Theory and evidence on the assessment of intention and responsibility

Determinants of reciprocal behavior

Urs Fischbacher

University of Konstanz

Thurgau Institute of Economics

Overview

- Key empirical evidence on social preferences
- Theories of social preferences
- Determinants of reciprocal behavior

Key evidence of social preferences

- Outcome oriented preferences
- Reciprocity
- Third party behaviors

Key evidence of social preferences

- Outcome oriented preferences
  - Prosocial, competitive
  - Inequity aversion
  - Measurable with social value orientation
- Reciprocity
- Third party behaviors

Key evidence of social preferences

- Outcome oriented preferences
- Reciprocity
  - Positive
    - Trust game
    - Gift exchange game
    - Conditional cooperation
  - Negative
    - Ultimatum game
    - Punishment in public goods games
- Third party behaviors

Key evidence of social preferences

- Outcome oriented preferences
- Reciprocity
- Third party behaviors
  - Indirect reciprocity (A→B; C→A; repeated)
  - Generalized reciprocity (A→B; B→C; repeated)
  - Third party punishment (and reward) in one-shot situations

Facts (largely) consistent with the Standard Prediction

- High offers in ultimatum game
- Generosity in indirect reciprocity – except at the end
- Trust in the trust game

Theories of social preferences

- Bounded rationality
- Outcome oriented models
- Reciprocity models

Outcome Oriented Models

- $U_i = U_i(\pi_i, \pi_{-i})$
- Utility differs from payoff (in experiment)

- How does U_i depend on π_{-i}
- Share: $\pi_i/\Sigma\pi_j$ (BO)
- all differences: $\pi_i-\pi_j$(FS)
- Efficiency: $\Sigma\pi_j$; maximin: $\min(\pi_j)$ (CR)

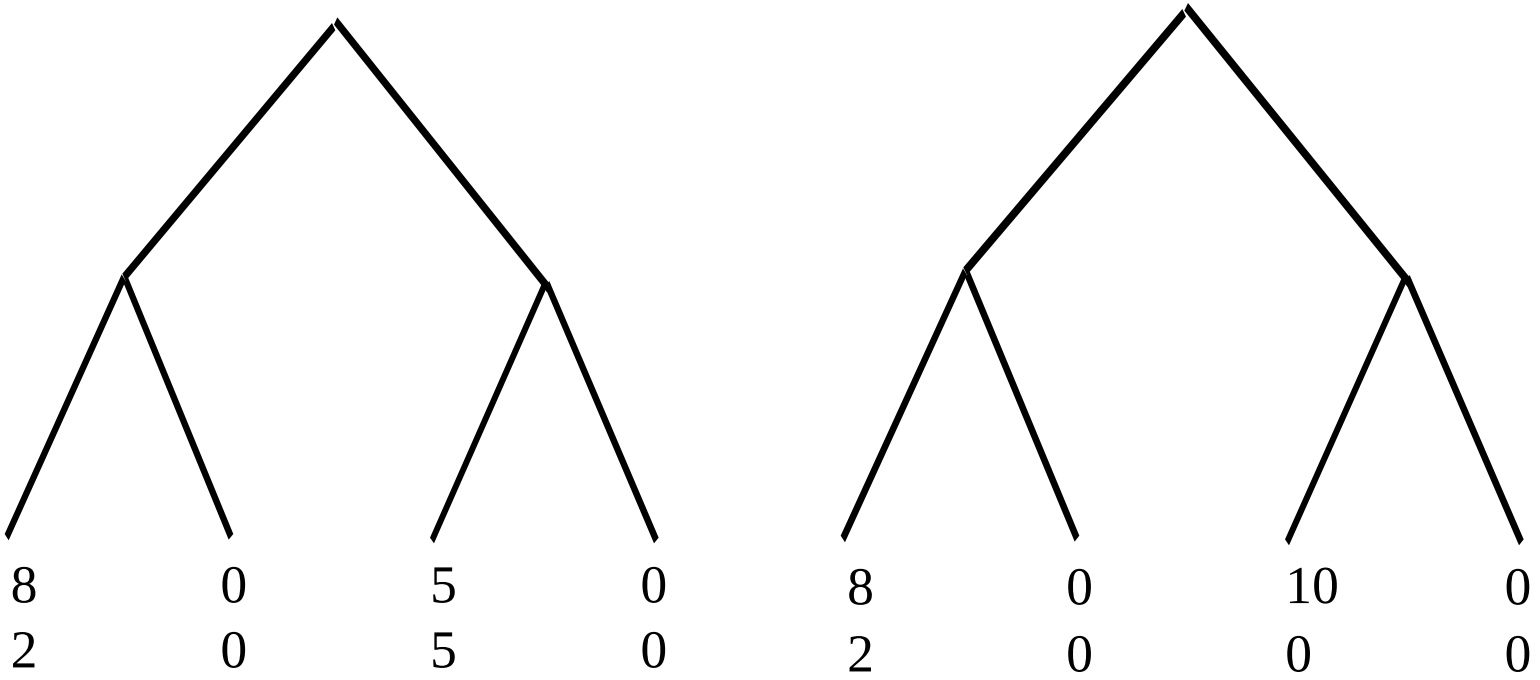
- Positive: altruism
- Negative: status seeking
- Hump-shaped (around 1/n or 0): inequity aversion (BO, FS)

The strength of inequity aversion models

- Theory can explain
  - Fair offers in ultimatum game
  - Unfair offers in competition

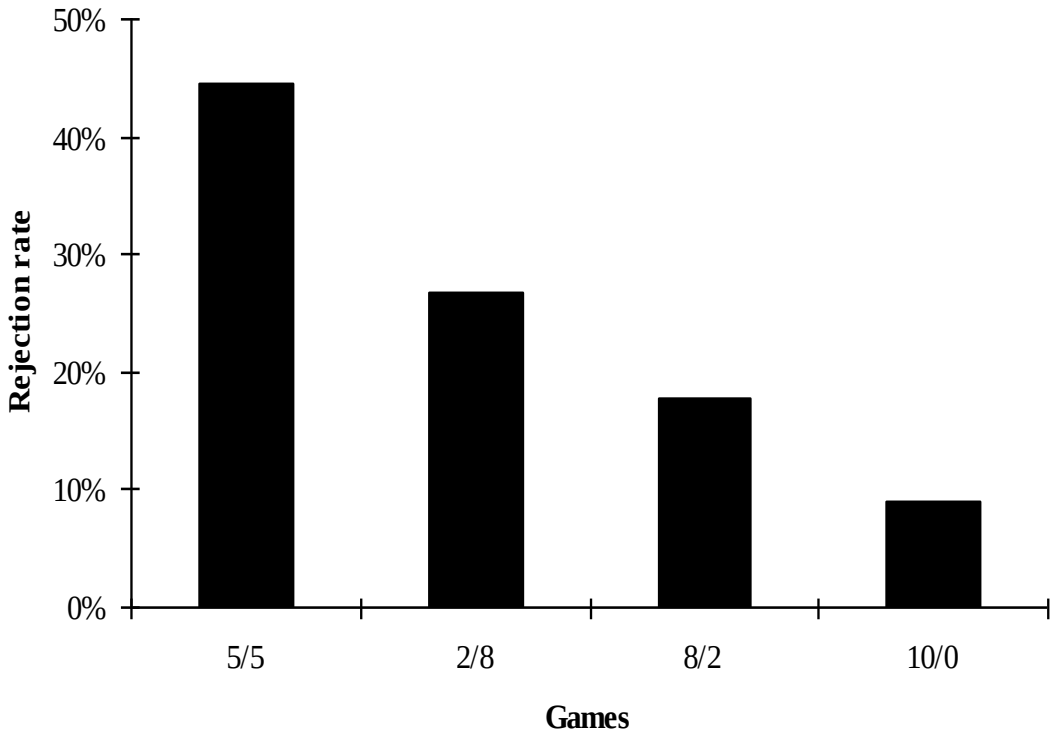
  - … with the same distribution of parameters.

A limit of outcome oriented theories



Q3: Experimental Results (N=45)

Figure 2
Rejection rate of the (8/2)-offer across games



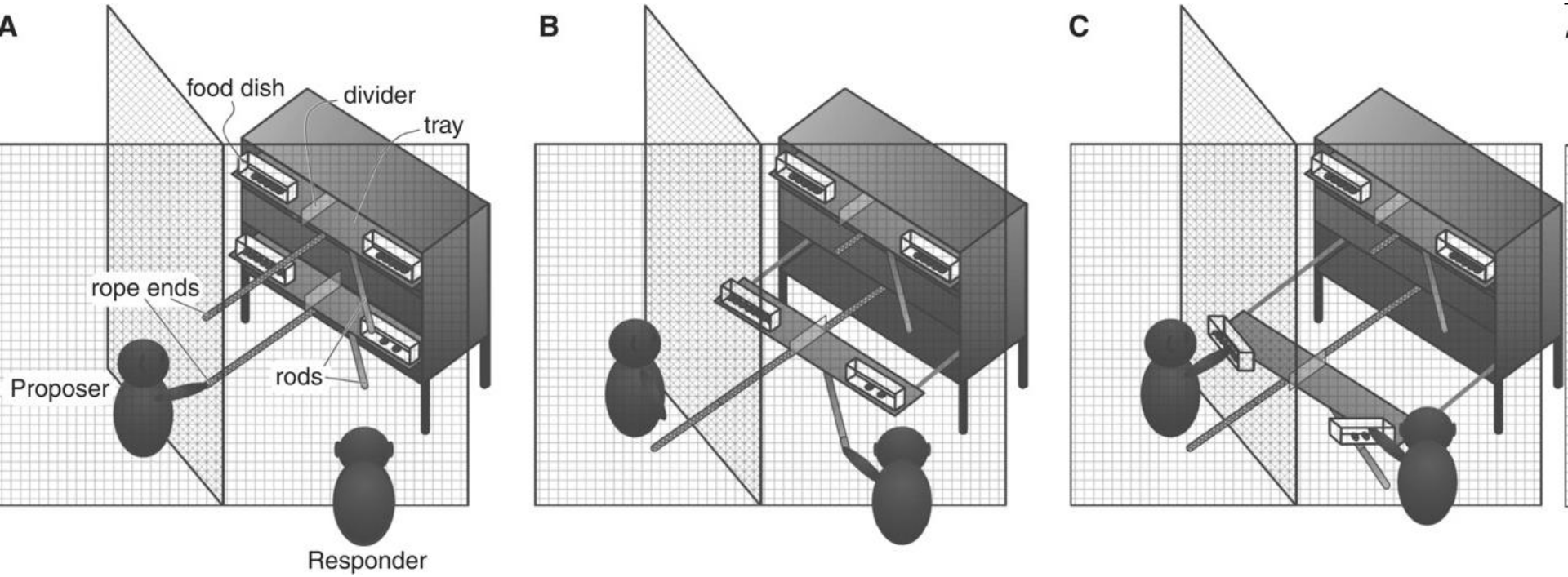
- Theories that are based only on outcomes cannot explain the different rejection rates.

