Introduction to Social Preferences⇌

- Social preferences can be defined as "ways that people's utility depend *directly* on the well-being, motives, and beliefs of others".⇌
- Among the most famous passages in economics is from Adam Smith's *The Wealth of Nations*:⇌
  - It is not from the benevolence of the butcher, the brewer, or the baker that we expect our dinner, but from their regard for their own interest. We address ourselves not to their humanity, but to their self-love, and never talk to them of our necessities, but of their advantage.

↬

Social Preferences

- Research recognizing the centrality of self-interest without ignoring social preferences, illustrated by Dawes and Thaler (1988):⇌
  - "In the rural areas around Ithaca it is common for farmers to put some fresh produce on the table by the road. There is a cash box on the table, and customers are expected to put money in the box in return for the vegetables they take.⇌ The box has just a small slit, so money can only be put in, not taken out.⇌ Also, the box is attached to the table, so no one can (easily) make off with the money.⇌ We think that the farmers have just about the right model of human nature.⇌ They feel that enough people will volunteer to pay for the fresh corn to make it worthwhile to put it out there.⇌ The farmers also know that if it were easy enough to take the money, someone would do so."

↬

Social Preferences

While moving now to the 'field', most evidence gathered from lab.⇌

- Legitimate concerns about extrapolability from one (campus *or* non-campus) setting to any other (campus or non-campus) setting.⇌

Note: Ignoring social preferences requires more than (wrong) claim that self interest **always** much more powerful. ⇌

- Rather, *also* requires⇌
  - No situations where actors can significantly affect well-being of others with only small effect on own well-being.⇌
- Note: Reference Dependence massively important in social prefs⇌
  - We ignore.

↬

Social Preferences

I am attitudy and puzzled by much of the research.$\rightleftharpoons$

- A huge industry in experiments on social preferences.$\rightleftharpoons$
- Most active area of both experimental and behavioral economics.$\rightleftharpoons$
- But not the most successful.$\rightleftharpoons$
  - Little integration of any lab lessons into mainstream economics.$\rightleftharpoons$
  - Movement towards conceptual tightness disappointingly slow.$\rightleftharpoons$
  - Movement towards serious empirical conclusions (including acknowledging ones staring us in the face) distressingly slow.$\rightleftharpoons$
  - And 100% self-interest not as far off the mark as many other assumptions behavioral economists seek to modify, so benefits lower.

$\looparrowright$

Social Preferences

More attitude:⇌

- Scientifically sinful to not disentangle obvious confounds when they *are* obvious, and when they are easy to disentangle.⇌
- Multiple theories for the same data rather than creating data to narrow down theories has been egregious.⇌
  - Even when data needed is really really really *really* easily created.⇌
- Part of reason: people sense lab data missing something of people's intuition about the world.⇌
  - Can't get envy in lab...mimic it with other social preferences!⇌
  - But that suggests not to use the lab—it does not justify adding confounds to the lab so that the data match your intuition.

↬

Social Preferences

Two examples, & competing hypotheses.⇌

- Ultimatum game and Prisoner's Dilemma⇌

	Accept	Reject
Share	5,5	5,5
Grab	8,2	0,0

	C	D
C	4,4	0,5
D	5,0	1,1

⇌

- Variants of these two gamesdominated much of the earlier social-preferences literature ...⇌
  - Despite being awful games to do science of social preferences with.⇌
- Now we have Dictator and Trust games!

↬

Social Preferences

Consensus empirical fact: people often (sacrifice money to) reject lop-sided offers in ultimatum game.$\rightleftharpoons$Why?$\rightleftharpoons$

- H1a: failed, mischievious, or lazy selfishness.$\rightleftharpoons$
- H2a: punishing obnoxious/unfair behavior.$\rightleftharpoons$
- H3a: hate coming out behind random other subject.$\rightleftharpoons$

Consensus empirical fact: people often cooperate with others in one-shot prisoner's dilemma, but *only if* they think partner will. $\rightleftharpoons$Why?$\rightleftharpoons$

- H1b: mistargeted attempt at repeated-game selfishness.$\rightleftharpoons$
- H2b: positive reciprocity, rewarding other.$\rightleftharpoons$
- H3b: implementing equitable outcome.$\rightleftharpoons$

