

Lecture 10: Social preferences and neural evidence

1. Economic theory has never been committed to the idea that people are narrowly selfish, in the sense of caring only about their material self interest. Utility functions can take any factors as arguments, including elements that are factors just insofar as they affect the welfare of others. It is an obvious fact about people that most of them will pay significant costs, under ranges of circumstances, to help – and to hinder – the fortunes of specific other people known to them and (more rarely and less reliably), the fortunes of abstractly conceived and anonymous people.
2. To say that people aren't narrowly selfish doesn't take us very far in the direction of positive theory. Humans are hyper-social animals, with various kinds of social motivations that interact in complex ways. We can conceptually distinguish at least the following:
  - (i) *Weak reciprocity*: A's utility is augmented when she treats B as B treated her, or, if there is no direct history between A and B, as she observed B treating C.
  - (ii) *Strong reciprocity*: A's utility is augmented when she justifiably punishes B for violating a norm of weak reciprocity, in interaction with A or with another.
  - (iii) *Norm sensitivity*: A's utility is negatively impacted when she believes a norm has been violated, by others or by herself. Violations of a norm are evidence that the norm is collapsing, so violations of a given norm have diminishing marginal effect on A's utility.

- (iv) *Benevolence*: A's utility is augmented when she gets information that leads her to expect an improvement in B's welfare.
- (v) *Solidarity*: A's utility is augmented when she observes a signal – which might consist in her own action – that she and B are coordinating their individual utility maximization so as to co-identify as a 'team'.
- (vi) *Inequality aversion*: A's utility is negatively impacted when she observes an outcome of an interaction that produces a distribution of assets at stake that is far from equal division, unless this is due either to proportionate skew in the ratio of quality of the parties' contributions, or the disadvantaged party violated a norm.
- (vii) *Character maintenance*: A's utility is negatively impacted when she acts in a way that signals to others that an autobiographical characteristic in which she has invested, and which she thinks will encourage others to interact with her, is unreliable; her utility is augmented when she acts in a way that signals to others that an autobiographical characteristic that has been associated with her, and which she thinks discourages others from interacting with her, is unreliable.

Each of these possible aspects of A's motivational structure will be reflected differently in the utility function we would write down for her. A utility function that simultaneously reflected *all* of them would be consistent with a very wide and complex range of behaviors – perhaps too wide and complex for empirical testing. In experiments, we therefore try to constrain situations so that only one of the hypothesized factors above will dominate the others and govern choice. Note, however, that we would expect the

motivations to be tightly intertwined. Trying to fabricate situations in which only one of them is relevant might lead to putting people in situations they regard mainly as *weird*. How much do we learn about normal human behavior by studying the behavior of confused people?

3. Note that Fehr, in the chapter at hand, rules out motivation (vii) as not really being a *social* preference, since it expresses an agent's concern for the value of her reputation for her *own* well being. We should expect motivation (vii) to be active whenever people play repeated games, and games with identified opponents whose views they might wish to influence. Fehr therefore argues that repeated games and non-anonymous games are non-ideal vehicles for studying social preference in his restricted sense. One takes the point. But two critical concerns are relevant here. First, Ken Binmore has argued that, in advance of significant learning, people tend to play one-shot games as if they were repeated games. This is simply because they lack experience of one-shot anonymous games, which rarely happen outside the lab. Second, Fehr's concept of social preference doesn't clearly distinguish among motivations (i) – (vi). I think he is aware of this, but we should bear it in mind in following him through his interpretations of the neuroeconomic findings.
4. Fehr also argues that the PD game is poorly structured for investigating social preferences. A player may defect in a PD either because he is narrowly selfish or because he thinks that his opponent is narrowly selfish. Thus one cannot make clean inferences about utility functions from observation of PD play. Note that experimental PDs furnish the best evidence for Binmore's concern above: as people learn to play one-shot PDs, they learn to play Nash equilibrium. This is probably not because they're narrowly selfish, but because they learn that non-Nash outcomes can't be reliably predicted

by any consistent model of the players.

