful report of each individual's benefit of a public project.
- It is also non-dictatorial (no individual imposes his/her own preferences on the group).
- Isn't this result contradictory with Gibbard-Satterrthwaite theorem?
  - No. For the VCG mechanism to work, we need to restrict the set of admissible preferences to those representable with a quasilinear utility function.
  - We are, hence, giving up the U property.

Revelation of individual preferences

- We have already encountered the VCG mechanism:
  - In the chapter on public goods in EconS 501.
  - Otherwise, I recommend you to read section 9.5 in JR
    - This section is quite long, so if you don't have time, please read pp. 461-465.

Revelation of individual preferences

- More on mechanism design (as a summer reading):
  - Chapter 9 in JR (nice worked-out examples).
  - One of the last chapters in Tadelis.
  - Once you are done with the above two: chapter 23 in MWG.
  - Hungry for more?
    - Lecture notes by Tilman Borgers (Yale), and
    - Lecture notes by Matthew Jackson (Stanford).