H1a/b: great pedigree, popularity, and wrongness.$\rightleftharpoons$
H2a, H2b, *H3a,* and *H3b* all reasonable. $\rightleftharpoons$

- And could all be true

↬

Social Preferences

Laboratory evidence has inspired many models of "social preferences".⇌

- Three classifications, "distributional preferences", "intentions-based preferences," and "other belief-based preferences".⇌
- Note: economists prone towards particular assumption: altruism, either targeted (children) or need-based (charity).⇌
  - Both are majorly right. Economists should and do study targeted altruism, and nobody objects to the investigation of charity, etc.⇌
- But lab research has emphasized fairness, reciprocity, etc.

↬

Social Preferences

Distributional Preferences$\rightleftharpoons$

- Represent Person 0's preferences by $U_0(\pi_0, \pi_1, \pi_2, ...)$, $\rightleftharpoons$
  - where π_k is Person k's "material" utility/payoffs. Begin with:$\rightleftharpoons$
- "Disinterested distributional preferences": Allocations people choose for others, when their choices do not affect their own outcome?

↬

Social Preferences

Formally, $W_0(\pi_1, \pi_2, ...)$. Examples of (extreme) preferences:$\rightleftharpoons$

- Surplus-maximizing: $W_0 = \sum_k \pi_k$.$\rightleftharpoons$
  - (If π_k really "material hedonic return", then this is utilitarianism.)$\rightleftharpoons$
- Rawlsian/maximin: $W_0 = Min\{\pi_k\}_k \rightleftharpoons$
- Egalitarian: $W_0 = -\sum_k (\pi_k - \overline{\pi})^2 \rightleftharpoons$

Note: $\rightleftharpoons$

- Rawlsian preferences have a form of aversion to inequality, but they are monotonically increasing.$\rightleftharpoons$
- "Egalitarian preferences" are qualitatively more extreme dislike of unequal outcomes; you'd actually lower people's payoffs.)$\rightleftharpoons$

These are of course unrealistically extreme forms.

$\looparrowright$

Social Preferences

"Non-disinterested" distributional preferences:$\rightleftharpoons$ $U_0(\pi_0, \pi_1, \pi_2, ...) =$

$$(1 - k - l)\pi_0$$

$$+(k \cdot W_0(\pi_0, \pi_1, \pi_2, ...)$$

$$+(l \cdot D_0(\pi_0 - \pi_1, \pi_0 - \pi_2, ...),$$

where $k, l, k + l \in [0, 1]$.$\rightleftharpoons$

- Of course, components are interpretation; don't observe separately. $\rightleftharpoons$
- What are they?$\rightleftharpoons$
- Answer: $\rightleftharpoons$ self interest, $\rightleftharpoons$ disinterested principles$\rightleftharpoons$ "social comparison"

Social Preferences

CR's simplified parameterization of two-person preferences, based on FS.$\rightleftharpoons$

- $U_B(\pi_A, \pi_B) \equiv \rho\pi_A + (1-\rho)\pi_B$ when $\pi_B \geq \pi_A$.$\rightleftharpoons$
- $U_B(\pi_A, \pi_B) \equiv \sigma\pi_A + (1-\sigma)\pi_B$ when $\pi_B \leq \pi_A$.$\rightleftharpoons$

Exaggerates the kinkiness of preferences.$\rightleftharpoons$

- This is largely for simplification.$\rightleftharpoons$
- Surely generally true that $\rho \geq \sigma$.$\rightleftharpoons$
- But the rest up for empirical grabs.

