Moral preferences
Moral preferences

Harsanyi and Rawls argue for theories of social justice (equivalently, fairness) based on the choices that agents would make for society in the original position, behind a veil of ignorance.

...without knowing their own social and economic positions, their own special interests in the society, or even their own personal talents and abilities (or their lack of them). — Harsanyi (1975) —

Harsanyi and Rawls come to quite different conclusions, not because they view the original position differently, but because they treat uncertainty quite differently (Rawls denies orthodox decision theory).
Harsanyi’s (1953, 1955) model for moral value judgments

Suppose an agent wants to make a moral value judgment about the relative merits of two alternative social systems.

... act in such a way as if he assigned the same probability to his occupying each social position under either system...

... then, he would clearly satisfy the impartiality and impersonality requirements to the fullest possible degree. – Harsanyi (1978) –

The agent has two different sets of preferences: personal preferences and moral preferences (preferences in the original position).
Two observations

[1] Both Harsanyi and Rawls insist that moral preferences must conform to certain rationality requirements, and hence must have a special form – as opposed to personal preferences, which merely reflect taste.

[2] Harsanyi and Rawls – and many other writers – view the original position as a purely hypothetical environment, and hence view moral preferences as a purely intellectual construct.
Our point of departure from the work of Harsanyi and Rawls – and the enormous literature they spawned – comes from two observations:

[1] Choice behavior/preferences *behind* the veil of ignorance can be decomposed into choice behavior/preferences *in front of* the veil of ignorance:

choices that involve only personal consumption under uncertainty and choices that involve social consumption – but no uncertainty.

[2] Choices behind the veil of ignorance *can* be presented – and choices in the other two environments as well – in a controlled *laboratory* setting.
The linkage between preferences behind and in front of the veil of ignorance provides new ways of interpreting the theory of justice:

not just as a *normative* theory, but also as a *descriptive* theory and even as a *prescriptive* theory.

This linkage means that moral preferences *cannot* occupy such a privileged position – modulo certain assumptions, they are completely determined by risk preferences and social preferences.
Template for analysis

- Consider choice behavior by a single agent in each of three environments.

- Each choice has consequences for *self* (the agent) and for an (unknown) *other*.

- We consider only environments that involve binary choices and equiprobable lotteries.

- The results extend to more general choices and lotteries, and to unknown probabilities as well.
Consider lotteries over outcomes \([a, b]\), where \(a\) is consumption for *self* and \(b\) is consumption for *other*.

For our purposes, it suffices to consider binary lotteries with equal probabilities:

\[
(0.5)[a, b] + (0.5)[c, d]
\]

where \(a, b, c, d \geq 0\). Write \(L\) for the space of all such lotteries, and identify \(L\) with the convex cone \(\mathbb{R}_+^4\).
Define closed convex subcones of $\mathcal{L}$:

$$\mathcal{R} = \{( .5)[a, 0] + ( .5)[c, 0]\},$$

$$\mathcal{S} = \{( .5)[a, b] + ( .5)[a, b]\},$$

$$\mathcal{M} = \{( .5)[a, b] + ( .5)[b, a]\}.$$

We can interpret choice in each of the environments as choice in one of the corresponding cones by making an obvious identification:
- **Risk:** identify $\mathbb{R}_+^2$ with $\mathcal{R}$ by
  \[ (x, y) \mapsto (.5)[x, 0] + (.5)[y, 0]. \]

- **Social:** identify $\mathbb{R}_+^2$ with $\mathcal{S}$ by
  \[ (x, y) \mapsto (.5)[x, y] + (.5)[x, y]. \]

- **Moral:** identify $\mathbb{R}_+^2$ with $\mathcal{M}$ by
  \[ (x, y) \mapsto (.5)[x, y] + (.5)[y, x], \]
  which coincides exactly with Harsanyi’s (1953, 1955) formalization of the original position.
Research questions

[1] What is the relationship between moral preferences and personal/social (altruistic) preferences?

[2] How can behavior behind [Harsanyi’s] veil of ignorance be characterized experimentally?

[3] Is behavior behind a veil of ignorance consistent with the utility maximization model?

[4] Can the underlying moral preferences be recovered from observed choices?
Assumptions

Given a preference relation $\succeq$ on $L$, write $\succeq_{\mathcal{R}}$, $\succeq_{\mathcal{S}}$, $\succeq_{\mathcal{M}}$ for its restrictions to $\mathcal{R}$, $\mathcal{S}$, $\mathcal{M}$, respectively.

$[i]$ $\succeq$ satisfies the usual requirements: completeness, transitivity, reflexivity, continuity, and the Sure Thing Principle.

