The Structural Evolution of Morality

J. McKenzie Alexander

Department of Philosophy, Logic and Scientific Method
London School of Economics and Political Science

Outline

- Part I. The Structural Evolution of Morality
  - Trust
  - Fairness
- Part II. Evolutionary Game Theory as a Tool for the Moral Philosopher

Part I.

The Structural Evolution of Morality

The core claims.

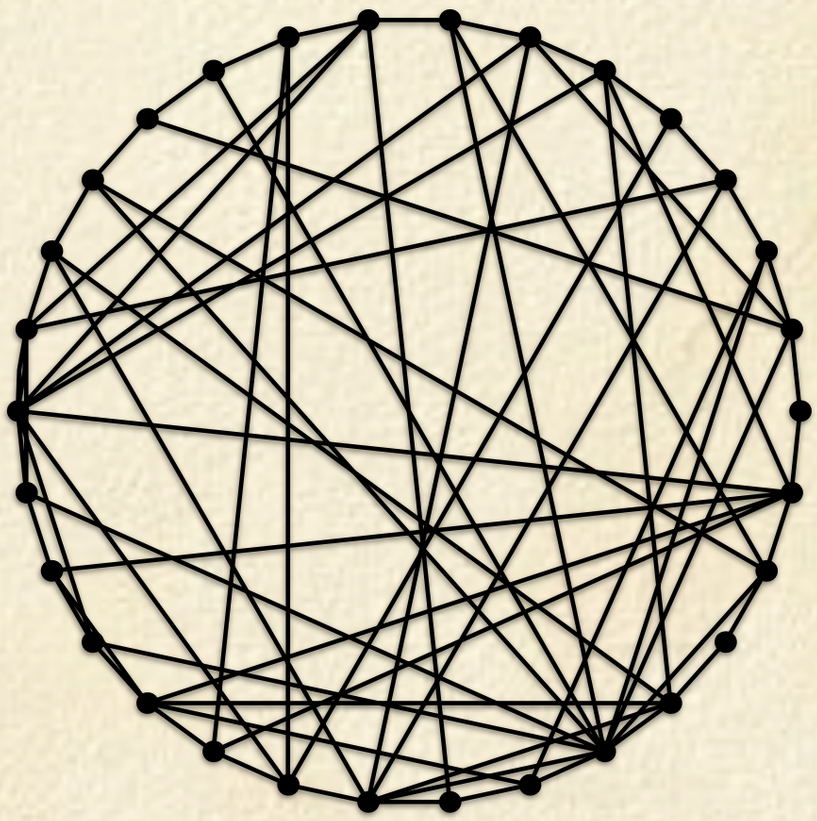
1. Morality provides a set of heuristics that, when followed, serves to produce the best expected outcome, for each of us, over the course of our lives, given the constraints placed by other people.

2. Our moral theories make the behavioural recommendations they do because of evolution and the structured nature of society.

- The kind of evolution I am talking about is *cultural evolution*, by which I mean nothing more than change in belief over time.

Local interaction models

- A local interaction model is a game played on a graph (or graphs).
  - Vertices in the graph represent persons.
  - Edges in the graph represent social relations.