$\looparrowright$

Social Preferences

A bunch of (not necessarily mutually-exclusive) examples:$\rightleftharpoons$

- $\rho = \sigma = 0 \implies$ pure self interest.$\rightleftharpoons$
- $\rho = 1, \sigma = 0 \implies$ disinterested pure Rawlsian$\rightleftharpoons$
- $\rho = \sigma = \frac{1}{2} \implies$ dinterested surplus maximizer$\rightleftharpoons$
- $1 \geq \rho \geq \sigma \geq 0 \implies$ Social-Welfare Preferences$\rightleftharpoons$
- $1 \geq \rho \geq 0 \geq \sigma \implies$ inequity averse$\rightleftharpoons$
- $\rho \geq 1 \geq 0 \geq \sigma \implies$ egalitarian?$\rightleftharpoons$
- $0 \geq \rho \geq \sigma \implies$ competitive?$\rightleftharpoons$

$\hookrightarrow$

Social Preferences

Lab Evidence on Social Preferences⇌

- In all examples, proportions choosing each of two money combinations (usually U.S. pennies or Spanish pesatas) are drawn underneath the amounts.⇌
  - From Kritikos & Bolle (2001), Charness & Grosskopf (2001), and Charness & Rabin (2002,2005):⇌
- Caution:⇌
  - These slides somewhat dated⇌
  - More evidence on some of this⇌
  - (Although remarkably little)

↬

Some Lab Data on Distributional Prefs

Evidence on disinterested distributional preferences?

- There is virtually no $-stakes evidence on disinterested preferences!
- From Charness and Rabin (2002):

C chooses (A,B) allocation of	(400,400)	vs.	(750,375)	
	.46		.54	
C chooses (A,B) allocation of	(400,400)	vs.	(1200,0)	
	.82		.18	

Keep these in mind.

- Especially the first one
- Disinterested care a lot about "equity"

Social Preferences

Evidence on Non-Disinterested Distributional Preferences?

- Quick, selective examples meant to illustrate. $\rightleftharpoons$
- (But meant not to be misleading.)$\rightleftharpoons$
- To ask purely distributional preferences, free of reciprocity, consider first only "dictator" games.$\rightleftharpoons$
  - My attitude begins here ... $\rightleftharpoons$
  - disentangle rather than confound.

$\looparrowright$

Social Preferences

What is the evidence on ρ?

B chooses (A,B) allocation of	(200,700)	vs.	(600,600)
Old C-R:	.27		.73

B chooses (A,B) allocation of	(0,800)	vs.	(400,400)
Old C-R	.78		.22
New C-R	.56		.44
New C-R with requests by A	.45		.55

My impressions from the accumulated experimental evidence:

- The average or median ρ is about .4.
- About 10% of subjects have $\rho < 0$.

↬

Social Preferences

Evidence on σ?$\rightleftharpoons$

- Shockingly (unforgivably) little evidence on this disentangled from negative reciprocity.$\rightleftharpoons$
  - Do people Pareto-damage when other has done nothing wrong?$\rightleftharpoons$
- Warning: might be misleading, because compared to other experiments, relatively little "Pareto damage".$\rightleftharpoons$

$\looparrowright$

Social Preferences

Small sacrifice to avoid coming out behind?⇌

B chooses (A,B) allocation of	*(1200,625)*	*vs.*	*(600,600)*	⇌
C–G	*88%*		*12%*	

B chooses (A,B) allocation of	*(750,400)*	*vs.*	*(375,375)*	⇌
C-R	*77%*		*23%*	

That 23% is the largest % I know of for reciprocity-free strict Pareto-damaging sacrifice, across all experiments ever run.

- Recent evidence, still absorbing, finds more.⇌
- (And certainly also more evidence of less.)

↬

Social Preferences

Significant sacrifice to avoid coming out behind?$\rightleftharpoons$

B chooses (A,B) allocation of	*(4,1)*	*vs.*	*(0,0)*	$\rightleftharpoons$
K-B	*88%*		*12%*	
B chooses (A,B) allocation of	*(3,2)*	*vs.*	*(0,0)*	$\rightleftharpoons$
K-B	*100%*		*0%*	
B chooses (A,B) allocation of	*(800,200)*	*vs.*	*(0,0)*	
C-R	*100%*		*0%*	

$\hookrightarrow$

Social Preferences

Small sacrifice to come out *further* behind?⇌

B chooses (A,B) allocation of C-G	*(625,625) 33%*	*vs.*	*(1200,600) 67%*	⇌
B chooses (A,B) allocation of C-R	*(400,400) 55%*	*vs.*	*(750,375) 45%*	⇌

- - Recall 1: Only 23% choose (375,375) over (750,400).⇌
  - Recall 2: 46% of *disinterested* choose (400,400) over (750,375).⇌
- **So:** more people sacrifice to come out **behind** when it really helps other than will sacrifice to avoid coming out behind! ⇌
- **And**: most who do not sacrifice to help:⇌ are refusing out of sincere Rawlsian motives, not because selfish or because hate coming out behind.

