Fairness and Reciprocity

Michele Albach
March 26th, 2019

Questions to Answer

- What does it mean to act fairly?
- When do people act fairly?
    - What games exhibit fairness?
- Why do people act fairly?
    - What factors affect fairness?
- How can fairness be modelled?
    - Which models best support observed evidence?
- Future work?
    - What still needs to be done?

Outline

- Motivation
  - Games demonstrating fairness
- Modelling Fairness
  - Intentions-based models
  - Outcome-based models
  - Combining Intentions and Outcomes
- Comparing Models
- Recent Work
- Future Work

Motivation

Games demonstrating fairness

Fun Game: The Ultimatum Game (UG)

Güth *et al.*, 1982

There are two roles: the proposer and the responder

- The proposer has been given some amount (say $10) and must make an offer to split the amount with the responder
- The responder may either accept or reject the offer
- If the offer is accepted, the money is split according to the offer
- If the offer is rejected, both players receive nothing
- Play this game at least once as each role

Equilibrium in the Ultimatum Game

Güth *et al.*, 1982

- Since something is better than nothing, the responder should accept any positive offer
- Knowing this, the proposer should offer the smallest amount possible
- Experimental data does not support this equilibrium, why?
    - Because proposers want to be fair?
    - Because proposers are afraid that their offer will be rejected?
    - Other reasons?

UG Results

Güth *et al.*, 1982

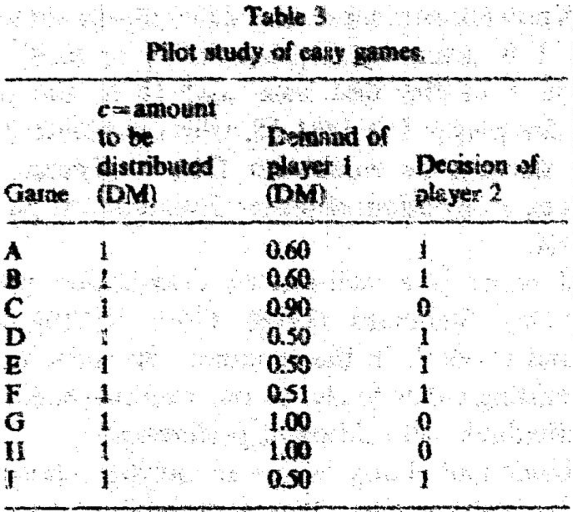
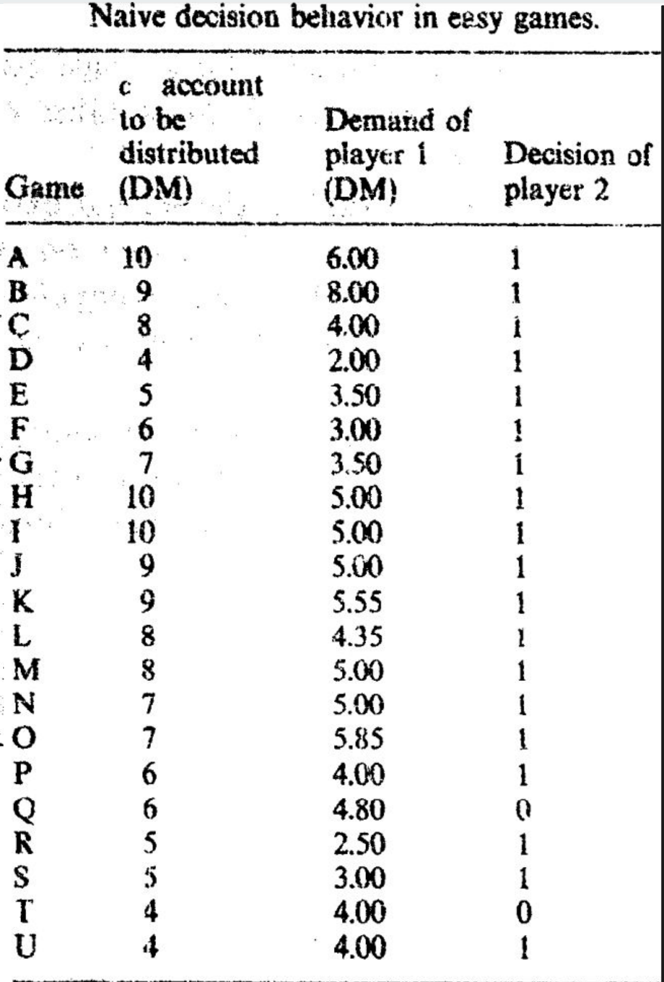