$[ii]$ $\succeq$ satisfies (weak) independence:

$$[a, b] \succeq c \in S [a', b'] \text{ and } [c, d] \succeq s [c', d']$$

$$\Rightarrow \quad (.5)[a, b] + (.5)[c, d] \succeq (.5)[a', b'] + (.5)[c', d']$$

(not the usual independence axiom and does not have the usual consequences).
Next, we make two assumptions about social preferences:

[iii] **Worst outcome:** \([a, b] \succeq_S [0, 0]\) for every \([a, b] \in S\).

[iv] **Self-regarding:** for each outcome \([a, b]\) there is an outcome \([s, 0]\) such that \([s, 0] \succeq_S [a, b]\).

[i] and [ii] are rationality requirements (should not necessarily be given any philosophical interpretation).

[iii] and [iv] limit the extent to which the self is (respectively) spiteful or altruistic toward other; they seem very natural requirements but they are not entirely innocuous.
**Result:** Every preference relation $\succeq$ on $L$ that satisfies $[i]-[iv]$ is determined by its restrictions $\succeq_R$ and $\succeq_S$.

**Proof:** Fix an outcome $[x, y]$. Because $\succeq_S$ is self-regarding, there is some $s$ such that $[s, 0] \succeq_S [x, y]$.

Define the *selfish equivalent* of $[x, y]$ by

$$\sigma[x, y] = \inf \{ s : [s, 0] \succeq_S [x, y] \}.$$  

Continuity and worse outcome guarantee that $[\sigma[x, y], 0] \sim_S [x, y]$. and by construction,

$$[a, b] \sim_S [\sigma[a, b], 0] \text{ and } [c, d] \sim_S [\sigma[c, d], 0].$$
independence guarantees that

\[(.5)[a, b] + (.5)[c, d] \sim (.5)[\sigma[a, b], 0] + (.5)[\sigma[c, d], 0].\]

Hence

\[(.5)[a, b] + (.5)[c, d] \succeq (.5)[a', b'] + (.5)[c', d']\]

\[\iff\]

\[(.5)[\sigma[a, b], 0] + (.5)[\sigma[c, d], 0] \succeq \mathcal{R}(.5)[\sigma[a', b'], 0] + (.5)[\sigma[c', d'], 0]\]

which decomposes preferences over \(\mathcal{L}\) into preferences over \(\mathcal{S}\) (selfish equivalents) and preferences over \(\mathcal{R}\), as desired.
Given a linear budget constraint, we identify choice behavior in the Social Choice environment as

- **selfish** if the choice subject to every budget constraint is of the form $[y, 0]$ – giving nothing to *other*.

- **symmetric** if $(a, b)$ is chosen subject to $px + qy \leq w$ iff $(b, a)$ is chosen subject to the mirror-image budget constraint $qx + py \leq w$. 
Corollary I: If the preference relation $\succeq$ satisfies [i] and [ii] and choice behavior in the $S$ is selfish then choice behavior in $R$ coincides with choice behavior in $M$.

Proof: Monotonicity and continuity guarantee that purely selfish behavior implies that $[x, 0] \sim_S [x, y]$ for every $x, y$. Independence implies that

$$(.5)[y, 0] + (.5)[x, 0] \sim (x, y) + (.5)[y, x].$$

It follows immediately that $\succeq_R$ and $\succeq_M$ coincide from whence choices in the Risk and Veil of Ignorance environments coincide, as asserted.
Corollary II: If the preference relation $\succeq$ satisfies [i] and [ii] and choice behavior in $\mathcal{S}$ is symmetric, then choice behavior in $\mathcal{S}$ coincides with choice behavior in $\mathcal{M}$.

Proof: Suppose that $(a, b)$ is chosen from some budget set $B$ for the Social Choice environment, so that $(b, a)$ is chosen in the mirror image budget set $B'$.

Say that $(c, d)$ is chosen from the budget set $B$ for the Veil of Ignorance environment, and that $(c, d) \neq (a, b)$.
Because \((c, d) \in B\), it follows that
\[
(\cdot 5)[c, d] + (\cdot 5)[d, c] \succeq \mathcal{M} (\cdot 5)[a, b] + (\cdot 5)[b, a].
\]

Independence implies that
\[
[c, d] \succeq_S [a, b] \text{ or } [d, c] \succeq_S [b, a],
\]
which is inconsistent with the fact that \((a, b)\) (resp. \((b, a)\)) is chosen from the budget set \(B\) (resp. \(B'\)).

It follows that risk attitude is irrelevant in the Veil of Ignorance environment, as asserted.
Experimental analysis

- Subjects in the experiments were recruited from all classes at UCLA and Yale Law School.

- Each decision problem is presented as a choice from a two-dimensional budget line.

- A choice \((x, y)\) from the budget line represents an allocation between accounts \(x, y\) (corresponding to the horizontal and vertical axes).