↬

Social Preferences

Costlessly take $ from other to avoid behind?⇌

B chooses (A,B) allocation of	*(X,0)*	*vs.*	*(0,0)*	
K-B	*75%*		*25%*	⇌
B chooses (A,B) allocation of	*(900,600)*	*vs.*	*(600,600)*	
C-G	*67%*		*33%*	⇌
B chooses (A,B) allocation of	*(750,400)*	*vs.*	*(400,400)*	
C-R	*68%*		*32%*	⇌
B chooses (A,B) allocation of	*(2000,400)*	*vs.*	*(400,400)*	
C-R	*82%*		*18%*	⇌

- More than I ever expected... $\frac{1}{3}$ taking $ away from other.⇌
  - But recall: behindness aversion says motive super-strong here.

↬

Social Preferences

Crude summary accumulated experimental evidence:$\rightleftharpoons$

- About 30% $\sigma < 0$, about 70% $\sigma > 0$.$\rightleftharpoons$
- Median $\overline{\sigma} > 0$, but very few $|\sigma| >> 0$.$\rightleftharpoons$
- Including very few $\sigma << 0$.$\rightleftharpoons$

Update on original examples and hypotheses? So far:$\rightleftharpoons$

- Little indication that rejections in the ultimatum game have much to do with "behindness aversion". $\rightleftharpoons$
  - Maybe it is all about negative reciprocity instead. $\rightleftharpoons$
- But lots of eidence of sharing when ahead ... very consistent with PD being from "inequity aversion".$\rightleftharpoons$
  - Does not preclude positive reciprocity.

$\hookrightarrow$

Social Preferences

Intentions-Based Preferences⇌

- People may care not *just* about outcomes, but with rewarding and punishing good and bad behavior.⇌
  - In bilateral context, we think of this as reciprocity. ⇌
- I emphasize **"just"** to make clear: ⇌
  - conceptually incoherent to have preferences that are just about reciprocity ... ⇌
  - people **must** have some notion of good and bad outcomes in order to be reciprocal about anything. ⇌
  - Rabin (1993), Dufwenberg-Kirchsteiger embed bad distributional⇌
  - Falk-Fischbacher, Charness-Rabin combine more serious distributional preferences with reciprocity.⇌

↬

Social Preferences

How to test for the role of intentions? $\rightleftharpoons$

- Very simple method:$\rightleftharpoons$
  - With common knowledge to the players, Player A chooses between outcome X and giving Player B the choice from $\{Y, Z\}$.$\rightleftharpoons$
  - Player B's preferences between Y and Z across situations, or when Player A has no choice, reflects B's distributional preferences.$\rightleftharpoons$
  - But the way Player B's preferences between Y and Z depend on changes in X reflect his reciprocal/intentions-based preferences.

$\looparrowright$

Social Preferences

Begin with the dark side: What induces "Pareto-damaging" behavior?⇌

- Behavior that hurts some or all without helping anybody.⇌

A chooses	*or*	*lets B choose*	*(800,200)*	*vs.*	*(0,0)*
No choice			*100%*		*0%*
fairer than	*(800,200)*		*81-92%*		*8-19%*

⇌

- Note: discussing only B behavior.⇌
  - A behavior in papers.⇌
- Reminder: these data not typical evidence in the literature, where more Pareto damage is typically observed.

↬

Social Preferences

A chooses	*or let*	*B choose*	*(750,400)*	*vs.*	*(375,375)*
No choice			*77%*		*23%*
fairer than	*(750,400)*		*71%*		*29%*
(400,750)			*80%*		*20%*

$\rightleftharpoons$

Increase is statistically significant, but not large. Note that seems that B does not punish A for not wanting to be on the short end herself.