5. Fehr also suggests that games used to investigate social preferences should be sequential, to reduce the extent to which players are called upon to model their opponents based on general theories about the population. Note that this aspect of tasks can never be entirely eliminated, because observation of a move in a game is almost always consistent with more than one imputation of a utility function.
6. We don't know nearly as much as we'd like to about the intrapersonal stability of social preferences. Are social preferences more like character traits or more like ecological channels (that is, social-environmental constraints on behavior that influence people's behavior as a function of the extent to which they're perceptually cued)? Note that the interpretation of social preferences as ecological channels is a *more* radical departure from descriptive individualism than is the interpretation of social preferences as standing aspects of people's utility functions.
7. People's behavior suggests wide heterogeneity with respect to social preferences. Knowledge of this should inspire rational agents in games to make moves designed to test their models of one another. For example, in a repeated ultimatum game a proposer might start with a stingy offer in order to explore the preferences of the other player – a 'fair' offer elicits no information. Of course, in a true one-shot game such probing is pointless. But what if people underestimate the extent of heterogeneity in the population? Or suppose that they model social preferences as character traits when they're really more like cue-elicited ecological channels? In either case people might be motivated to run tests in early encounters that they use as an inferential basis for strategy

choice in later encounters.

8. Fehr mentions evidence of *betrayal aversion* – fear of being betrayed by another person. This may lead people to demonstrate less trust (in, e.g., one-shot anonymous TGs) than people’s social preferences would otherwise lead them to choose.
9. With some appreciation of the inferential minefield in which we’re working, let’s now follow Fehr’s review of the neuroeconomic evidence to date.
10. We have already discussed the Sanfey *et al* experiment with punishment behavior in UGs. Fehr suggests that the activation in anterior cingulate they observed as associated with receipt of stingy offers may be related either to emotional resentment or to motivational conflict between desire to punish and desire to harvest a reward: ACC activation may reflect the effort of cost comparison. Given background knowledge about dlPFC, it is suggested that its activation here indicates cognitive control of an impulse to inflict punishment. Tangential support for this hypothesis is suggested by a later experiment in which PFC activity was differentially associated with stingy offers that were accepted, as contrasted with stingy offers that were rejected.
11. Knoch *et al* did just the right experiment in this context. They used TMS to interfere with dlPFC activation during UG play. Instead of increasing the rejection rate on stingy offers, as the original impulse control hypothesis predicts, the acceptance rate went up. Furthermore, attenuation of dlPFC activation had less impact on acceptance rates of stingy offers generated by computer opponents.

12. On this basis two further (rival) hypotheses are constructed: (i) disruption of dlPFC interferes with subjects' disposition to perceive stingy offers as unfair; (ii) disruption of dlPFC interferes with subjects' dispositions to control their selfish impulses. The second hypothesis is supported by evidence that TMS interference with dlPFC doesn't disrupt subjects' reported judgments about unfairness.
13. It is striking to consider the idea of older brain areas as narrowly selfish maximizers, while a new brain area encourages pro-social behavior – by inspiring people to inflict revenge! (Let this serve as a warning against thoughtlessly importing folk expectations about what kinds of behavior count as impulsive and what kinds count as reflective.)
14. In light of evidence that vmPFC is implicated in integration of expectations about emotional costs and benefits, Fehr favors the following hypothesis as the current best interpretation of all these data on subjects' responses to UG offers during TMS disruption of dlPFC: “[L]ow-frequency TMS of right dlPFC induces an impairment in the integration of the emotional cost of accepting an unfair offer. Such impairment could be caused by possible *network effects* of TMS that diminish the functioning of the vmPFC” (emphases added). Let us reflect on the implications of interpretations of neural probes that appeal to network effects.
15. Later in the chapter (pp. 226-228), Fehr reports experiments in which subjects show enhanced dlPFC and OFC activation when they play UGs than when they play Dictator games, and when they play UGs against people than when they play UGs against computers. This is consistent with the hypothesis that dlPFC mitigates impulses to

maximize narrow self-interest.