Q3: Proposer Behavior

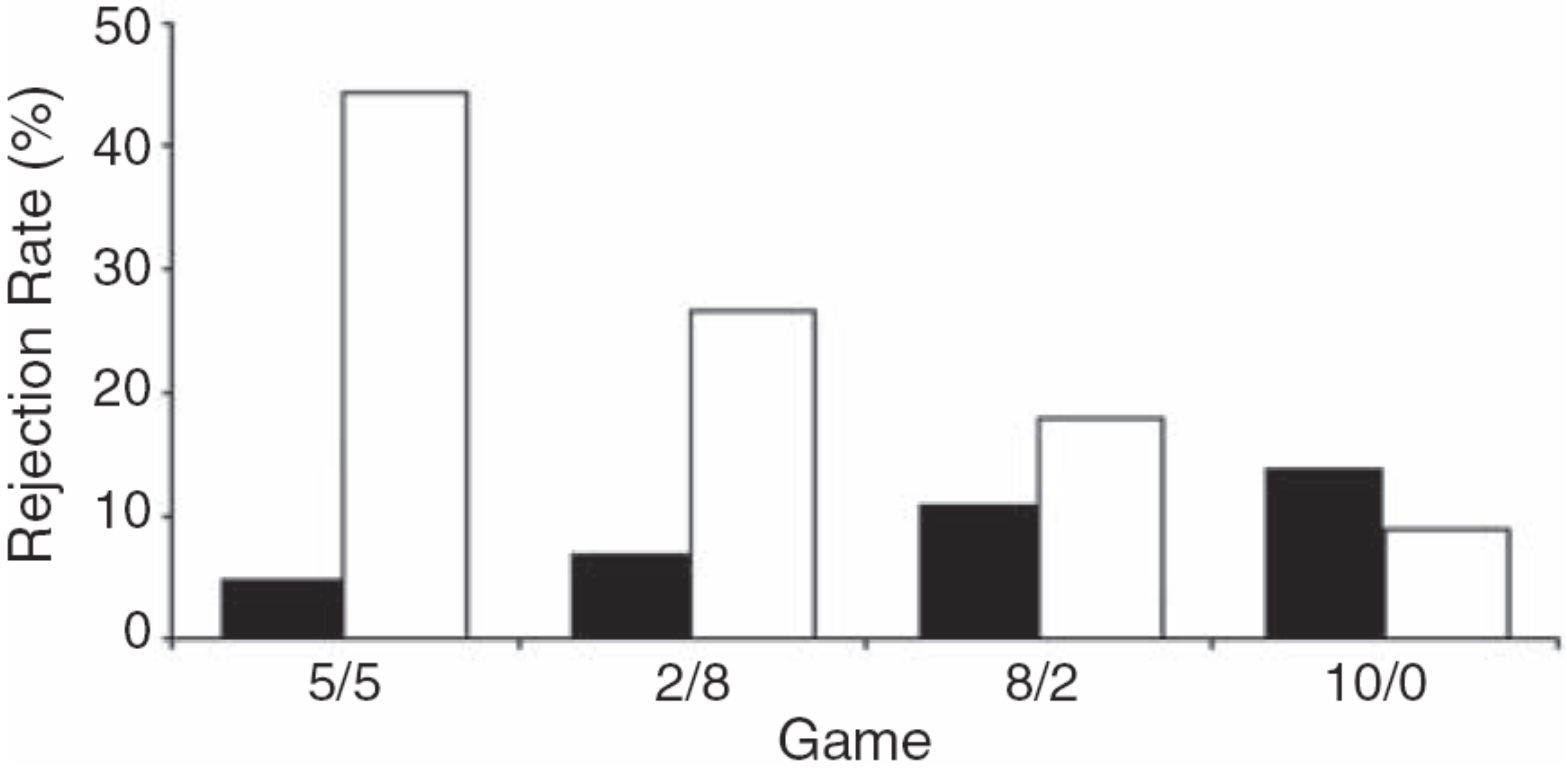
Alternative	Rejection rate of 8/2	Rejection rate of alternative	Choice of 8/2
5/5	44%	0%	31%
2/8	27%	2%	73%
8/2	16%	20%	-
10/0	9%	89%	100%

- Proposer behavior is compatible with selfishness, but also with preferences for fairness.

Do Chimpanzees care about intentions Jensen, Call, Tomasello (2007)



Jensen, Call, Tomasello (2007)



- Black bars: Chimpanzees, White bars: humans

Non-Outcome oriented theories

- Reciprocity
  - Reward kind behavior
  - Punish unkind behavior
- What is kind or unkind?
  - Outcome (inequity aversion)
  - Intention
    - Alternative based (Rabin, 1993)
    - Type based (Levine, 1998)
  - Accountability (Konow, 2003)

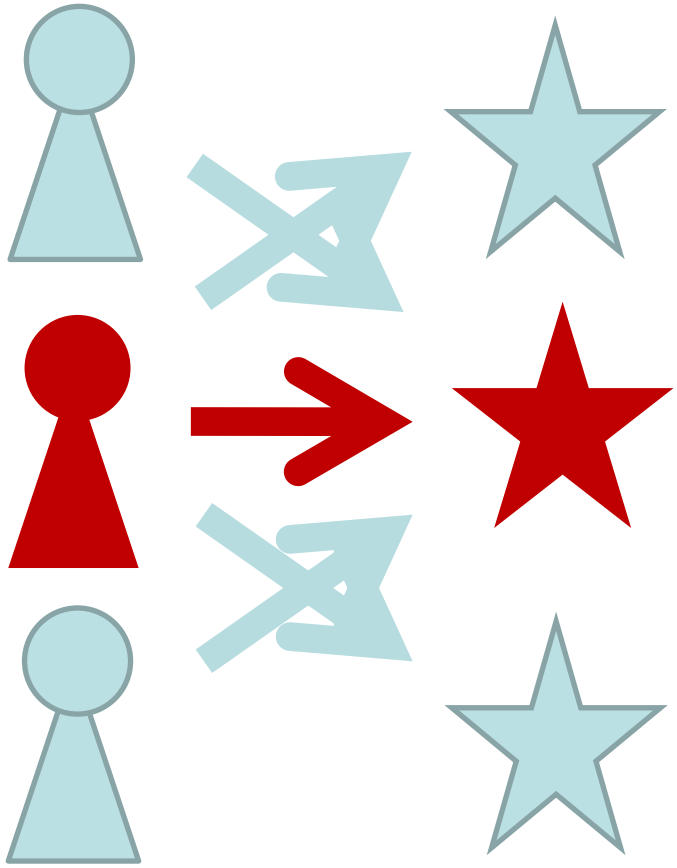
What is blameworthy?

- Definition:
  "To be morally responsible for something, say an action, is to be worthy of a particular kind of reaction—praise, blame, or something akin to these—for having performed it."
  [Stanford Encyclopedia of Philosophy]

- The analysis of reward and punishment patterns allows the investigation of moral judgments.
  - Public good games (Ostrom, 1992, Yamagishi, 1986; Fehr and Gächter, 2000)
  - Ultimatum game (Güth et al. 1982)
  - Third party punishment (Fehr, Fischbacher 2003)

Approaches to responsibility

- Is the result of an action intended by the person?
  - …or did the person have another intention.
  - Intention for side effects.
- Is the person causally responsible for the result?
  - …or did other persons or factors cause the result?
  - Responsibility in decision chains.



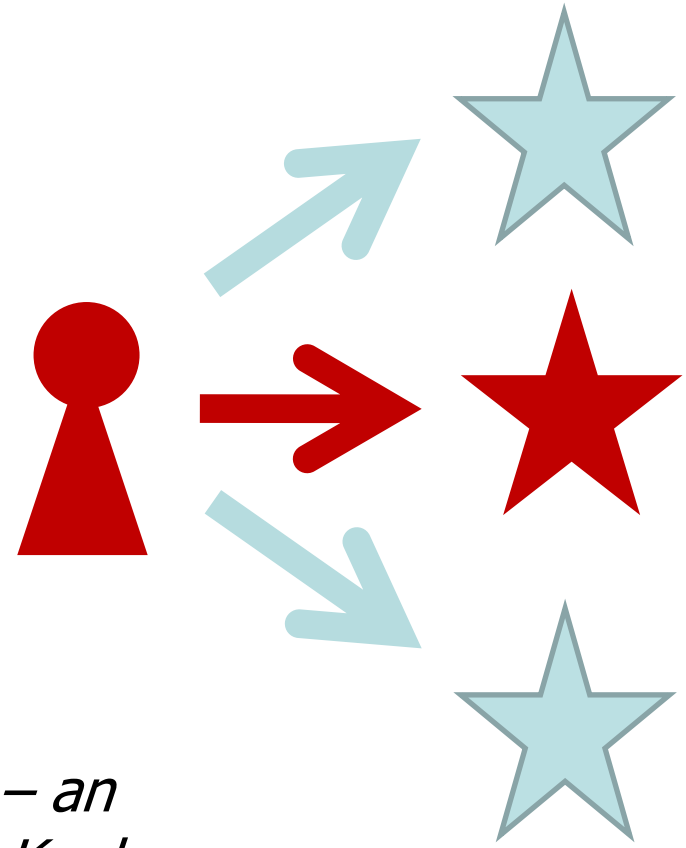
Goal of the talk

- Link theoretical constructs of intention and responsibility with actual behavior.

- Part I: Intention
  - Folk concepts of intention
  - Economic concept of intention
  - Reward and punishing behavior
- Part II: Responsibility
  - Theoretical measure of causal responsibility
  - Delegation and punishing behavior

Intention

- Folk concepts of intention
- Economic concept of intention
- Punishing behavior



Attribution of Externalities – an Economic Approach to the Knobe Effect

With Verena Utikal

A Story

The vice-president of a company went to the chairman of the board and said, 'We are thinking of starting a new program. It will help us increase profits, but it will also harm the environment.' The chairman of the board answered, 'I don't care at all about harming the environment. I just want to make as much profit as I can. Let's start the new program.'

So the company started the new program, increased profits and the environment was harmed.

Now ask yourself:

Did the chairman of the board *intentionally* harm the environment?