$\hookrightarrow$

Social Preferences

When B makes choices where $ for A at stake, but not $ for B, how respond to goodness, badness?⇌

A chooses	*or*	*lets B choose*	*(750,400)*	*vs.*	*(400,400)*
No choice			*≈ 60%*		*≈ 40%*
(550,550)			*≈ 55%*		*≈ 45%*
(750,0)			*94%*		*6%*

⇌

- What is going on? ⇌
  - 1st vs. 2nd row? ⇌
  - 1st vs. 3rd?⇌
- Surprisingly little punishment, even when free. ⇌
  - A puzzle.⇌
- **But behindness aversion virtually vanishes when A has been kind.**

↬

Social Preferences

Crude summary of accumulated experimental evidence:$\rightleftharpoons$

- Not much Pareto-damage without reciprocity $\rightleftharpoons$
- Increase in Pareto-damage by B if A is mean/unfair$\rightleftharpoons$
- Stronger indication of diminishing Pareto-damage if A behaves nicely.

$\rightsquigarrow$

Social Preferences

How does good and bad behavior by one player affect the other player's inclination to engage in helpful sacrifice?⇌

	A chooses	*or*	*lets B choose*	*(750,375)*	*vs.*	*(400,400)*	
	No choice			*46%*		*54%*	
⇌	*(750±50,0)*			*37%*		*63%*	⇌
	(550,550)			*11%*		*89%*	

- Comment on 1st vs. 2nd line: ⇌
  - Whoa!⇌
- Lots and lots of data since we didn't believe the results. ⇌
  - It's robust. ⇌
  - Statistically significantly in opposite direction as positive reciprocity.

↬

Social Preferences

Accumulated evidence: $\rightleftharpoons$

- Lots of helpful sacrifice.
- *Not* increased by other's good behavior.$\rightleftharpoons$
- But withdrawn if other misbehaves.$\rightleftharpoons$

Some experiments find positive reciprocity, but very few. $\rightleftharpoons$

- My bet: meta-analysis would show people that positive reciprocity in sense of behaving <u>better</u> towards somebody if she has been good is approximately **zero** in the lab. $\rightleftharpoons$
- The "concern withdrawal", sort of in between positive reciprocity and negative reciprocity as conventionally conceived, seems robust.

$\curvearrowright$

Social Preferences

- Explanations, by Bozos like Rabin (1993), that cooperation in laboratory PD is from positive reciprocity, is wrong.$\rightleftharpoons$
- Similarly: massive experiments on "trust" seem really to be about ρ — little indication that trust is rewarded independent of desire to share.$\rightleftharpoons$
- I believe in positive reciprocity$\rightleftharpoons$
  - and have thoughts why missing from lab.$\rightleftharpoons$
  - but it is not in evidence whatsoever in the lab.

$\looparrowright$

Social Preferences

- Approximately:⇌
  - In UG: All negative reciprocity, no behindness aversion.⇌
  - In PD: All equity, no positive reciprocity.⇌
- Over-strong statements? ⇌
  - Stark statements to counter-balance the apparent triumph of priors and theory and pet models over easy-to-see and overwhelming empirical evidence.⇌
- The point is **not** a horse race, ⇌
  - not a claim about which motives bigger.⇌
- Estimates in lab on behindness aversion and positive reciprocity.⇌
- Both are (slightly) backwards.

↬

Social Preferences

Back to belief-based utility?⇌

- Reciprocity models (and cousins/descendants a la Levine "Spite") are clearly belief-based.⇌
- But so are the new models based on image.⇌
- Clearly right direction⇌
  - Even if some of them formulated in complacent terms.⇌
  - And must be "seeded" by distributional assumptions.⇌
  - Because all models must.

↬

Social Preferences

Great early experiment by Dana, Weber, and Kuang (2007):$\rightleftharpoons$

- (other, self) of $A \equiv (\$5, \$5)$ vs. $B \equiv (\$1, \$6)$.$\rightleftharpoons$
  - 26% choose $B \equiv (\$1, \$6)$.$\rightleftharpoons$
- (other, self) of $C \equiv (\$1, \$5)$ vs. $D \equiv (\$5, \$6)$. $\rightleftharpoons$
  - Presumptively, 100% choose D. $\rightleftharpoons$
  - And, in hypothetical treatment, 100% in fact chose it.$\rightleftharpoons$
- Then: people told to choose either \$5 or \$6 for self.$\rightleftharpoons$
  - But 50% chance $(\$\_\_, \$5)$ vs. $(\$\_\_, \$6)$ is A *vs.* B.$\rightleftharpoons$
  - 50% chance it is C *vs.* B.