Table 3

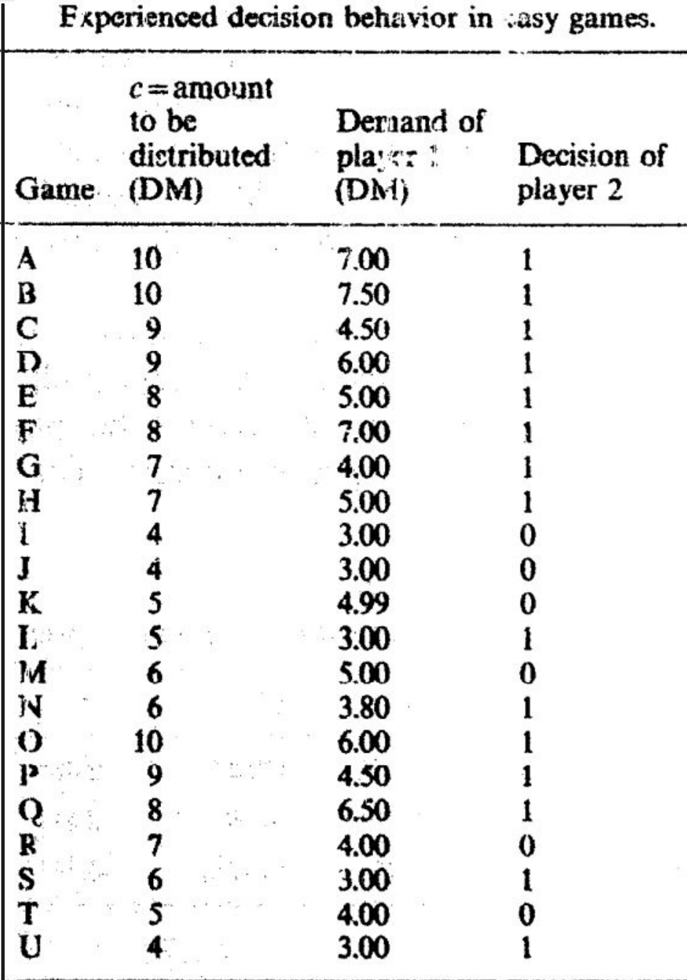
Pilot study of easy games.

Game	c = amount to be distributed (DM)	Demand of player 1 (DM)	Decision of player 2
A	1	0.60	1
B	1	0.60	1
C	1	0.90	0
D	1	0.50	1
E	1	0.50	1
F	1	0.51	1
G	1	1.00	0
H	1	1.00	0
I	1	0.50	1



Naive decision behavior in easy games.

Game	c account to be distributed (DM)	Demand of player 1 (DM)	Decision of player 2
A	10	6.00	1
B	9	8.00	1
C	8	4.00	1
D	4	2.00	1
E	5	3.50	1
F	6	3.00	1
G	7	3.50	1
H	10	5.00	1
I	10	5.00	1
J	9	5.00	1
K	9	5.55	1
L	8	4.35	1
M	8	5.00	1
N	7	5.00	1
O	7	5.85	1
P	6	4.00	1
Q	6	4.80	0
R	5	2.50	1
S	5	3.00	1
T	4	4.00	0
U	4	4.00	1



Experienced decision behavior in easy games.

Game	c = amount to be dictributed (DM)	Demand of player 1 (DM)	Decision of player 2
A	10	7.00	1
B	10	7.50	1
C	9	4.50	1
D	9	6.00	1
E	8	5.00	1
F	8	7.00	1
G	7	4.00	1
H	7	5.00	1
I	4	3.00	0
J	4	3.00	0
K	5	4.99	0
L	5	3.00	1
M	6	5.00	0
N	6	3.80	1
O	10	6.00	1
P	9	4.50	1
Q	8	6.50	1
R	7	4.00	0
S	6	3.00	1
T	5	4.00	0
U	4	3.00	1

Other games that exhibit fairness: Altruism

- Dictator Game (DG) (Forsythe *et al.*, 1994)
  - Same as the ultimatum game but the responder must accept
  - Proposers offer less, some offer nothing (36%), but some still offer positive amounts
    - So results from the ultimatum game are not only due to fairness
- Gift Exchange Game (GEG) (Fehr *et al.*, 1993)
  - An employer offers a 'wage' *w* to a worker
  - If accepted, the worker chooses an 'effort level' *e* to give in return
  - Employers cannot enforce effort levels
  - Employers receive a payoff of *ve-w* for some value of effort *v*
  - Workers receive a payoff *w-c(e)* for some effort cost function *c*
  - "At the individual level reciprocal behaviour is the dominant behavioural pattern" (Fehr *et al.*, 1993)
    - Workers give increasingly positive values for *e* with increasing values for *w*
  - Would this result change in single-shot *vs.* repeated games?
    - Gaechter and Falk, 2001
    - Effort levels increase with repeated interaction, but are also observed in single-shot games

Other games that exhibit fairness: Spitefulness

- Public Good Games (PGG) (Fehr and Gächter, 2000)
  - N subjects are each given an amount y and simultaneously choose to invest g_i ($0 \leq g_i \leq y$) into a public goods project
  - No-punishment treatment:
    - The payoff of each subject is $y - g_i + a\sum g_j$ where a is some per capita return on the project and g_j is the amounts contributed by the other subjects
    - a is set ($0 < a < 1 < na$) so that the best outcome is if all subjects contribute 100% of y
  - Punishment treatment:
    - In a second stage of the game, after all players see everyone else's contributions, players can choose to punish each other at a cost to their own payoff
  - Punishing others is a dominated strategy, so results should be the same in both treatments
  - Results:
    - Punishment occurs
    - Investments converge to zero over repeated interactions in the no-punishment treatment
    - Investments are on average 58% of y in the punishment treatment (and do not change over time)

Other games that exhibit fairness: Heterogeneity

- Trust Games (Berg *et al.*, 1995)
    - A trustor has some amount *y* and can choose to send *x* ($0 \le x \le y$) to the trustee, who actually receives *3x*
    - Then, the trustee can choose to send some amount *z* ($0 \le z \le 3x$) back to the trustor
    - Results:
        - Trustors sent varying amounts
        - Out of 28 trustees who were sent more than *x* = $1:
            - Some trustees sent back nothing or $1 (12)
            - Some trustees sent back more than what was sent to them (11)
        - So not all individuals act fairly, but some do

How to model fairness?