Another story

The vice-president of a company went to the chairman of the board and said, 'We are thinking of starting a new program. It will help us increase profits, and it will also help the environment.' The chairman of the board answered, 'I don't care at all about helping the environment. I just want to make as much profit as I can. Let's start the new program.'

So the company started the new program, increased profits and the environment was helped.

Now ask yourself:

Did the chairman of the board *intentionally* help the environment?

Answers to the Questions

N=180		help		
		intentionally	not intentionally	
harm	intentionally	16%	51%	67%
	not intentionally	3%	30%	33%
		19%	81%	

People are willing to blame other people for foreseen side-effects that are bad, but are not willing to praise them for foreseen side-effects that are good. (Knobe 2003)

Philosophical interpretations of this result

- Morally good and bad behavior (Knobe, 2006)
- Trade-off hypothesis (Machery, 2008)
- Relation to responsibility in the sense of deserving blame or praise (Wright and Bengson, 2009)

Trade-off hypothesis
Machery (2008)

- Joe was feeling quite dehydrated, so he stopped by the local smoothie shop to buy the largest sized drink available. Before ordering, the cashier told him that if he bought a Mega-Sized Smoothie he would get it in a special … cup. Joe replied, 'I don't care about a … cup, I just want the biggest smoothie you have.' Sure enough, Joe received the Mega-Sized Smoothie in a commemorative cup. — 45%
- Joe was feeling quite dehydrated, so he stopped by the local smoothie shop to buy the largest sized drink available. Before ordering, the cashier told him that the Mega-Sized Smoothies were now one dollar more than they used to be. Joe replied, 'I don't care if I have to pay one dollar more, I just want the biggest smoothie you have.' Sure enough, Joe received the Mega-Sized Smoothie and paid one dollar more for it. — 95%

Our questions with regard to the Knobe effect

- Is there an asymmetry in the assignment of praise and blame for positive and negative externalities?

  ➔ Experiment

- What theoretical concept of intention captures…
  - …the assignment of blame and praise.
  - …the attribution of intention in the Knobe questionnaire.

  ➔ Theory

Experimental approach

- Create the same setting as in the questionnaire.
  - Let third parties reward and punish.
- Vary the setting and investigate changes in the punishment pattern.
- Confirm the intention attribution as suggested by the punishment pattern with new vignettes.

Perception of the stories

Perception I

- Powerful firm
- Weak environment
- Strong negative externalities
- Weak positive externalities

Perception I - Experimental Design

- Player 1: firm (active agent)
- Player 2: environment (passive agent)
- Player 3: reader of the story, redistributor

HURT	X	Y
firm	50	
environment	50	

Player 3 is redistributor; can transfer points between firm player and environment player

Perception I - Experimental Design II

HURT	X	Y
firm	50	60
environment	50	30

Perception I

- Powerful firm
- Weak environment
- Strong negative externalities
- Weak positive externalities

HELP	X	Y
firm	50	60
environment	20	30

No Externalities	Y
firm	60
environment	30

Implementation of reward and punishment

- Reward and punishment are measured relative to the no externality treatment.
- Reward is equivalent to withdrawn punishment and vice versa.
- Reward and punishment are comparable.

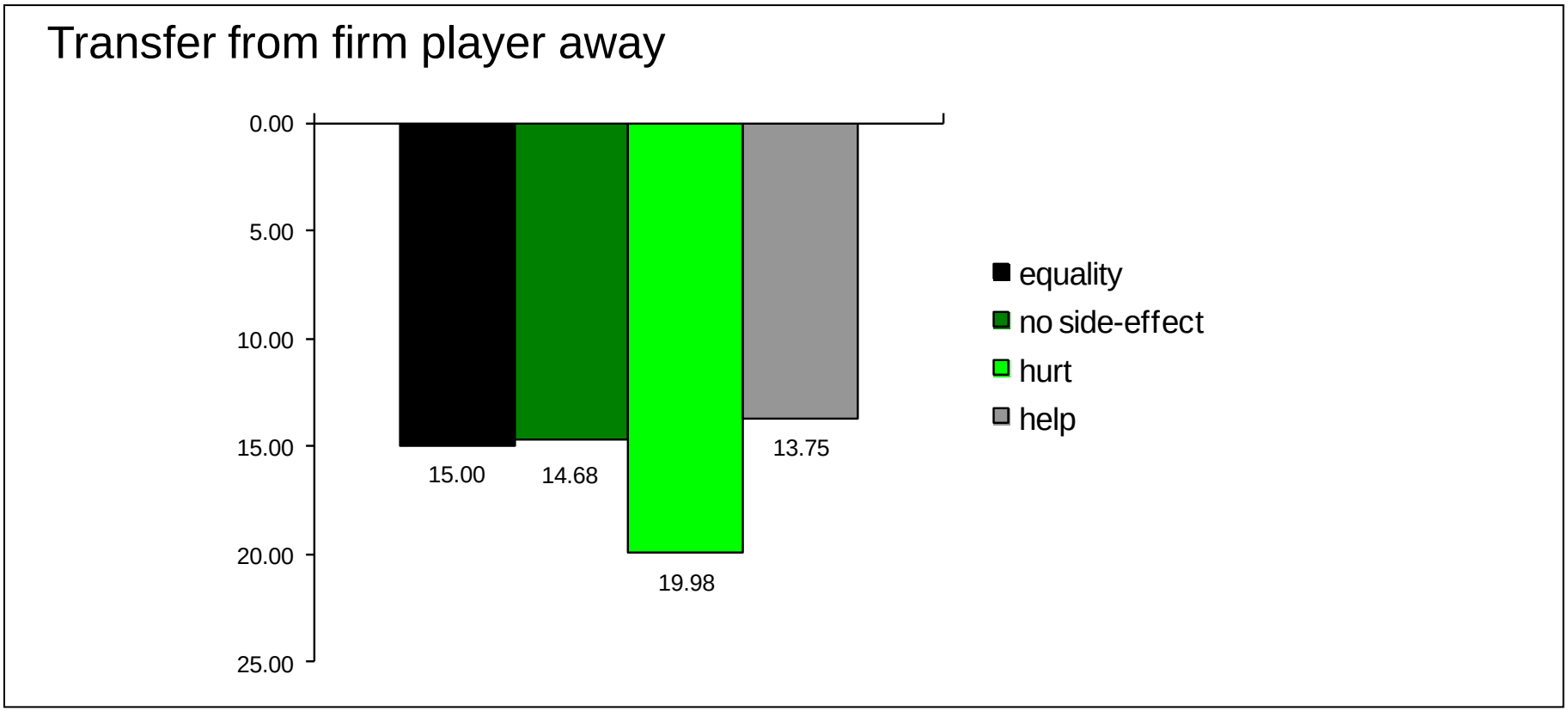
Procedures

- Experiment took about 30 minutes
- 180 participants (60 redistributors)

- For each game player 1 and 2 received the payoff from the allocation +/- points transferred by the redistributor
- Player 3 received 100 points for every period
- Transfer is costless
- 100 points =1 Euro (average earnings 3.77 Euro)

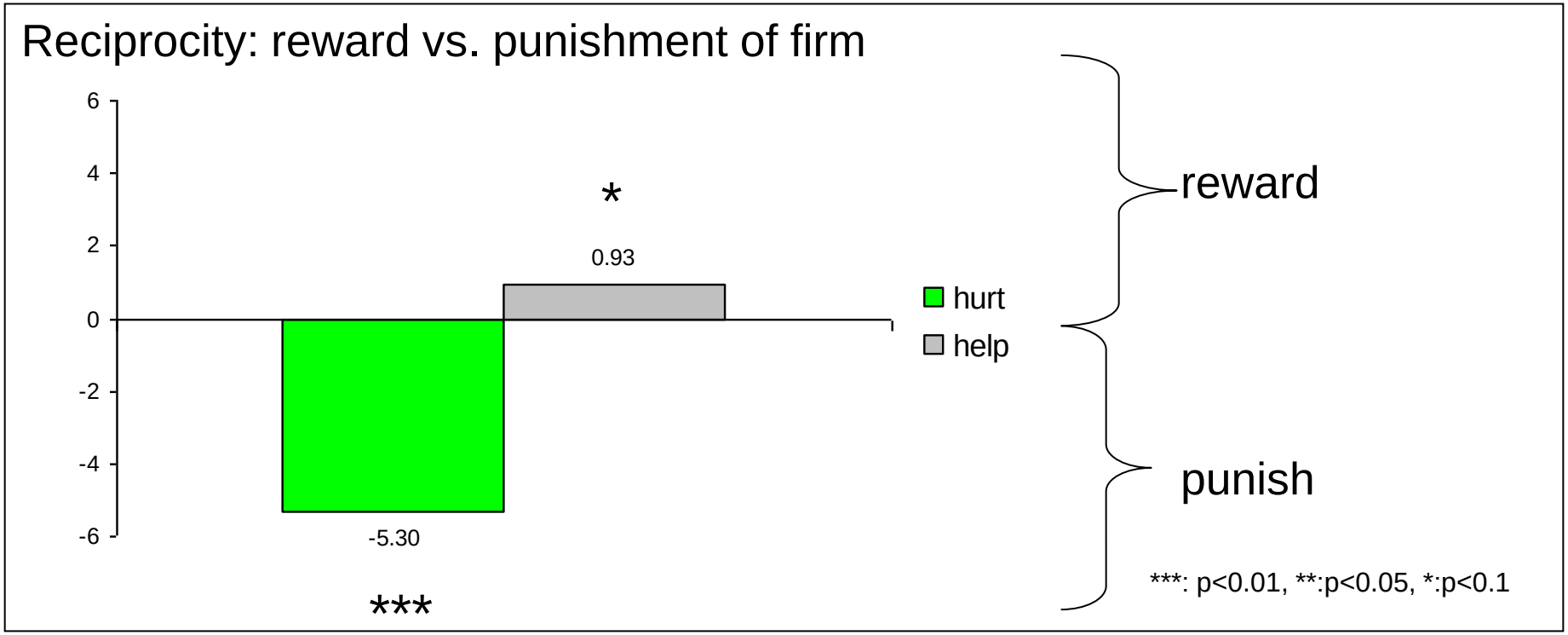
- Subjects did not receive feedback until the end of the experiment
- Randomized order
- For the redistributor, we used the strategy method

Perception I - Results (when Y is chosen)



		X (hurt)	Y	X (help)
Perception I	firm	50	60	50
	environment	50	30	20

Perception I - Results (when Y is chosen)



		X (hurt)	Y	X (help)
Perception I	firm	50	60	50
	environment	50	30	20

Perception II/III

In Perception I…

- Firm possesses higher economic status
- Strong negative externalities
- Weak positive externalities