16. Fehr next cites evidence that, controlling for reward associated with monetary gains, subjects show extra striatal activity when they achieve cooperative outcomes. This is of course taken to suggest that cooperation is rewarding in itself (consistent with utility factors (i), (iv) and (v) above).
17. “Social preference theories also predict that subjects prefer punishing unfair behavior, such as defection in public good and PD games, because leaving an unfair act unpunished is associated with higher disutility than bearing the cost of punishing an unfair act.” In this light, Fehr considers a TG experiment by de Quervain *et al* in which investors could subsequently punish trustees who betrayed them. In one condition punishment was costly to defectors, in the other (“symbolic”) condition it wasn’t. PET data found caudate activity positively associated with the costly punishment by comparison with the costless punishment. Singer *et al* compared male and female striatal activity when subjects could administer shocks to defectors in sequential PDs. Men and women showed activity in anterior cingulate – associated with empathy – when cooperative partners were exogenously shocked. But men, and not women, showed NAcc and OFC activation, correlated with their reported desires for revenge, when they administered shocks to defectors.
18. Fehr next reviews studies of the reward system while subjects have opportunities to make charitable donations. The system is activated both by receipt of money and, independently, by decisions to donate. Fehr also reports an experiment in which subjects show enhanced reward circuit activity when they observe a charity receive more money and they themselves also receive more money. I can’t determine

from the summary why this is supposed to be interesting.

19. Golnaz *et al* found that reward circuitry was more active when UG players received fair offers than when they received less fair offers of identical monetary value (controlling for subjects' wealth).
20. Do people behave altruistically toward others *because* this delivers augmented hedonic reward? Fehr reviews a few experiments in which striatal activity associated with altruistic choices predicts further altruistic choices in other conditions. Since a primary job of the reward system is learning to predict outcomes of choices, this may simply reflect the fact that more altruistic subjects have learned to monitor the effects of their altruism, which would hardly be surprising. I don't regard the 'causal evidence' Fehr reviews here as very persuasive.
21. Kosfeld *et al* found that TG players infused with oxytocin (OT) exhibited greater trust than controls, even though OT didn't affect their attitudes to risk in parametric decisions, and didn't affect their cognitive judgments about their chances of being repaid. Other work suggests that this effect works through OT inhibition of amygdala, which is interpreted as indicating that OT suppresses fear of betrayal.
22. Are emotional and reward processing associated with social preference distinct from evaluations of players' reputations? An experiment by Singer *et al* suggests that they are intertwined. Singer *et al* elicited PD players' responses to pictures of faces of people who had cooperated with them, as contrasted with pictures of unfamiliar faces. The faces associated with good reputations activated both reward circuits and those associated with 'emotional' responses. King-Casas *et al* found that caudate activity encodes learned

positive reputations of other players in repeated TGs. Baumgartner *et al* found they could block the behavioral impact of negative reputation learning in TG players by administering OT; furthermore, the area that showed reduced activation under OT treatment was caudate.

23. If neural encoding of reputational effects and reward system and amygdala activation in response to pro-social outcomes are tightly intertwined, what does this suggest about prospects for using neuroscience to isolate the distinct elements of utility listed at the beginning of this lecture?
  
24. People achieve cooperative outcomes far more reliably in face-to-face games than in anonymous ones. Fehr suggests that much of this is due to the enormous signaling bandwidth that facial movements have for people. It is worth reflecting on the implications of this for differences between life in spatially restricted communities and life in the virtual global workspace. Should we expect people to behave more like narrowly selfish maximizers as higher proportions of their interactions occur via electronic networks? Alternatively, might advancing use of video links (e.g., Skype, or integration of social networks with streamcasting) correct for this?