- There are two main categories for models of fairness (Fehr and Schmidt, 2003):
  - Intentions-based models
    - Players judge how kind their opponents are being by perceiving their intentions
  - Outcome-based models (social preference)
    - Players care about the outcomes that their opponents receive as well as their own outcome

Intentions-Based Models

Rabin Fairness (1993)

Rabin attempted to define the emotional responses behind fairness in 3 points:

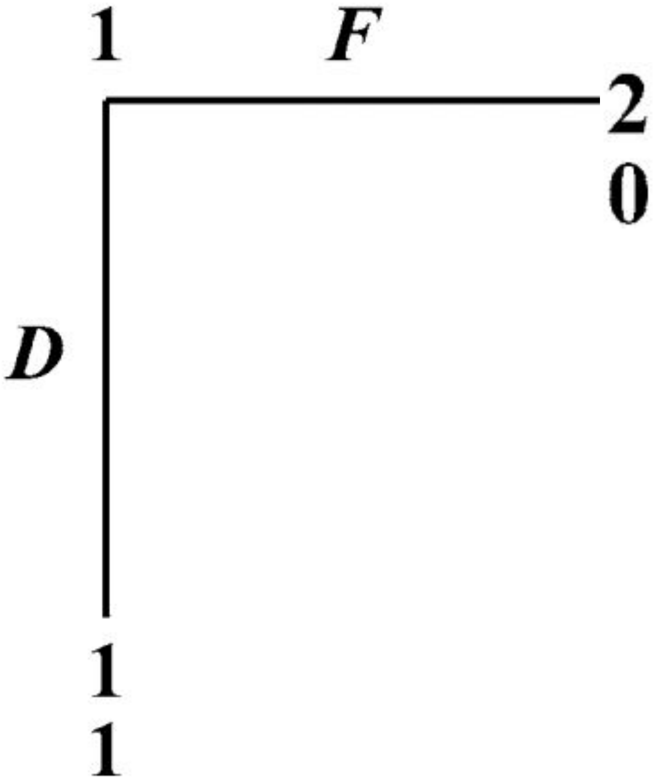
“

- People are willing to sacrifice their own material well-being to help those who are being kind
- People are willing to sacrifice their own material well-being to punish those who are being unkind
- Both [previous motivations] have a greater effect on behaviour as the material cost of sacrificing becomes smaller

“

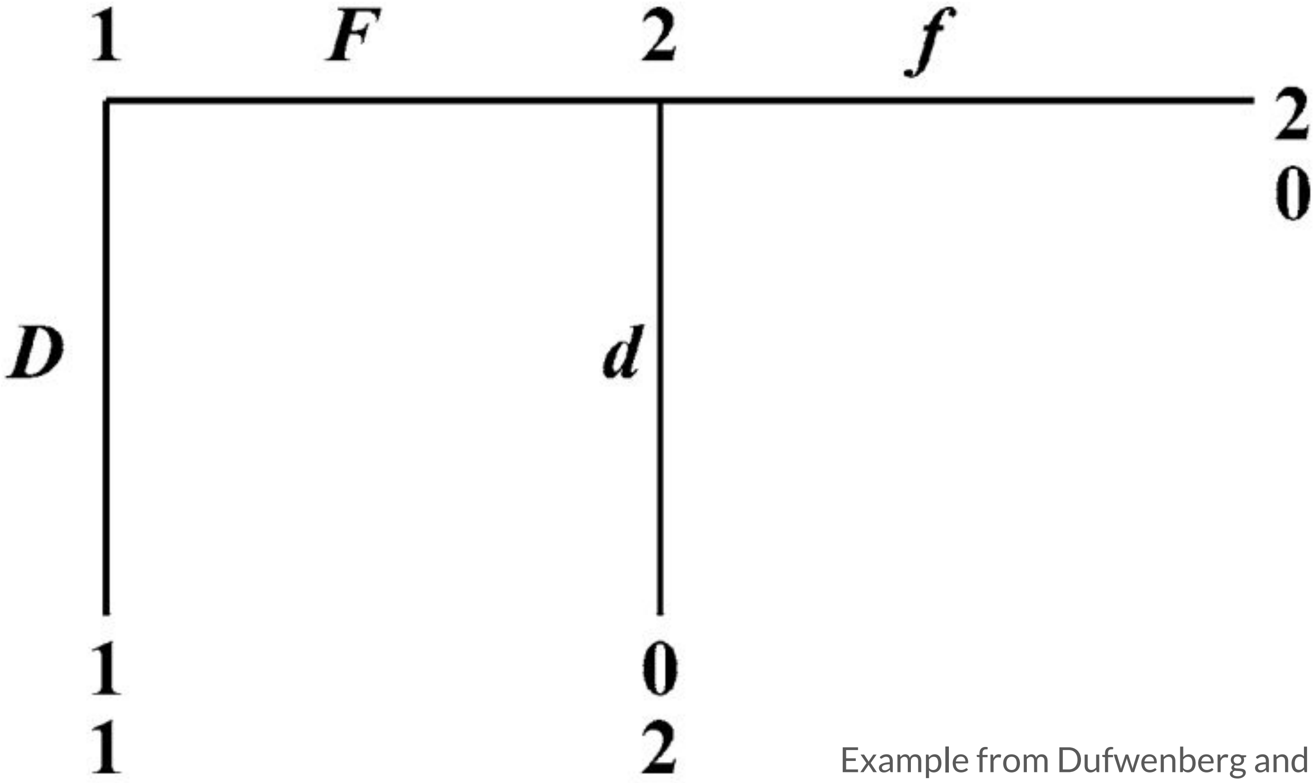
The first two points are the definition for reciprocity