Local interaction models

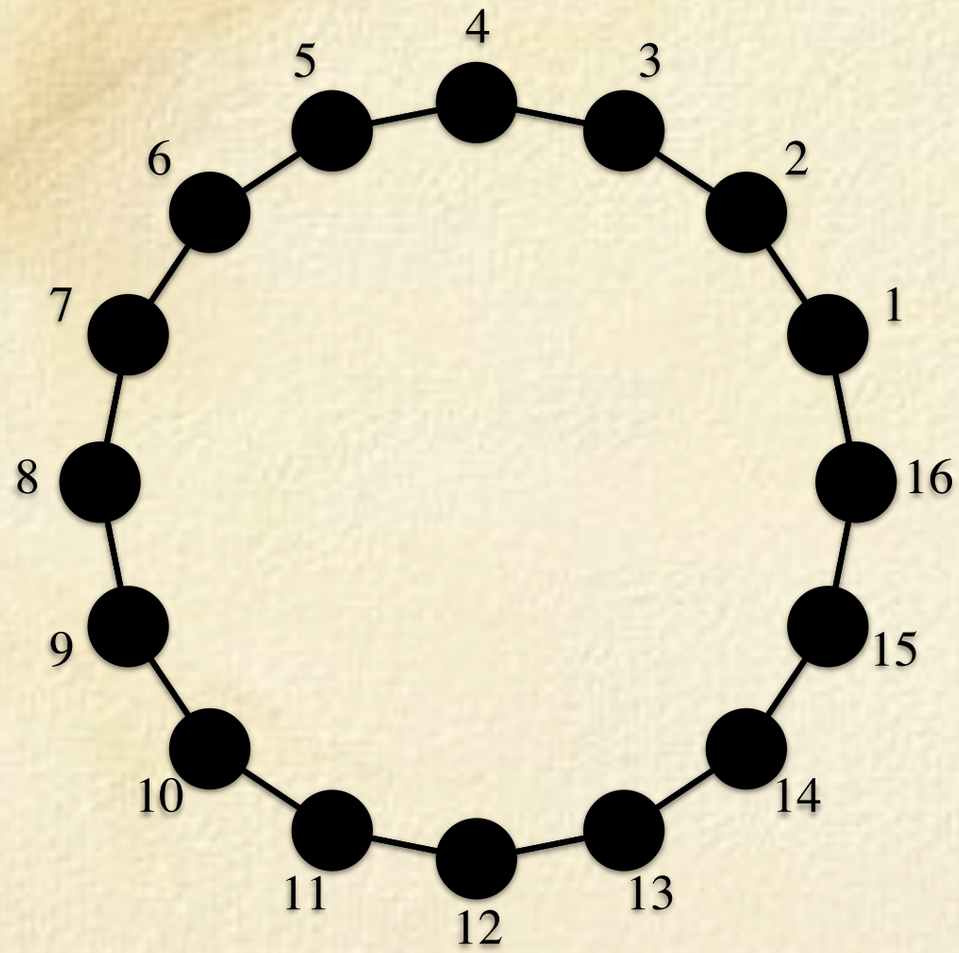
- How people learn:
  - In each round of play, persons play a game with everyone in their interaction neighborhood.
  - The total payoff received is the sum of each individual game.
  - After each round of play, a person *p* adopts the strategy of some person *q* in their update neighborhood, if *q's* total payoff is greater than *p*'s total payoff (and *q* was the best overall in *p*'s update neighborhood). This is a form of imitative learning.