What about other perceptions? For example…

- Environment possesses higher economic status
- Strong positive externalites

Perception II – Experimental Design

HURT	X	Y
firm		30
environment		60

Perception II

- Weak firm
- Powerful environment
- Strong negative externalities
- Weak positive externalities

HELP	X	Y
firm	20	30
environment	50	60

No Externalities	Y
firm	30
environment	60

Perception III – Experimental Design

HURT	X	Y
firm	20	30
environment	80	60

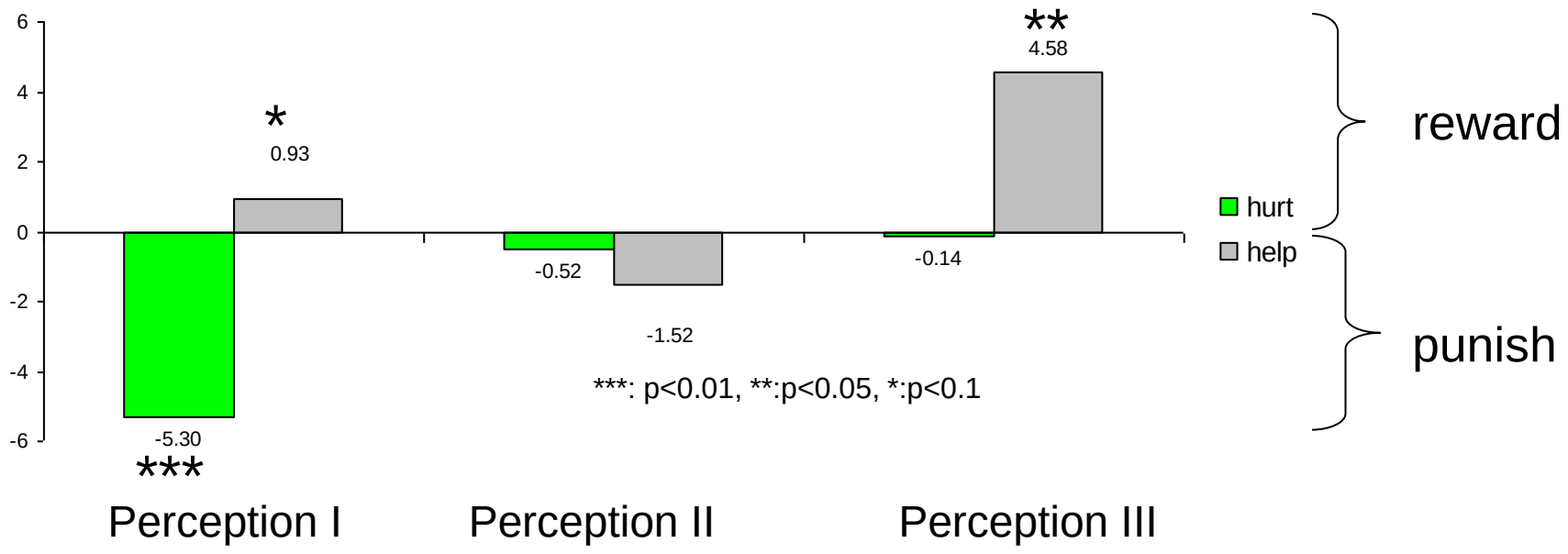
Perception III

- Weak firm
- Powerful environment
- Strong negative externalities
- Strong positive externalities

HELP	X	Y
firm	20	30
environment	20	60

No Externalities	Y
firm	30
environment	60

Reciprocity: Transfer to firm in Perceptions I-III



		X (hurt)	Y	X (help)
Perception I	firm	50	60	50
	environment	50	30	20
Perception II	firm	20	30	20
	environment	80	60	50
Perception III	firm	20	30	20
	environment	80	60	20

Are these results relevant for the Knobe effect?

- Knobe effect is about intentionality, not about blame and praise.
- New Knobe questions with our new perception:
  - Strong environment

A Story II

The vice-president of a small fast-food restaurant went to the chairman of the board and said, 'We are thinking of launching a new burger. It will help us increase profits, but it will also harm McDonalds next door.' The chairman of the board answered, 'I don't care at all about harming McDonalds. I just want to make as much profit as I can. Let's launch the new burger.'

So the company launched the new burger, increased profits and McDonalds next door was harmed.

Now ask yourself:

Did the chairman of the board *intentionally* harm McDonalds?

Another Story II

The vice-president of a small fast-food restaurant went to the chairman of the board and said, 'We are thinking of launching a new burger. It will help us increase profits, but it will also help McDonalds next door (for example due to higher pedestrian flow).' The chairman of the board answered, 'I don't care at all about helping McDonalds. I just want to make as much profit as I can. Let's launch the new burger.'

So the company launched the new burger, increased profits and McDonalds next door was helped.

Now ask yourself:

Did the chairman of the board *intentionally* help McDonalds?

Answers to the Questions

N=87		help		
		intentionally	not intentionally	
harm	intentionally	9%	18%	27%
	Not intentionally	6%	67%	73%
		15%	85%	

N=180		help		
		intentionally	not intentionally	
harm	intentionally	16%	51%	67%
	Not intentionally	3%	30%	33%
		19%	81%	

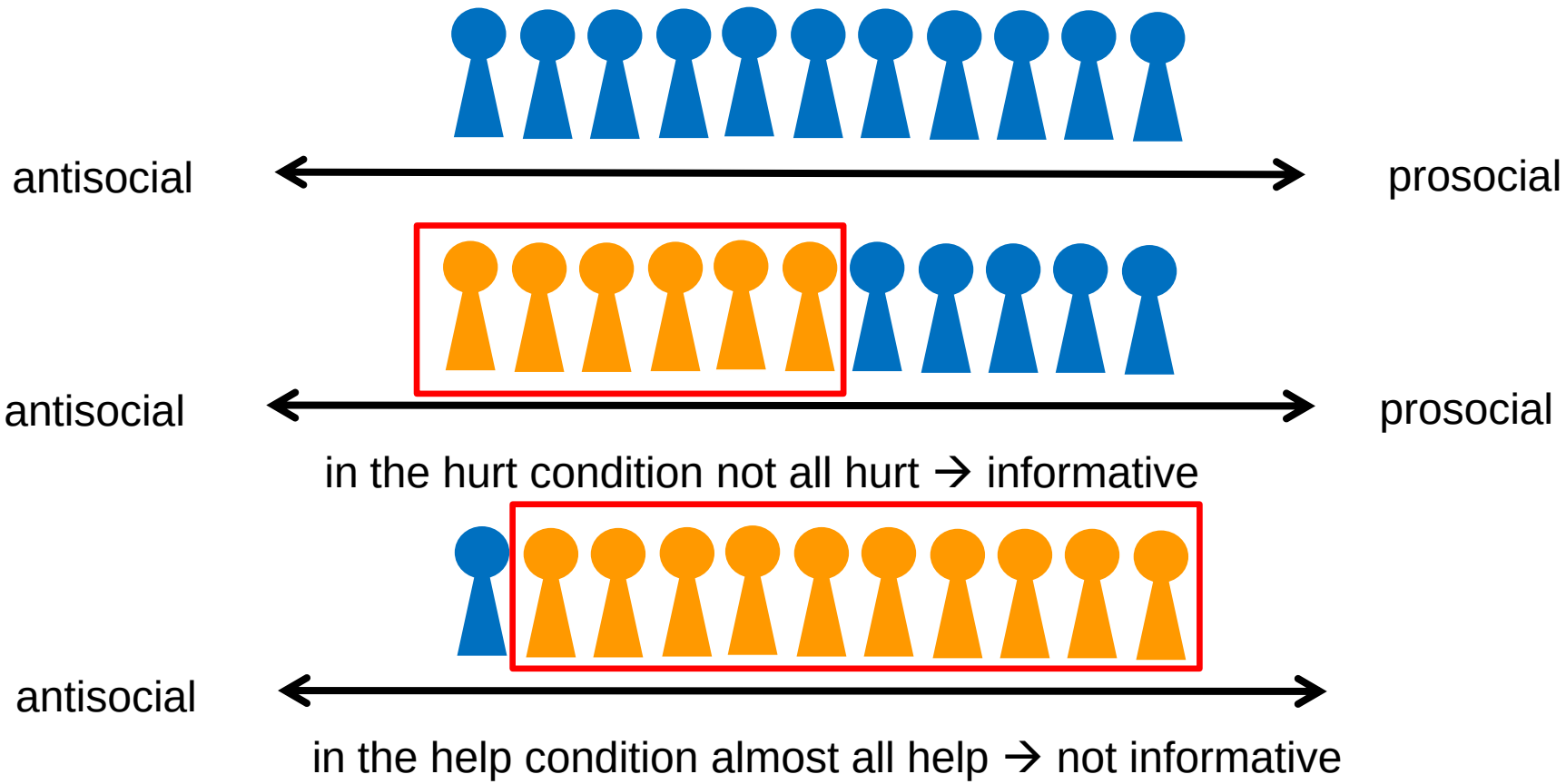
The Knobe effect and concepts of intention

- Is the action intentional?
  - Always in the context of a decision experiment.
  - This is not interesting.
- Is the side effect intentional?
  - If the side effect was a goal of the agent.
  - Is the side effect is motivated – part of the utility function.
  - If side effect is an externality:
    - Intentionally kind ⇔ kindness is intentional
    - Like in Levine (1998) model.