What does it mean to be kind?



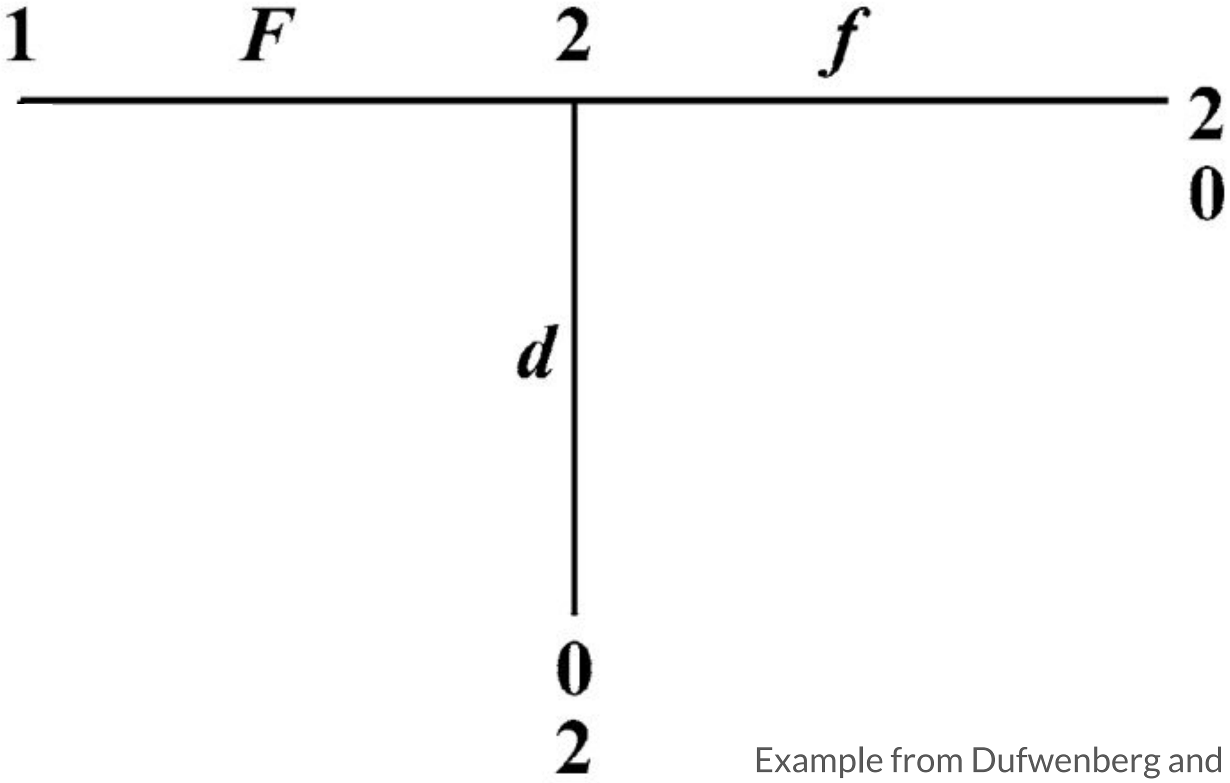
Example from Dufwenberg and Kirchsteiger, 2004

What does it mean to be kind?



Example from Dufwenberg and Kirchsteiger, 2004

What does it mean to be kind?