Trust

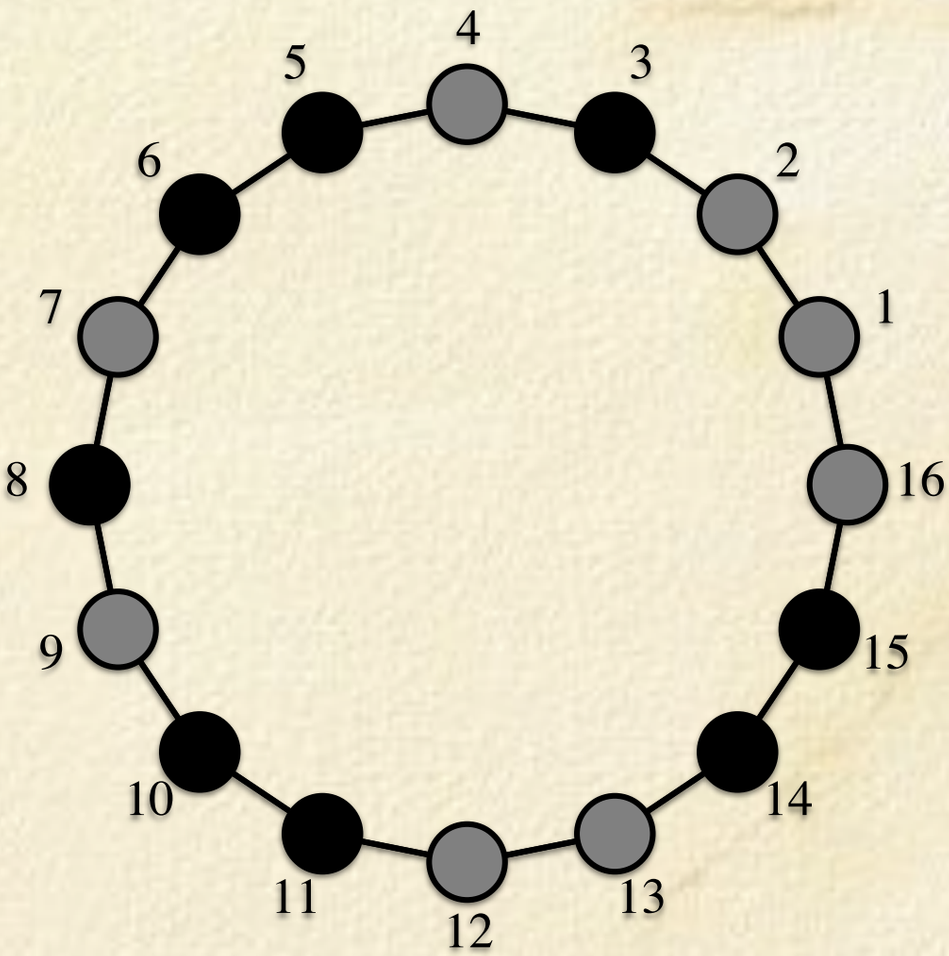
The Stag Hunt / Assurance game

Me \ You	Stag	Hare
Stag	x	0
Hare	y	z

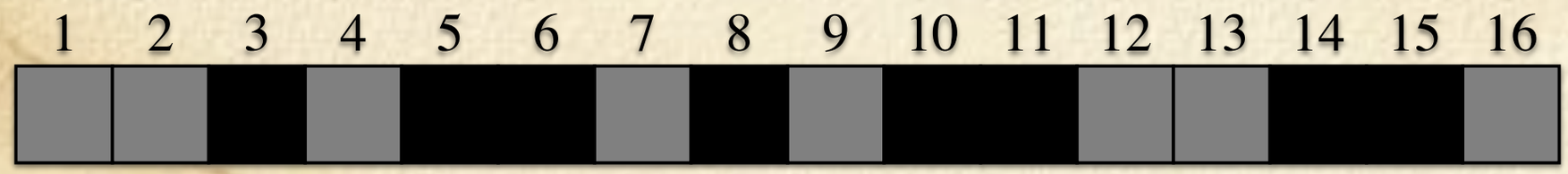
Stag Hunt on a ring