The Levine (1998) model

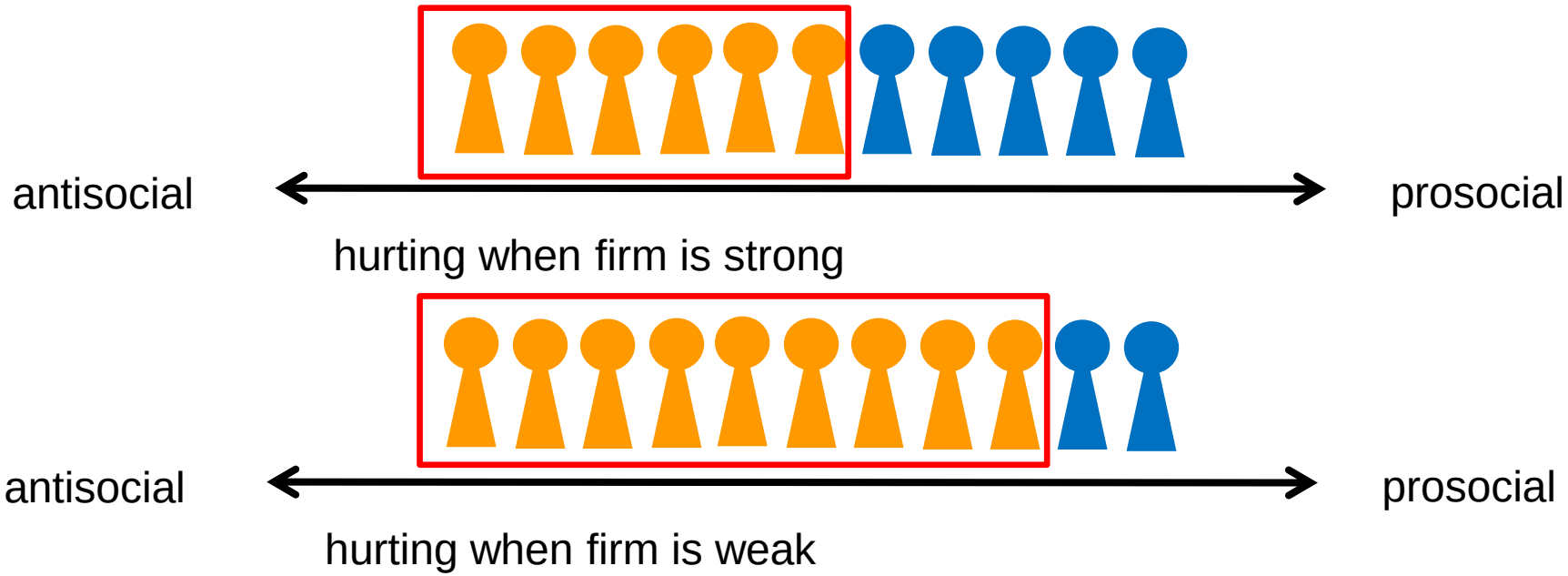
- People vary in how much they care about the other's payoff: a_i.
- This concern can be considered as the player's intention.
- People care about the other's intention, i.e., a_j.

The Levine Model - Illustration



The Levine (1998) model

- Robust implication:
  - The more people that cause an unfriendly outcome, the less unfriendly they are on average.
  - The more people that cause a friendly outcome, the less friendly they are on average.



The Levine model in the mini ultimatum games

Alternative	Choice of 8/2	Unkindness of 8/2
5/5	31%	high
2/8	73%	lower
8/2	-	low
10/0	100%	kind

Frequency of Player 1 choosing Y

	hurt	help
Perception I	53%	83%
Perception II	76%	92%
Perception III	83%	77%

- Levine-type reciprocity
  - Hurting is "special" in Perception I?
    - Only few hurt.
  - Helping is "special" in Perception III?
    - Only relatively few help.

The Levine style definition of intention and philosophical approaches

- Morally good and bad behavior (Knobe, 2006).
  - If an action is "moral", not everybody does it.
- Trade-off hypothesis (Machery, 2008).
  - If there is no trade-off, everybody does it.
  - Shows that intention for X can be defined for contexts outside of social preferences.
- Related to responsibility in the sense of deserving blame or praise (Wright and Bengson, 2009).
  - Attribution of intention determines blame or praise.

Conclusion

- Knobe vignette studies replicated in a language-independent way.
- Knobe effect with respect to blame and praise depends on economic determinants of the situation.
  - Economic status.
  - Size of positive externalities.

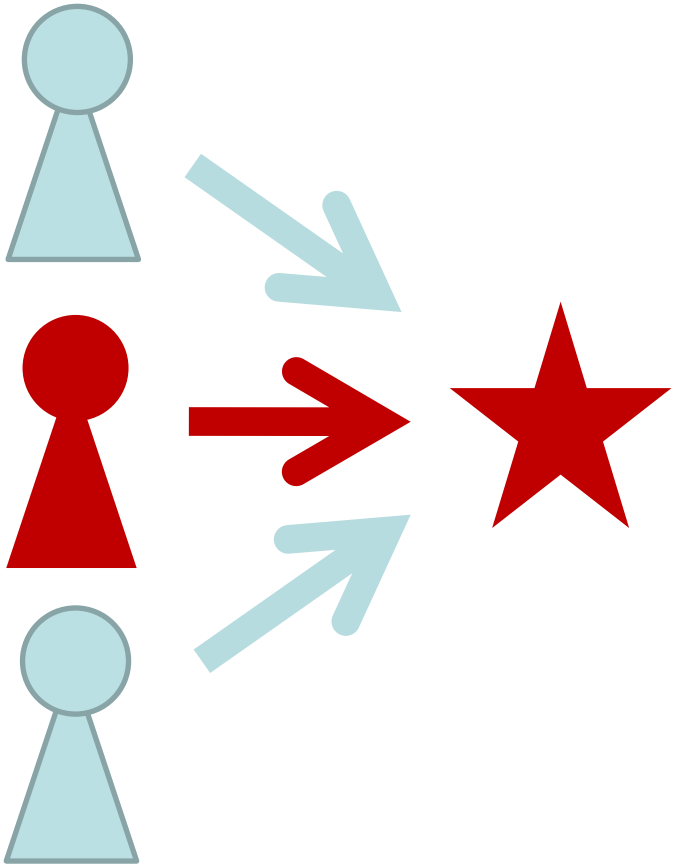
- The intention concept of Levine is consistent with the way people attribute intention in Knobe questionnaires.

Responsibility

- Responsibility in delegated decisions
- A measure of causal responsibility

Shifting the Blame: On Delegation and Responsibility

with Björn Bartling



Who is responsible for the disaster of Costa Concordia?

- Manufacturer of the ship?
  - Ship should not tilt.
- Shipping line?
  - Missing regulation and bad governance.
  - Incompetent captain.
- Captain?
  - Neglect of sea charts.
  - …

Picture: Tagesanzeiger; Bild:http://www.lexpress.fr/24henimage/costa-concordia-l-enquete_1078440.html

Introduction

- Research Question:
  - Who is held responsible for the outcome of a delegated decision, the person who delegated the decision right or the person who ultimately took the decision?
  - Does this affect how people delegate?
  - How can the concept of responsibility be formalized?

Introduction

- In political science the motive of blame shifting is much discussed since **Machiavelli** wrote:

  "Princes should delegate to others the enactment of unpopular measures and keep in their own hands the distribution of favours."

- Example: **chief restructuring officers** (CROs)

  CRO temporarily replaces CEO of a firm in financial distress.

  CROs bring specific expertise and experience, an outside perspective, and supplement incumbent managers in times of intensive work load.

  However, blame shifting might also play a role:

  "Change frequently requires difficult choices and unpopular decisions. The use of an interim executive to move through these decisions and then move on, allows new, permanent leadership to take the helm untainted by any residual negative feelings toward his or her predecessor"

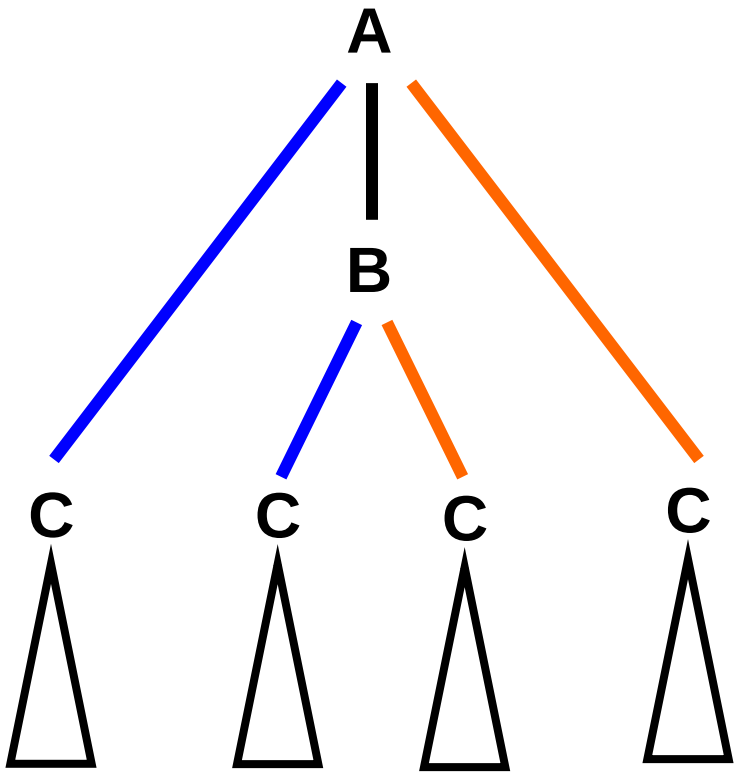
  [McShane Group, "Turnaround Consulting & Crisis Management"]

Overview

- We use punishment as a measure of responsibility attribution.
- We show that
  - responsibility can be shifted and that
  - responsibility shirking can be an important motive to delegate a decision right.
- We propose a simple, formal measure of responsibility and conduct an econometric comparison of different punishment motives (intentions, outcomes, responsibility)

Experimental Design

	Player A	Player B	Player C	Player C
fair allocation	**5**	**5**	**5**	**5**
unfair allocation	**9**	**9**	**1**	**1**



Player A can decide between the fair and the unfair distribution.

In treatments with delegation he can delegate the decision to player B. Then player B decides between the fair and the unfair distribution.

In treatments with punishment, one randomly chosen player C can assign costly punishment points to any other player; up to 7 points at the cost of 1 point.

Procedural Details

- All treatments were played one-shot.
- Strategy method for player C: all players C decide for all possible situations whether and how they want to punish.
- 824 subjects in 24 sessions. One belief elicitation session with 32 subjects. Subjects were students from Zurich, recruited with ORSEE (Greiner, 2004)
- Sessions without punishment lasted for about 45 minutes; sessions with punishment lasted for about 60 minutes.
- Programmed with z-Tree (Fischbacher, 2007)
- On average, subjects earned CHF 25 in sessions without punishment and CHF 23.25 in sessions with punishment, which includes a show-up fee of CHF 10.

Standard Prediction

Rationality and selfishness:

C

- Does not punish.

B

- B is unfair, unless B fears punishment.
- In particular, B is unfair if there is no punishment.