Example from Dufwenberg and Kirchsteiger, 2004

Rabin Fairness (1993)

- Only defined for 2-player normal form perfect information games (players *i* and *j*)
- Define:
  - a_i is player *i*'s action, b_i is the action that *j* believes *i* will play, and c_i is the action that *i* believes that *j* believes *i* will play
  - π is the material payoff function
  - $\pi_j^h(b_j)$ is player *j*'s highest possible payoff if they play b_i
  - $\pi_j^\ell(b_j)$ is player *j*'s lowest possible payoff out of non-Pareto-dominated points if they play b_j
  - $\pi_j^e(b_j) = [\pi_j^h(b_j) + \pi_j^\ell(b_j)]/2$ is the 'equitable payoff'
  - $\pi_j^{min}(b_j)$ is player *j*'s worst possible payoff if they play b_j
- Define a kindness function $f_i(a_i,b_j)$ measuring *i*'s kindness towards *j*:

$$f_i(a_i,b_j) \equiv \frac{\pi_j(b_j,a_i) - \pi_j^{\mathrm{e}}(b_j)}{\pi_j^{\mathrm{h}}(b_j) - \pi_j^{\mathrm{min}}(b_j)}.$$

$$\text{If } \pi_j^{\mathrm{h}}(b_j) - \pi_j^{\mathrm{min}}(b_j) = 0, \text{ then } f_i(a_i,b_j) = 0.$$

- Player *i*'s belief about how kind *j* is being to them is defined similarly as $f_j(b_j,c_i)$

Rabin Fairness (1993)

- Expands on Geanakoplos *et al.*’s (1989) model for ‘psychological games’
  - Allows utilities to depend on player’s beliefs as well as actions
- Adds the kindness function to utility:

$$U_i(a_i,b_j,c_i) \equiv \pi_i(a_i,b_j) + \tilde{f}_j(b_j,c_i)\cdot\left[1+f_i(a_i,b_j)\right]$$

- Next, uses Geanakoplos *et al.*’s concept of ‘psychological Nash equilibrium’ to define ‘fairness equilibrium’
  - (a_1,a_2) is a fairness equilibrium if for i = 1,2, a_i is best responding and $a_i = b_i = c_i$

Critiques of Rabin Fairness

- Limited to 2-player normal form games
- Assumes players are homogeneously fair
- Creates multiple and sometimes unrealistic fairness equilibria
  - Always at least one kind equilibrium and at least one unkind equilibrium
  - In UG, creates equilibria in which the responder receives more than 50% (Fehr and Schmidt, 2003)
- Is fairness actually more prominent with smaller material cost?
  - If so, could assume that fairness is less prominent with higher material cost
  - Research has found conflicting results:
    - Cameron, 1999 found that offers were still rejected at higher stakes
    - Anderson *et al.*, 2011 found that rejections decreased at higher stakes
  - Note: These studies were done in developing countries (Indonesia and Northeast India) to allow for higher payoffs
    - This brings into play questionable ethics and various factors that could affect results

Extending to Sequential N-player games

Dufwenberg and Kirchsteiger, 2004

- Allow beliefs to change, dependent on the history of the game
- Extend the kindness function to depend on history
  - Note that they remove Rabin's normalization for simplicity

$$\kappa_{ij}\left(a_i(h), \left(b_{ij}(h)\right)_{j\neq i}\right) = \pi_j\left(a_i(h), \left(b_{ij}(h)\right)_{j\neq i}\right) - \pi_j^{e_i}\left(\left(b_{ij}(h)\right)_{j\neq i}\right)$$

- Redefine utility to include reciprocity with all other players
  - $Y_{ij} > 0$ represents how much *i* cares about being reciprocal to *j*

$$\begin{aligned} &U_i\left(a_i(h), \left(b_{ij}(h), \left(c_{ijk}(h)\right)_{k\neq j}\right)_{j\neq i}\right) \\ &\quad = \pi_i\left(a_i(h), \left(b_{ij}(h)\right)_{j\neq i}\right) \\ &\qquad + \sum_{j\in N\setminus\{i\}} \left(Y_{ij} \cdot \kappa_{ij}\left(a_i(h), \left(b_{ij}(h)\right)_{j\neq i}\right) \cdot \lambda_{iji}\left(b_{ij}(h), \left(c_{ijk}(h)\right)_{k\neq j}\right)\right) \end{aligned}$$

- Define a sequential reciprocity equilibrium (SRE) similarly to fairness equilibrium
- Sebald, 2010 extends Dufwenberg and Kirchsteiger's model to allow for chance (nature player)

Outcome-Based Models

Altruistic or Spiteful?

Levine, 1998

- Gives all players a coefficient of altruism: $-1 < a_i < 1$
    - If $a_i > 0$ player *i* is altruistic, if $a_i < 0$ player *i* is spiteful, if $a_i = 0$ player *i* is selfish
- Update utility to incorporate other player's outcomes (u_j)

$$v_i = u_i + \sum_{j \neq i} \frac{a_i + \lambda a_j}{1 + \lambda} u_j$$

- Assumes lambda is the same for everyone
    - Estimates using ultimatum game data from Roth *et al.* (1991), finds lambda = 0.45
- Levine shows that his model can explain results from other games
    - Auction game, centipede, public good game
- Problems with this model
    - Cannot explain altruistic results from dictator games
    - Assumes individuals are consistently either altruistic or spiteful