Basic interaction structure



With strategies indicated



A more concise representation

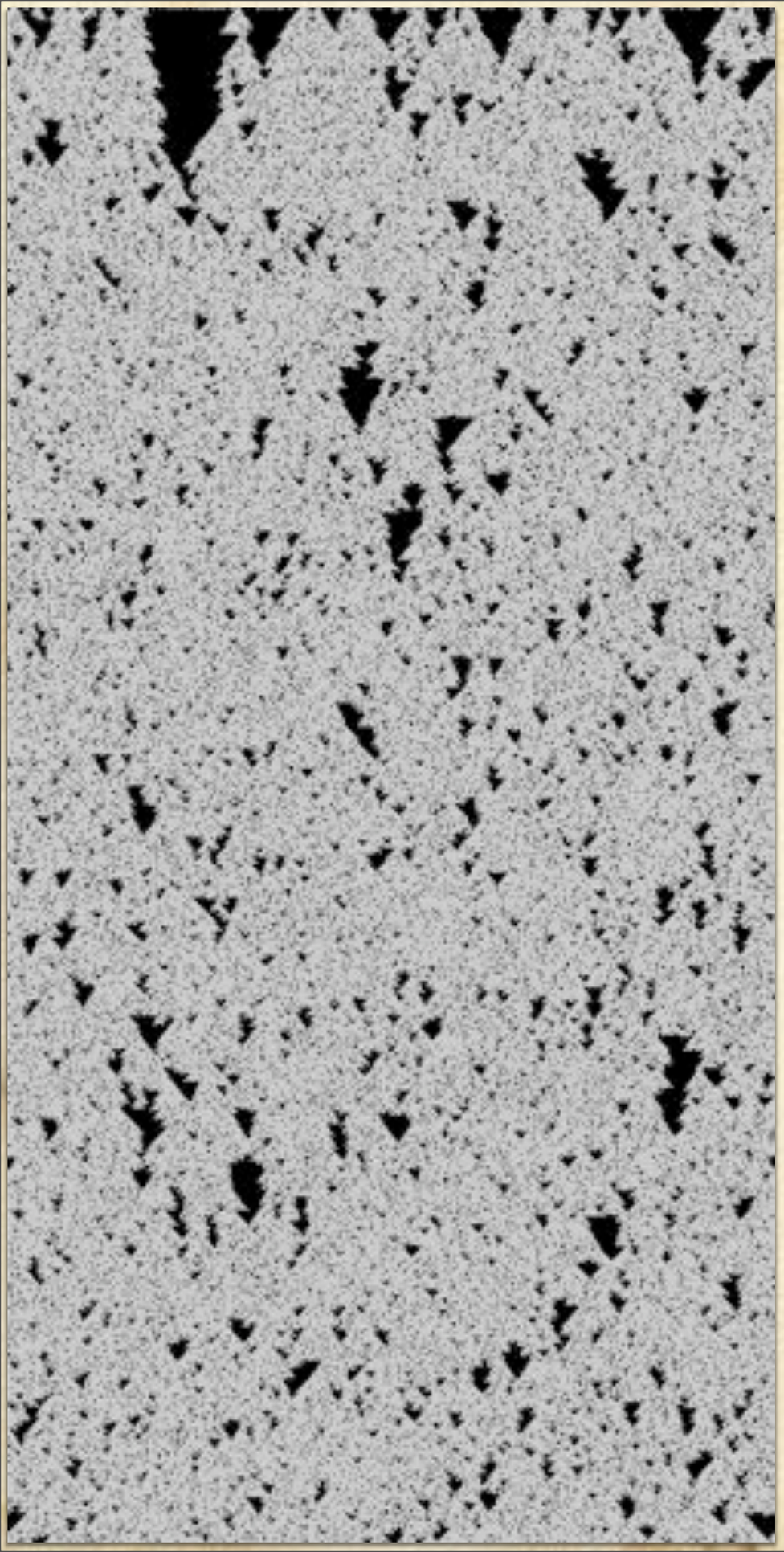
The Stag Hunt on a ring

- Interaction radius of 2
- Payoff matrix:

	S	R
S	3	0
R	2	2

Hunt Stag

Hunt Rabbit



The Stag Hunt

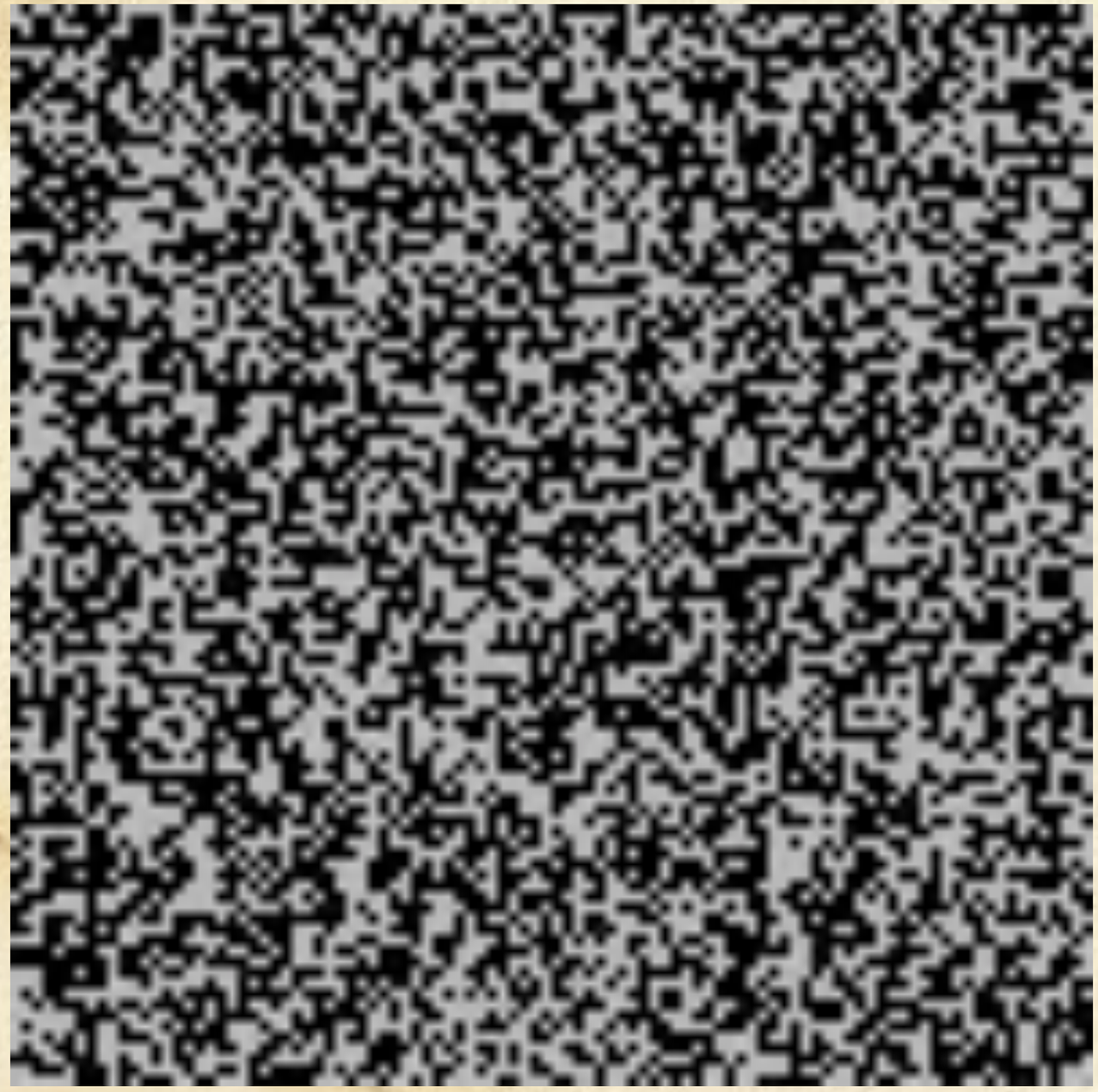
- The only experimentation permitted changes Stag Hunters into Rabbit Hunters, with a probability of 0.1.