- Choices are made through a simple point-and-click design using a graphical computer interface.
The actual payoffs of a particular choice in a particular environment/treatment are determined by the allocation to the $x$ and $y$ accounts:

- **Risk**: involves only pure risk; it is identical to the (symmetric) risk experiment of Choi, Fisman, Gale & Kariv (*AER*, 2007).

- **Social Choice**: involves only altruism; it is identical to the (linear) two-person dictator experiment of Fisman, Kariv & Markovits (*AER*, 2007).

- **Veil of Ignorance**: involves equiprobable binary lotteries over symmetric pairs of consumption for *self* and for *other*. 
The distributions of token shares aggregated across subjects

Social Choice

The tokens kept as a fraction of the sum of the tokens kept and given to other.
The distributions of token shares aggregated across subjects

Risk

The fraction of tokens allocated to the cheaper account.
The distributions of token shares aggregated across subjects

Veil of Ignorance

The fraction of tokens allocated to the cheaper account.
The distributions of token shares aggregated across subjects
UCLA

The distributions of token shares aggregated across subjects
Yale

Individual behavior

- The aggregate data tell us little about the choice behavior of individual subjects.

- Scatterplots of all choices of illustrative subjects – each entry plots $y/(x + y)$ as a function of $\log(p_x/p_y)$ in a particular treatment.

- There is no taxonomy that allows us to classify all subjects unambiguously.

- The characteristic of all our data is striking regularity within subjects and heterogeneity across subjects.
The relationship between the log-price ratio and the token share

$Y/(Y+X)$

$\ln(p)$

ID11

$Y/(Y+X)^{-1}$

$\ln(p)$

$Y/(Y+X)$

$\ln(p)$

\[ \frac{Y}{Y+X} \]

\[ \ln(p) \]

\[ X \text{ – Risk / } X \text{ – Social Choice / } X \text{ – Veil of Ignorance } \]
$\frac{Y}{Y+X}$

$\ln(p)$

ID80

$\frac{Y}{Y+X}$ - 1

$\text{Ln}(p)$

ID78

\[ \frac{Y}{Y+X} \]

\[ \text{Ln}(p) \]

\[ X \text{ – Risk / X – Social Choice / X – Veil of Ignorance} \]
$Y/(Y+X)$

$\ln(p)$

$Y/(Y+X)$

$\text{Ln}(p)$

ID3

$X - \text{Risk} / X - \text{Social Choice} / X - \text{Veil of Ignorance}$
ID143

Y/(Y+X)

Ln(p)

$Y/(Y+X)$

ID147

$\ln(p)$

$\frac{Y}{Y+X}$

$\ln(p)$

ID59

Testing the theory

- Many selfish subjects seem to display the same choice behaviors in the Risk and Veil of Ignorance environments, but a substantial number do not.

- Because of the nature of the data, “flexible” functional forms do not provide a plausible fit for the data.

- No satisfactory formulation to explain the “switching” between stylized behavior patterns exhibited by many subjects.

- Parametric approaches may be possible – keeping in mind that individual behaviors are extremely heterogeneous.
Non-parametric econometric approaches

Revealed preference

- The ratio of the CCEI score for the combined data set to the \textit{minimum} of the CCEI scores for the separate data sets.

- A measure of the extent to which choice behaviors in any two environments coincide.

- Unfortunately, this test is weak – cannot discriminate between Risk and Veil of Ignorance behavior of selfish and non-selfish subjects.
The distributions of CCEI scores for the combined data set

CCEI Risk and Veil of Ignorance

Fraction of subjects

Selfish  Non-selfish
Kolmogorov-Smirnov type tests

- A two-sample Kolmogorov-Smirnov tests of the equality of distributions of token and budget shares.

- The test is sensitive to differences in both location and shape of the empirical cumulative distribution functions of the two samples.

- Generalize the univariate Kolmogorov-Smirnov statistics for bivariate samples (Adler and Brown, 1986).
There are subjects who fail Corollary I (selfish but display different behaviors in the $\mathcal{R}$ and $\mathcal{M}$) and others who fail Corollary II.

These subjects might have preferences over $\mathcal{L}$ that do not obey independence (or might not be consistent with utility maximization).

Individual preferences are very heterogeneous, ranging from utilitarian to Rawlsian.

Actual preferences “mix-and-match” behavior in ways that no extant theory would regard as justified.
Takeaways

• A positive account of preferences for both personal and social consumption in rich choice environments.

• Two methodological contributions:
  – The establishment of theoretical links between preferences in various environments.
  – An experimental technique that allows for the collection of richer data about preferences.

• The experimental platform and analytical techniques are applicable to many other types of individual choice problems.