Inequity Aversion

Fehr and Schmidt, 1999

- Motivated by Loewenstein *et al.*, 1989
  - Asked subjects to react to described situations in which they and another person would receive some payoffs
  - Found that individuals preferred equality over both disadvantageous and advantageous inequality
- Fehr and Schmidt allow for players who are averse to both disadvantageous and advantageous inequality
  - Assumes that they are more averse to disadvantageous inequality
- Define:
  - x_i as player *i*'s material payoff
  - β_i ($0 \leq \beta_i < 1$) represents *i*'s aversion to advantageous inequality
  - α_i ($\beta_i \leq \alpha_i$) represents *i*'s aversion to disadvantageous inequality
- Update utility to include aversion to inequality

$$U_i(x) = x_i - \alpha_i \frac{1}{n-1} \sum_{j \neq i} \max\{x_j - x_i, 0\} - \beta_i \frac{1}{n-1} \sum_{j \neq i} \max\{x_i - x_j, 0\}$$

- Fehr and Schmidt show that their model can explain game results
  - UG, market games, PGG
  - However, their model predicts too extreme results in some games like DG and GEG

ERC: Equity, Reciprocity, and Competition

Bolton and Ockenfels, 2000

- Uses a ‘social reference point’: the average of all player’s payoffs
- Define σ_i as player *i*’s relative share of the total payoff:

$$\sigma_i = \sigma_i(y_i, c, n) = \begin{cases} y_i/c & \text{if } c > 0 \\ 1/n & \text{if } c = 0 \end{cases} \qquad c = \sum_{j=1}^{n} y_j$$

- Extend utility to depend on σ_i
- Problem: using the average payoff *vs.* comparing to each opponent (Fehr and Fischbacher, 2004)
    - Performed ‘third-party’ dictator game
        - Player 1 is given 100 points and can choose to give some to player 2
        - Player 3 is given 50 points (⅓ of the total payoff)
        - After seeing player 1’s choice, player 3 can choose to punish them at a cost to their own total points
    - 26% (n = 46) of third-parties punished when player 1 offered player 2 less than 50 points
    - So players care about equity over all players, not just themselves

Intentions or Outcomes?

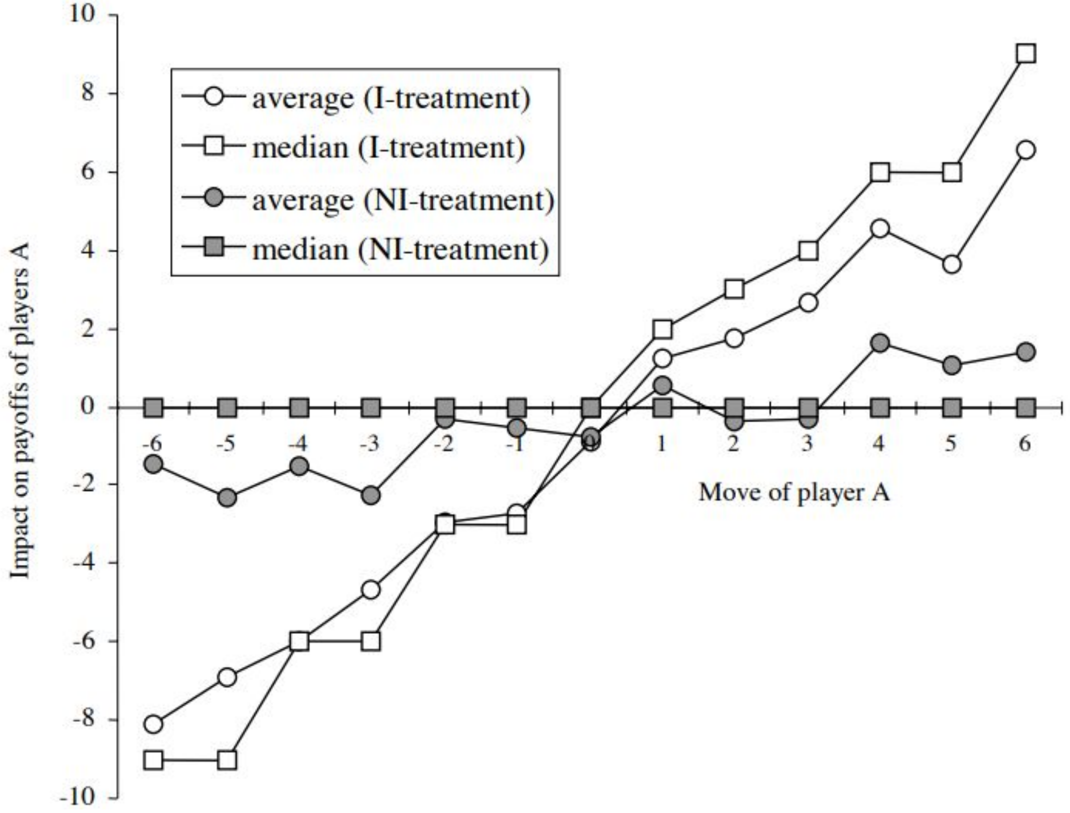
Intentions- *vs.* Outcome-Based Models

Falk, Fehr, and Fischbacher, 2008

- Moonlighting game (Abbink *et al.*, 2000)
    - Imagine an illegal moonlighter has taken a job from a client and been given funds to complete it
    - The moonlighter can choose to either complete the job or to take the funds and run
    - Next, the client can choose to either pay the moonlighter, do nothing, or attempt to punish the moonlighter at the cost of their own wellbeing
        - The activity is illegal so that any initial contract cannot be legally enforced
- Falk, Fehr, and Fischbacher's version of the game:
    - Each player is given 12 points
    - In the first stage, player A can choose to give or take up to 6 points to/from player 2
        - If they give *x* points (complete the job), player B receives 3*x* points
    - In the second stage, player B can choose to give or remove (punish) up to 18 points to/from player A
        - For every point removed, player B loses ⅓ of a point
    - Two treatments:
        - Intention treatment - as described above
        - No-intention treatment - player A's action decided by random device