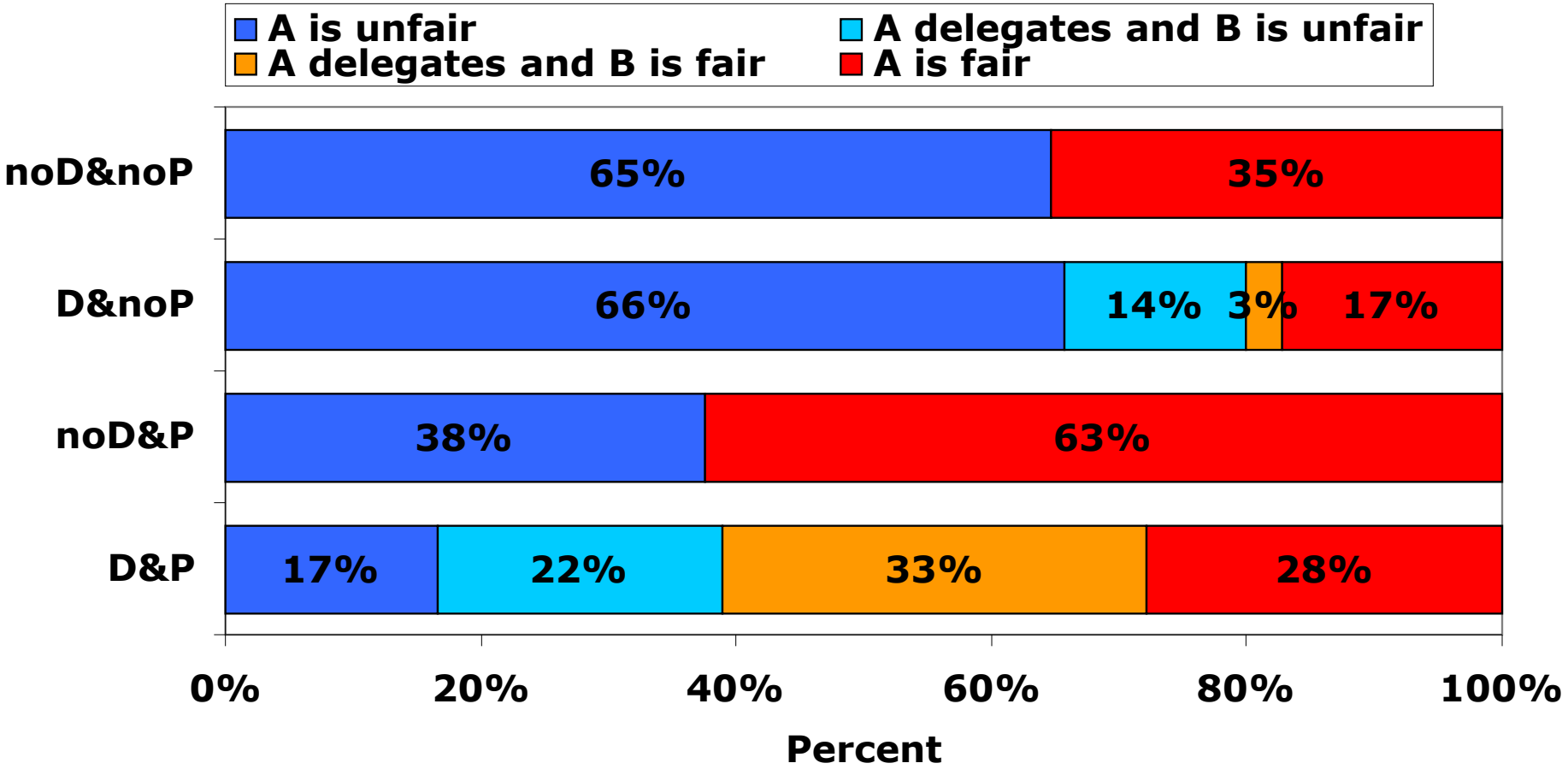
A

- A is unfair, unless A fears punishment.
- In particular, A is unfair if there is no punishment.
- Does not delegate in the no-punishment condition.

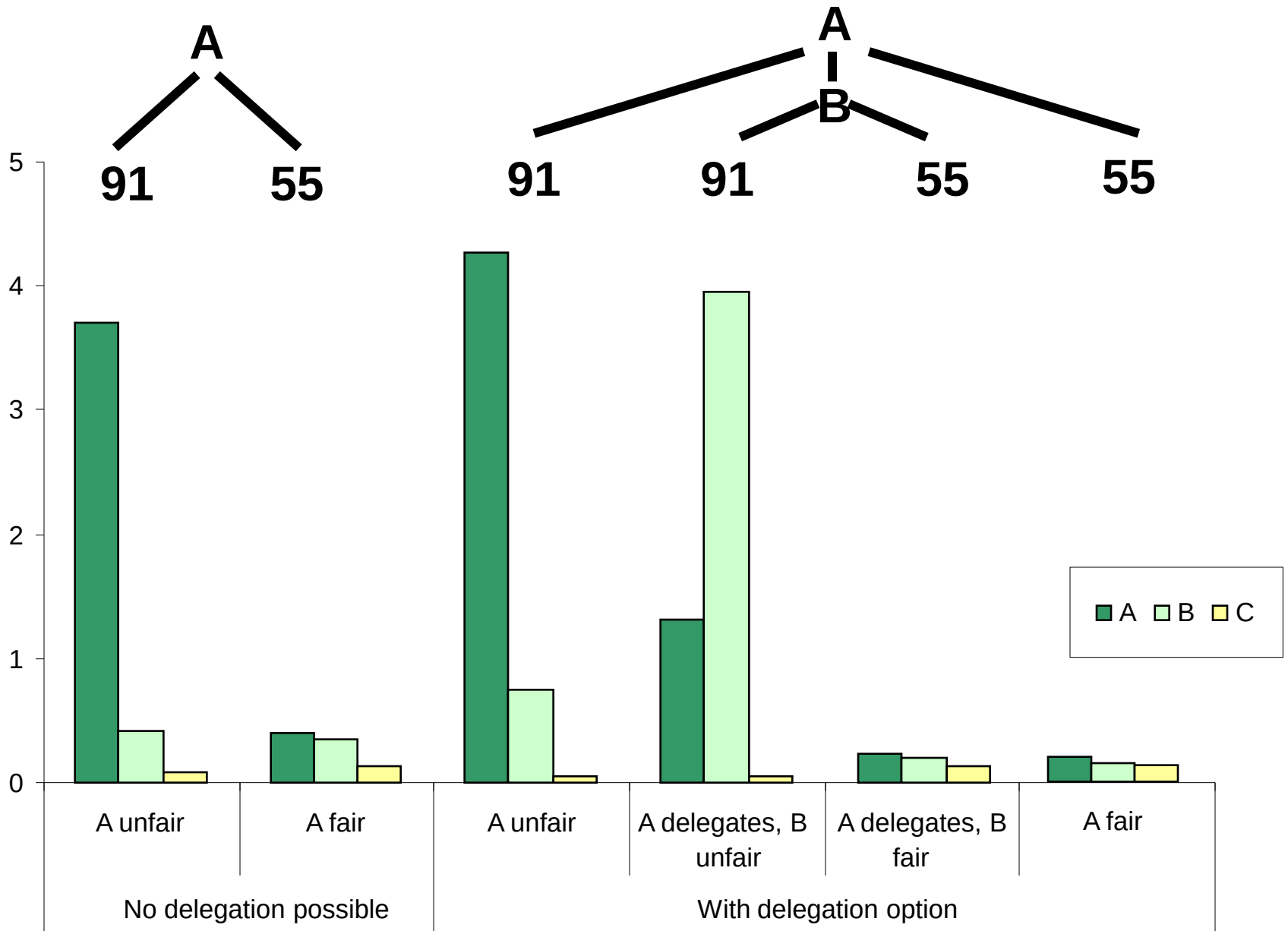
Counter Intuitions

- Some people are fair.
  - Dictator game (Forsythe et al. 1994).
- People do not want to behave unfairly, even when they would like an unfair outcome.
  - Exit in dictator game (Dana, Cain, Daws, 2005).
- People punish if they are treated unfairly.
  - Güth et al. (1982), Fehr and Gächter (2000).
  - The punishment pattern informs us about attributed responsibility.