$\looparrowright$

Social Preferences

- 37% choose to Not reveal, take \$6.$\rightleftharpoons$
- 7% choose to Not reveal, take \$5.$\rightleftharpoons$
- 56% choose to Reveal.$\rightleftharpoons$
- If revealed and saw (\$5,\$5) vs. (\$1,\$6),$\rightleftharpoons$
  - 25% choose (\$1, \$6)$\rightleftharpoons$
- If revealed and saw (\$1,\$5) vs. (\$5,\$6),$\rightleftharpoons$
  - 90% choose (\$5, \$6). (So 10% are either silly or nasty)$\rightleftharpoons$
- And so if choice turned out to be (\$5,\$5) vs. (\$1,\$6), 63% (rather than 26%) ended up choosing (\$1,\$6), either by choosing it after seeing it, or by choosing not to see it.

$\looparrowright$

Social Preferences

Interpretation?⇌

- "Moral Wiggle Room". Authors (incessantly) point out cannot explain by distributional preferences. Why not reveal payoffs?⇌
- One interpretation: self-image preservation.⇌
- Grossman (2008) and Lazear, Malmendier, and Weber (2007), Tadelis, etc.: variants of people seeming to have some variant of belief-based social preferences.
  - Shame? ⇌
  - Social signaling? ⇌
  - Self signaling? ⇌
- People care how obvious it is to themselves and to others that they are being selfish or virtuous.

↬

Social Preferences

Final point .... $\rightleftharpoons$

- "Wiggle room $\longrightarrow$ non-selfish behavior may be less important than it appears" is not an awful intuition.$\rightleftharpoons$
- But it's not as compelling as may seem.$\rightleftharpoons$
- A world where people convince themselves they are right to justify selfish behavior may be *more* different from simple selfishness than is simple simple social preferences.$\rightleftharpoons$
- Bargaining selfish people vs. bargaining altruists vs. bargaining self-righteous people?

$\looparrowright$

Social Preferences

Modeling Reciprocity? Hard.$\rightleftharpoons$

- I mean hard in both the sense of technically difficult and in terms of really finding robust models. Other- or self-signaling use signaling models (and moral wiggle room and related experiments seem to require weirdness; Rabin (1995) "moral rules" vs. "moral preferences" with weird assumption about treatment of information. Reciprocity uses "psychological games" by Geanakoplos, Pearce, and Stacchetti (1989), or related approaches, such as Levine's signaling approach or steady-state equilibrium a la Charness & Rabin.$\rightleftharpoons$
- Technical issues in modeling reciprocity: Both as psychological reality and as modeling strategy, not just outcomes, nor even cleverly defined to include game, but beliefs.$\rightleftharpoons$
- How might we capture the role of volition and intentions in models of social preferences?

$\hookrightarrow$

Social Preferences

Claim: Often no $U_j(\pi_i, \pi_j)$ can capture preferences.$\rightleftharpoons$

- Nor, to slightly generalize, can capture by writing payoffs solely function of outcomes.$\rightleftharpoons$
- Point is not that you can't do in terms of own payoff. Duh. The point is that it is not a function of everybody's payoff ... but different things. In any framework, psychological games of type-signaling games, need beliefs in there.$\rightleftharpoons$
- Think about the following game. Fully hypothetical; no data. Going to 3-person for ease of illustration.

$\hookrightarrow$

Social Preferences

- Want you to think about the following game (using intuition, like I am, since nobody has run the game to my knowledge). Going to 3-person for ease of illustration.⇌

	1 goes L 2 chooses		1 goes R 3 chooses	
	L	R	L	R
1's Payoff	10	12	11	5
2's Payoff	10	12	5	5
3's Payoff	10	0	5	5

⇌

What do you think Player 3 will do if Player 1 goes R? What are the issues?⇌

- If Player 3 thinks Player 2 would have gone L?⇌
- If Player 3 thinks Player 2 would have gone R?⇌
- What is not right about those questions?