Intentions- *vs.* Outcome-Based Models

Falk, Fehr, and Fischbacher, 2008



- Clearly, intentions matter
  - Most players neither rewarded or punished when there were no intentions
  - Models that ignore intentions cannot be entirely accurate
- Outcomes also matter
  - Some players still rewarded or punished despite the lack of intentions
  - Models that purely use intentions cannot be entirely accurate either
- Try incorporating both?

Combining Intentions and Outcomes

Falk and Fischbacher, 2006

- Define for 2 player game (They expand to N-player in their appendix):
    - π is the material payoff function
    - *n* is the current node
    - s_i is player *i*'s action, s_i' is the action that *i* believes *j* will play, and s_i'' is the action that *i* believes that *j* believes *i* will play
        - Note this is similar to a_i, b_j, c_i in Rabin fairness
    - Define an 'intention factor' δ (0 ≤ δ ≤ 1)
        - δ = 1 means that an outcome was produced intentionally by player *j*, δ < 1 means less or no intentions
        - This value depends on if player *j* had other options
    - Define an 'outcome term' Δ_i to be the player *i*'s expected difference between their payoff and their opponents payoff
        - Positive for advantageous, negative for disadvantageous

$$\Delta(n) := \pi_i(n, s_i'', s_i') - \pi_j(n, s_i'', s_i')$$

Combining Intentions and Outcomes

Falk and Fischbacher, 2006

- Define:
  - The ‘kindness term’ φ is the product of the intention factor and the outcome term
  - f is some end node
  - $v(n,f)$ is the node following n on the path to f
  - The ‘reciprocation term’ σ represents i’s kindness to j for an action in node n

$$\sigma(n,f) := \pi_j(\nu(n,f), s_i'', s_i') - \pi_j(n, s_i'', s_i')$$

  - The ‘reciprocity parameter’ ρ_i represents i’s tendency to play reciprocally
- Update utility

$$U_i(f) = \pi_i(f) + \rho_i \sum_{\substack{n \to f \\ n \in N_i}} \varphi(n)\sigma(n,f)$$

- Falk and Fischbacher show that their model can explain game results
  - UG, GEG, DG, PGG, Prisoner’s Dilemma

Recap of Models

- Intentions-Based
    - Rabin, 1993
        - Presented first kindness function using beliefs
        - Used Geanakoplos *et al.*’s ‘psychological game’ to allow utility to depend on beliefs and define ‘fairness equilibria’
        - Relatively simple
        - Only for 2-player normal form games
        - Creates multiple and sometimes unrealistic equilibria
    - Dufwenberg and Kirchsteiger, 2004
        - Extended Rabin fairness to N-player sequential games
        - Further extended by Sebald, 2010 to allow chance plays
- Outcome-Based
    - Levine, 1998
        - Assumed players are either altruistic or spiteful using ‘coefficient of altruism’
        - Cannot explain results from Dictator Games

Recap of Models

- Outcome-Based (Cont)
  - Fehr and Schmidt, 1999
    - Assumes that players are averse to both disadvantageous and advantageous inequity
    - Sums the differences between player’s payoffs
    - Uses individual parameters α_i and β_i to allow for heterogeneity in players
    - Ignores intentions
    - Relatively simple
  - Bolton and Ockenfels, 2000
    - Similar to Fehr and Schmidt, but uses a ‘relative share’ comparison to the average payoffs
    - Assumes that players only care about their relative payoff, not the distribution across all other players
    - Fehr and Fischbacher, 2004 show that this assumption is incorrect using the ‘third-party’ dictator game
    - Also ignores intentions
- Combining Intentions and Outcomes
  - Falk and Fischbacher, 2000
    - Works for N-player extensive-form games
    - Use both an ‘outcome term’ describing the difference in outcomes and an ‘intention factor’ determining the intentions of the other player
    - Very complex

More Recent Work

- Motives (Orhun, 2015)
  - Examines reactions to kind actions that could be strategically motivated
  - Finds that players are less likely to reward kind actions in the case when that action could have been chosen strategically
  - Highlights the importance of motives as well as intentions
- Kindness through blame (Çelen *et al.*, 2017)
  - Formalizes the idea of *blame* as if an opponent's action is better or worse than what the player would do in their shoes
  - Redefines the kindness function using blame
- Hidden intentions (Friehe and Utikal, 2018)
  - Examines reactions to when players attempt to hide their intentions
  - After choosing either a kind or unkind action, allows players the option of paying to decrease the chance that their opponent will know their choice
  - Finds that hiding intentions is considered to be unkind, but not as much as overt unkind actions