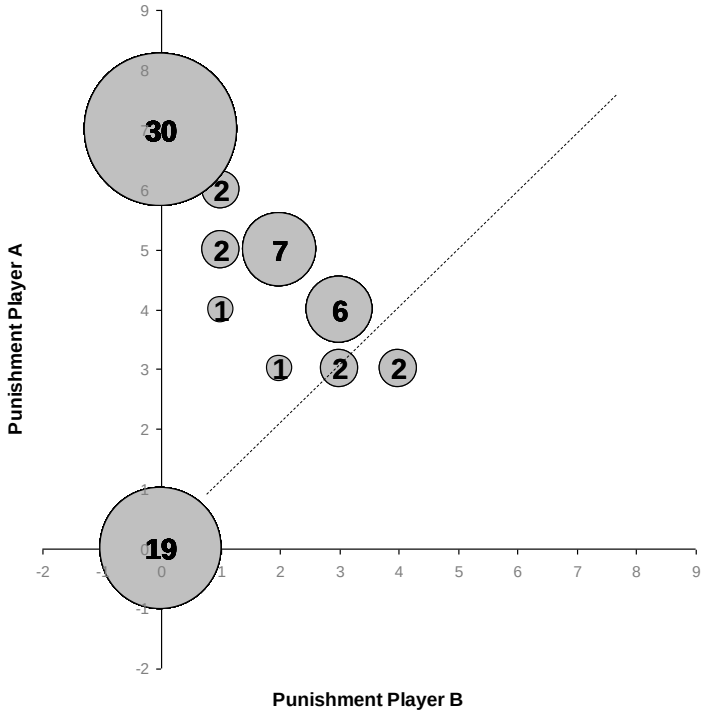
Decisions by Players A and B

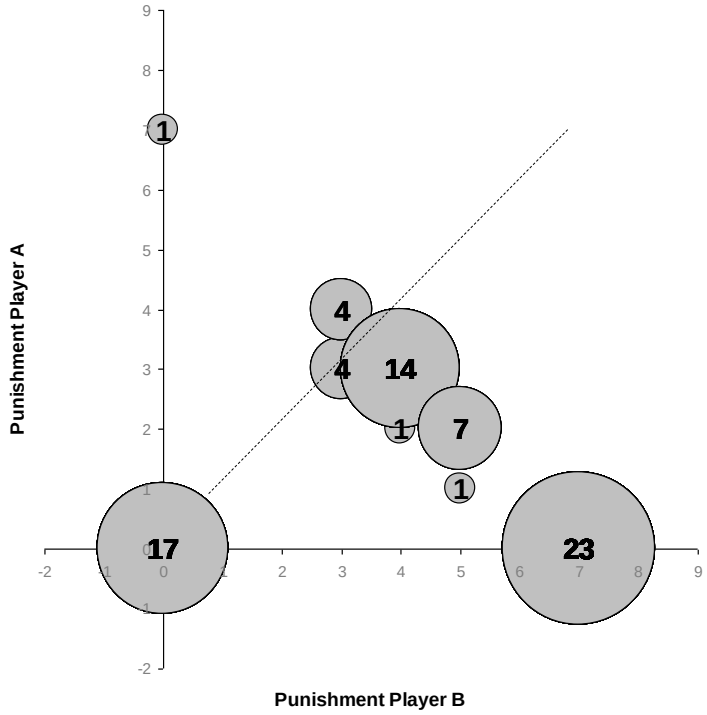


Average Punishment Points

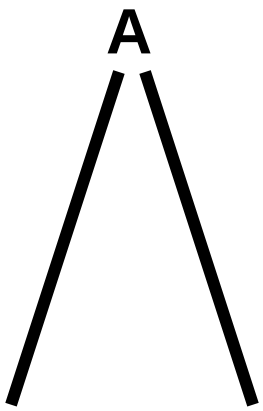


Punishment in Treatment D&P





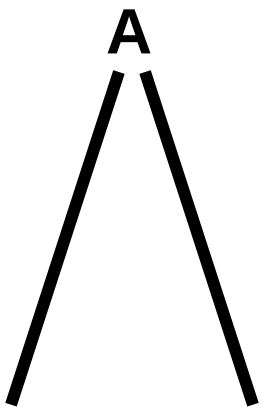
Motives for Punishment



		unfair	fair
Outcome	A	H	-
	B	H	-

Desire for punishment: H= High; M= medium: - = no

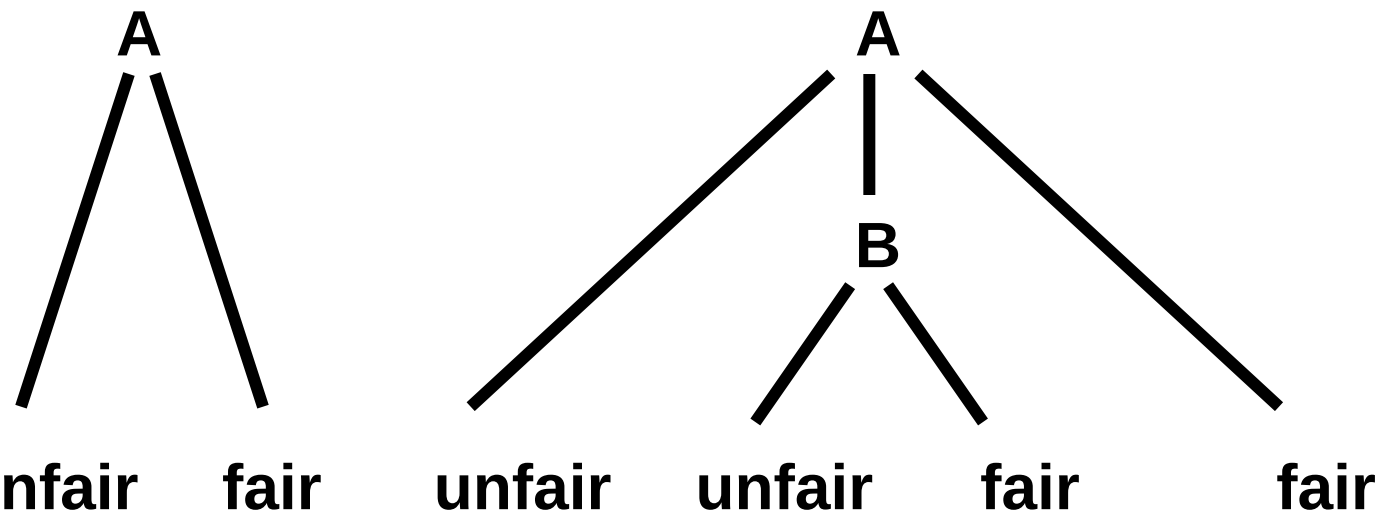
Motives for Punishment



		unfair	fair
Outcome	A	**H**	-
	B	**H**	-
Intention	A	**H**	-
	B	-	-

Desire for punishment: H= High; M= medium: - = no

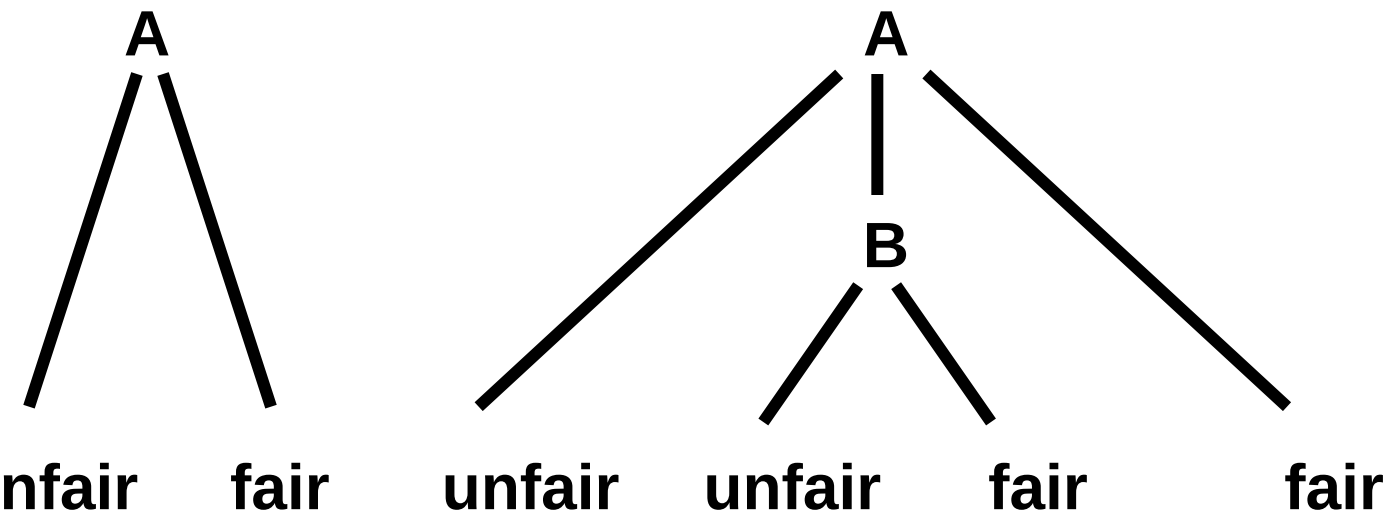
Motives for Punishment



Outcome	A	**H**	**-**	**H**	**H**	**-**	**-**
	B	**H**	**-**	**H**	**H**	**-**	**-**
Intention	A	**H**	**-**				
	B	**-**	**-**				

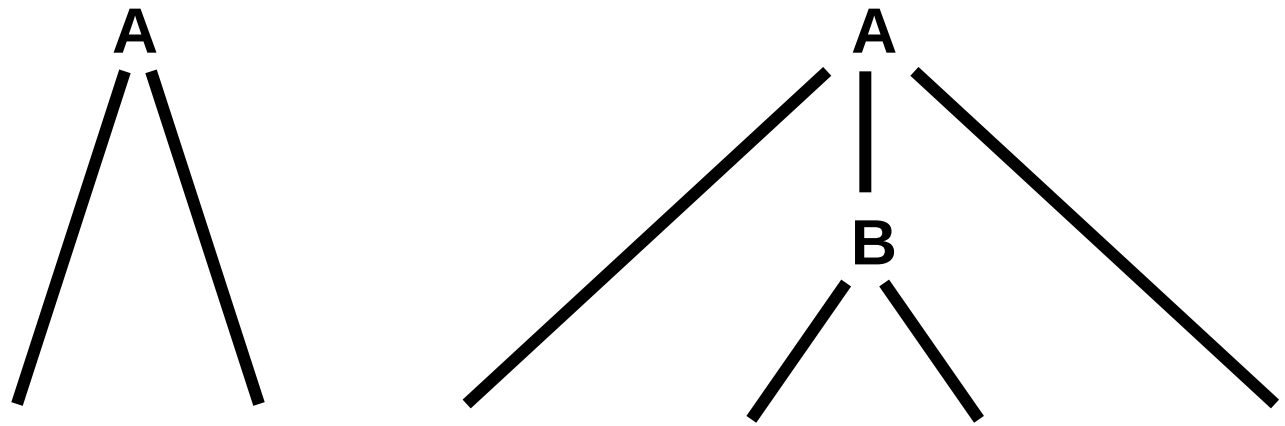
Desire for punishment: H= High; M= medium: - = no

Motives for Punishment



Outcome	A	**H**	**-**	**H**	**H**	**-**	**-**
	B	**H**	**-**	**H**	**H**	**-**	**-**
Intention	A	**H**	**-**	**H**	**M**	**M**	**-**
	B	**-**	**-**	**-**	**H**	**-**	**-**

Desire for punishment: H= High; M= medium: - = no



		unfair	fair	unfair	unfair	fair	fair
Outcome	A	H	-	H	H	-	-
	B	H	-	H	H	-	-
Intention	A	H	-	H	M	M	-
	B	-	-	-	H	-	-
Result	A	3.7	0.4	4.3	1.3	0.2	0.2
	B	0.4	0.3	0.7	3.6	0.2	0.2

Observations

- Outcome-oriented models fail to explain why B is not punished when A is unfair.
- Intention-based models fail to explain why the punishment of A after delegation depends on the behavior of B.

- Idea:
  1. There is punishment only when the outcome is unfair; i.e. there is no punishment when an unkind action has no consequences.
  2. People are punished according to their *responsibility* for the outcome.

Types of Responsibility

- Prospective responsibility: Who has the duty to achieve an outcome?
    - Responsibility can be assigned. What is the optimal way to assign responsibility? (Prendergast 1995, Sliwka, 2004)
- Retrospective responsibility: Who caused an outcome?
    - …

Notions of Retrospective Responsibility

Retrospective responsibility: Who/what caused an outcome?

- Responsibility – the person or luck (bad luck)?
  - Responsibility for the outcome components under one's control.
  - Strict egalitarianism – Liberal egalitarianism – Libertarianism (Experimental studies by Konow, 2000; Frohlich et al. 2004; Fleurbaey, Maniquet, 2005; Cappelen et al. 2006).