Future Work

- Better comparison of existing models
  - Difficult to find extensive list of current models
  - Unsure if work exists comparing them all
- Continue to combine models
  - Varying models are all good for different reasons/in different scenarios
- Find additional factors that affect fairness
  - Factors like motives and hidden intentions are interesting, perhaps more
  - For example: Mood? Relationship with opponent (stranger or friend)? Experience?
    - I have not extensively searched for existing work on these topics

References

Abbink, K., Irlenbusch, B. and Renner, E. (2000). The moonlighting game. *Journal of Economic Behavior & Organization*, 42(2), pp.265-277.

Andersen, S., Ertaç, S., Gneezy, U., Hoffman, M. and List, J. (2011). Stakes Matter in Ultimatum Games. *American Economic Review*, 101(7), pp.3427-3439.

Berg, J., Dickhaut, J. and McCabe, K. (1995). Trust, Reciprocity, and Social History. *Games and Economic Behavior*, 10(1), pp.122-142.

Bolton, G. and Ockenfels, A. (2000). ERC: A Theory of Equity, Reciprocity, and Competition. *American Economic Review*, 90(1), pp.166-193.

Cameron, L. (1999). Raising the Stakes in the Ultimatum Game: Experimental Evidence from Indonesia. *Economic Inquiry*, 37(1), pp.47-59.

Çelen, B., Schotter, A. and Blanco, M. (2017). On blame and reciprocity: Theory and experiments. *Journal of Economic Theory*, 169, pp.62-92.

Dufwenberg, M. and Kirchsteiger, G. (2004). A theory of sequential reciprocity. *Games and Economic Behavior*, 47(2), pp.268-298.

Falk, A., Fehr, E. and Fischbacher, U. (2008). Testing theories of fairness—Intentions matter. *Games and Economic Behavior*, 62(1), pp.287-303.

Falk, A. and Fischbacher, U. (2006). A theory of reciprocity. *Games and Economic Behavior*, 54(2), pp.293-315.

Fehr, E. and Fischbacher, U. (2004). Third Party Punishment and Social Norms. *SSRN Electronic Journal*.

Fehr, E. and Gächter, S. (2000). Cooperation and Punishment in Public Goods Experiments. *American Economic Review*, 90(4), pp.980-994.

Fehr, E., Kirchsteiger, G. and Riedl, A. (1993). Does Fairness Prevent Market Clearing? An Experimental Investigation. *The Quarterly Journal of Economics*, 108(2), pp.437-459.

Fehr, E. and Schmidt, K. (2003). *Theories of fairness and reciprocity: evidence and economic applications*. Cambridge: Cambridge University Press, pp.1-56.

Fehr, E. and Schmidt, K. (1999). A Theory of Fairness, Competition, and Cooperation. *The Quarterly Journal of Economics*, 114(3), pp.817-868.

Forsythe, R., Horowitz, J., Savin, N. and Sefton, M. (1994). Fairness in Simple Bargaining Experiments. *Games and Economic Behavior*, 6(3), pp.347-369.

Friehe, T. and Utikal, V. (2018). Intentions under cover – Hiding intentions is considered unfair. *Journal of Behavioral and Experimental Economics*, 73, pp.11-21.Gaechter, S. and Falk, A. (2001). Reputation or Reciprocity? An Experimental Investigation. *CESifo*, [online] Working Paper No. 496. Available at: http://hdl.handle.net/10419/75722 [Accessed 26 Mar. 2019].

Geanakoplos, J., Pearce, D. and Stacchetti, E. (1989). Psychological games and sequential rationality. *Games and Economic Behavior*, 1(1), pp.60-79.

Güth, W., Schmittberger, R. and Schwarze, B. (1982). An experimental analysis of ultimatum bargaining. *Journal of Economic Behavior & Organization*, 3(4), pp.367-388.

Levine, D. (1998). Modeling Altruism and Spitefulness in Experiments. *Review of Economic Dynamics*, 1(3), pp.593-622.

Loewenstein, G., Thompson, L. and Bazerman, M. (1989). Social utility and decision making in interpersonal contexts. *Journal of Personality and Social Psychology*, 57(3), pp.426-441.

Orhun, A. (2019). Reciprocating to Strategic Kindness. [online] pp.1-42. Available at: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.730.6816 [Accessed 26 Mar. 2019].

Rabin, M. (1993). Incorporating Fairness into Game Theory and Economics. *American Economic Association*, [online] 83(5), pp.1281-1302. Available at: https://www.jstor.org/stable/2117561 [Accessed 26 Mar. 2019].

Roth, A., Prasnikar, V., Okuno-Fujiwara, M. and Zamir, S. (1991). Bargaining and Market Behavior in Jerusalem, Ljubljana, Pittsburgh, and Tokyo: An Experimental Study. *American Economic Review*, 81(5), pp.1068–1095.

Sebald, A. (2010). Attribution and reciprocity. *Games and Economic Behavior*, 68(1), pp.339-352.