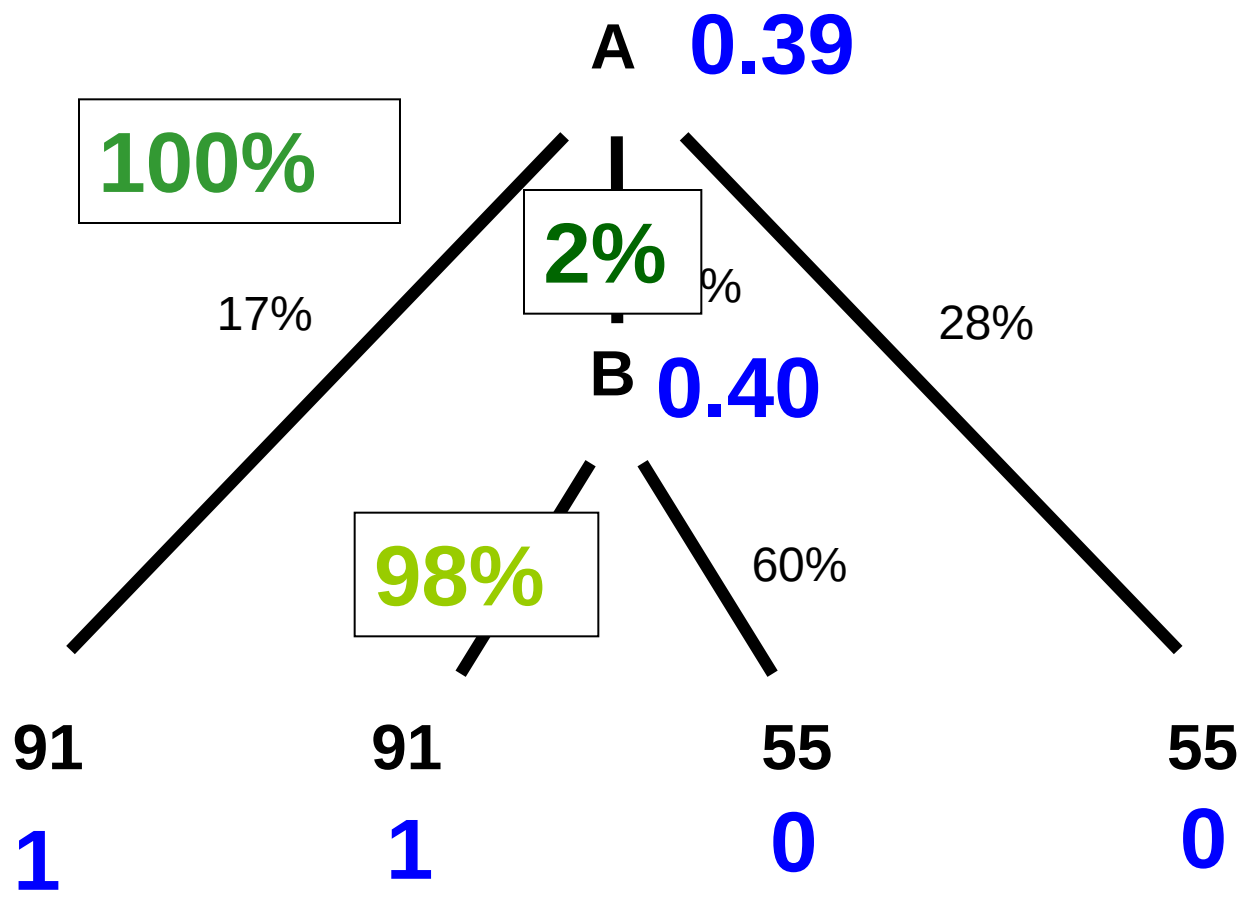
- Responsibility in a decision chain. Who of different people is responsible, in particular if a decision can be delegated?
  - Camerer (2003, p. 56): Define the last-moving player who affects player i's payoff as the only one 'responsible' for i.
  - Our concept…

A Measure of Causal Responsibility

- The measure formalizes a player's responsibility for an outcome (e.g. the unfair allocation) of a game.

- In a nutshell:
  - Calculate how the moves of a player change the probability that the unfair allocation occurs, given some belief, and given that the unfair allocation indeed occurs.
  - Calculate the relative influence of a player (excluding nature).
  - Allow for belief heterogeneity/uncertainty.

A Measure of Responsibility Example



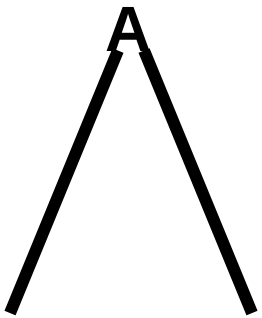
- Determine probability of unfair outcome.
- How much does a move affect this probability?

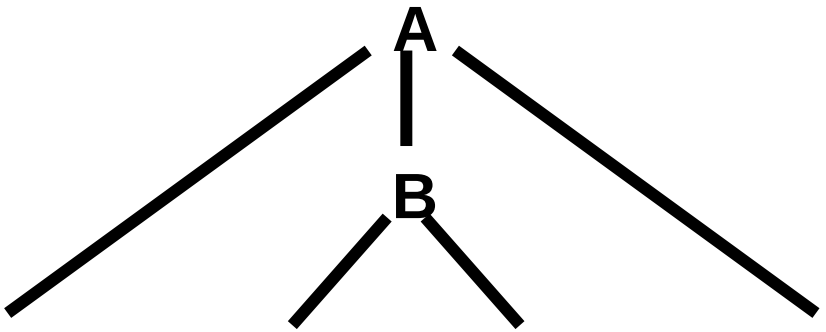
⇒ If only A decides, A is fully responsible.

⇒ If A delegates and B is unfair, almost all responsibility is attributed to B.

⇒ If the outcome is fair, nobody is responsible.

… Beliefs based on experiment.



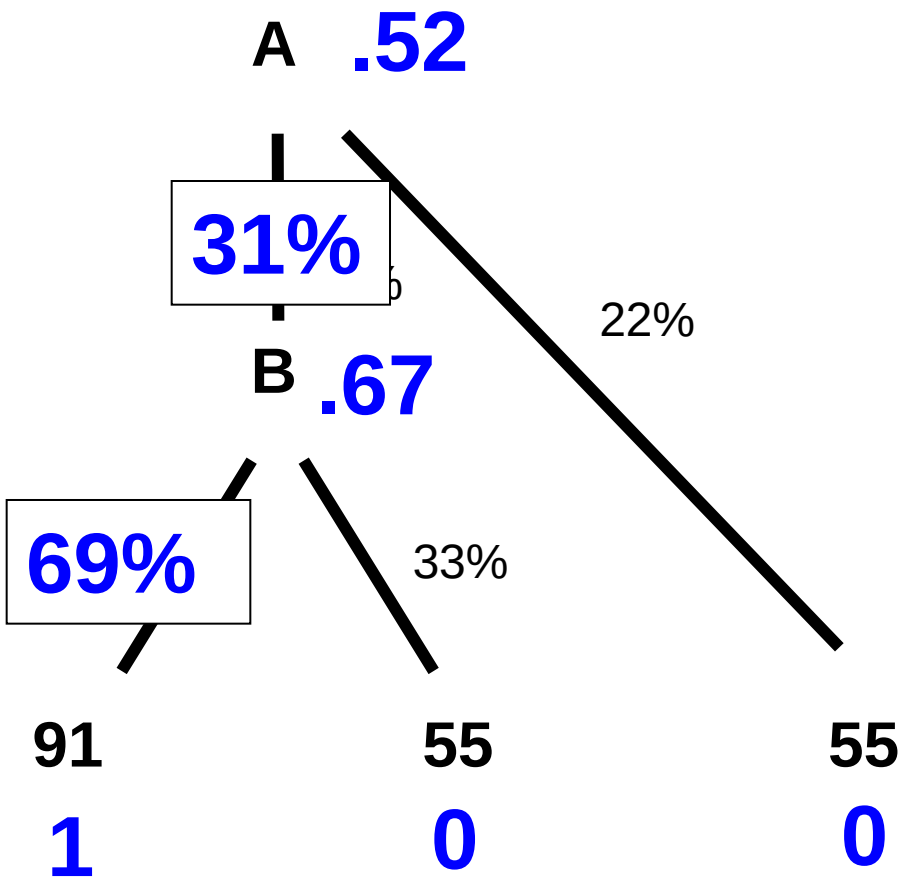


		91	55	91	91	55	55
Outcome	A	1	-	1	1	-	-
	B	1	-	1	1	-	-
Intention	A	1	-	1	.34	.34	-
	B	-	-	-	1	-	-
Respon-sibility	A	1	-	1	.02	-	-
	B	-	-	-	.98	-	-
Result	A	3.7	0.4	4.3	1.3	0.2	0.2
	B	0.4	0.3	0.7	3.6	0.2	0.2

Control Treatments

- Is it possible to shift the blame to a dice?
  - ➢Player A can be fair or unfair or he can delegate to a dice that has a 60:40 chance of the fair distribution.
- Does the choice set have an influence on responsibility and punishment?
  - ➢Player A can only be fair or delegate.

The Calculation of Responsibility



Econometric Comparison of the Different Punishment Motives

- Regressions with the model predictions as regressors.

- Outcome:
  - 1 if unfair, 0 if fair.
- Intention:
  - 1 (0) if the action is taken that increases the probability of the unfair outcome by the most (least).
  - Values in between if more actions are possible.
- Responsibility
  - 0 if outcome fair.
  - Otherwise responsibility with beliefs from experiment.

Outcome	2.036			0.483
	(0.105)***			(0.089)***
Intention		2.733		0.244
		(0.160)***		(0.141)*
Responsibility			3.748	3.141
			(0.198)***	(0.265)***
Constant	0.252	0.286	0.395	0.221
	(0.043)***	(0.042)***	(0.041)***	(0.043)***
Observations	1792	1792	1792	1792
R-squared	0.21	0.29	0.42	0.43

OLS regression, Players A and B
Robust standard errors in parentheses:
* significant at 10%; ** significant at 5%; *** significant at 1%

Conclusions with respect to responsibility

- Responsibility can be shifted by delegation, but only to people.
- Responsibility shirking is a strong motive for the delegation of a decision right.
- A simple formal measure of causal responsibility is supported by the data.

Why do people care about intention and responsibility?

- Punishment is altruistic if it creates incentives for behaving well.
  - …and do not harm people who need no incentive.
- Do not punish people without bad intention.
  - One only has to punish those who do not have a concern for the others.
- Punish those with the highest responsibility
  - …those with the highest impact on the outcome.
  - Nature cannot be incentivized.

Final Conclusions

- Intentions and causal responsibility are important determinants of punishment behavior.
- The concepts can be formalized.
- The formalizations correspond to concepts suggested by philosophers.

- The punishment patterns based on the attribution of intentions and causal responsibility are consistent with *altruistic* punishment.

Thank you for your attention