Hunt Stag

Hunt Rabbit

The Stag Hunt

Interactions with 4 nearest neighbors

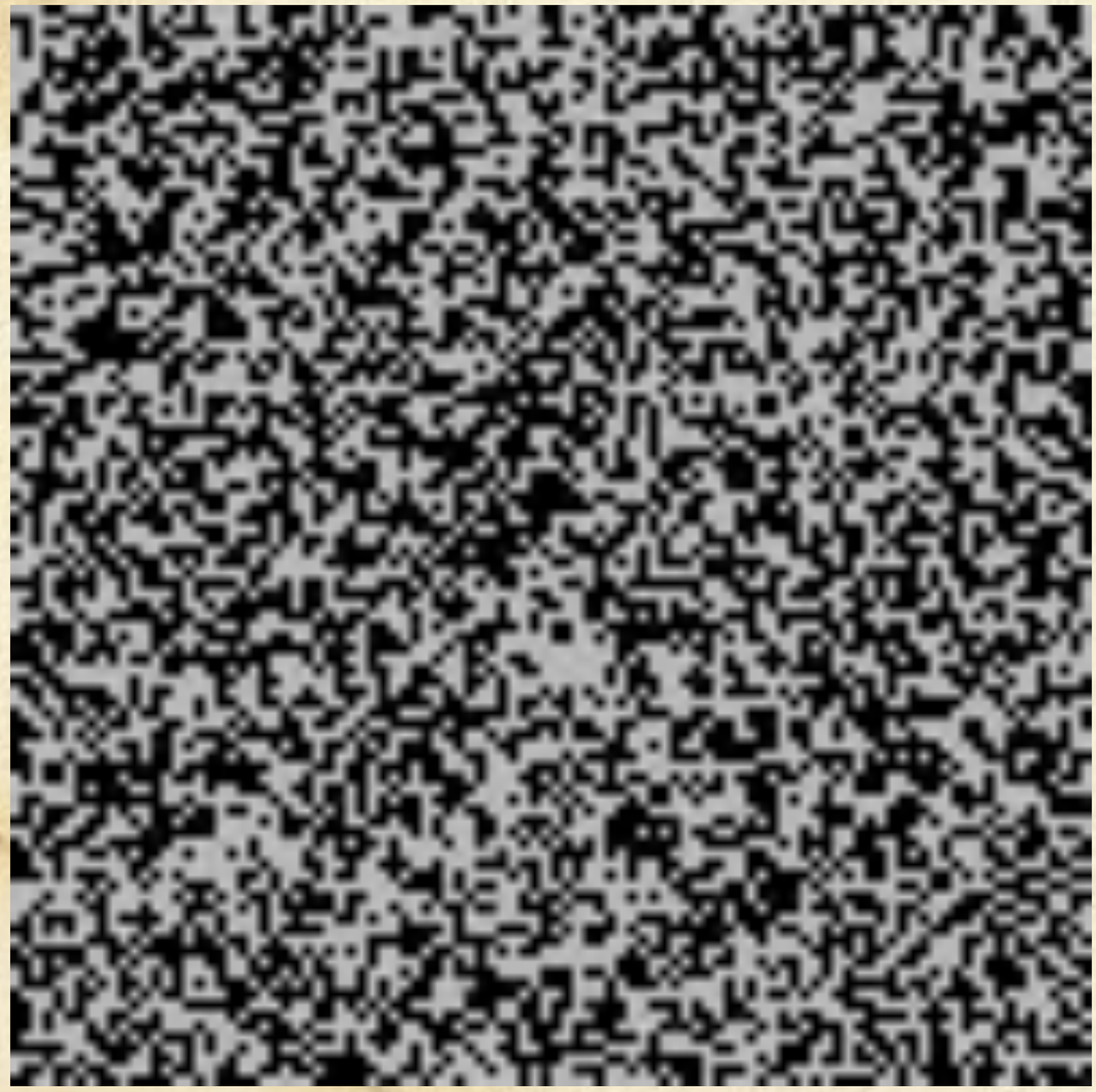


Hunt Stag

Hunt Rabbit

The Stag Hunt

Interactions with 8 nearest neighbors



Hunt Stag

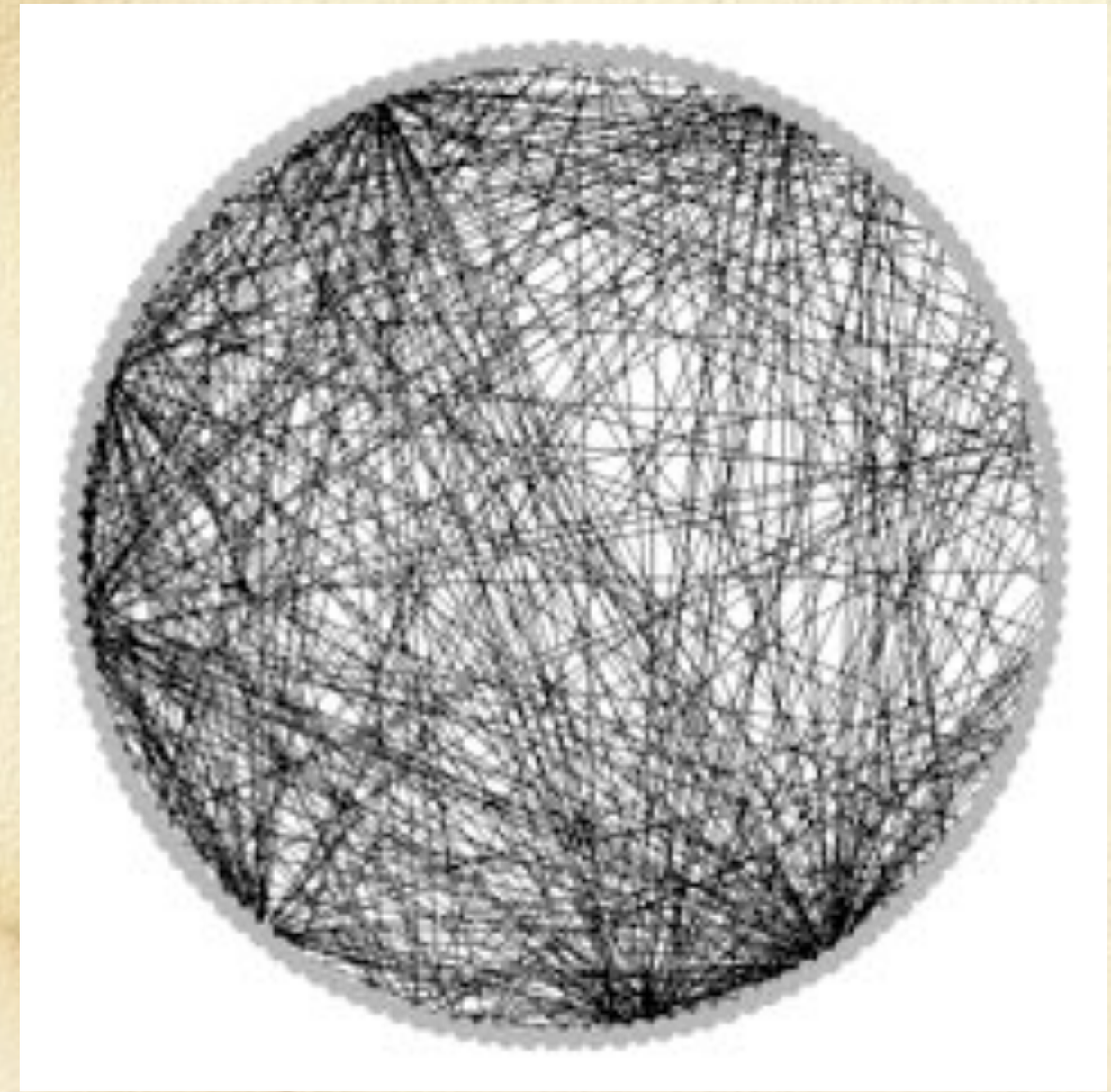
Hunt Rabbit

Stag Hunt on “Realistic” networks

- It is difficult to collect data concerning the shape of actual human social networks.

- The Stanford GraphBase (Knuth, 1993) contains definitions of graphs representing the social network of acquaintanceship for all the characters in several major novels

The *Anna Karenina* social network

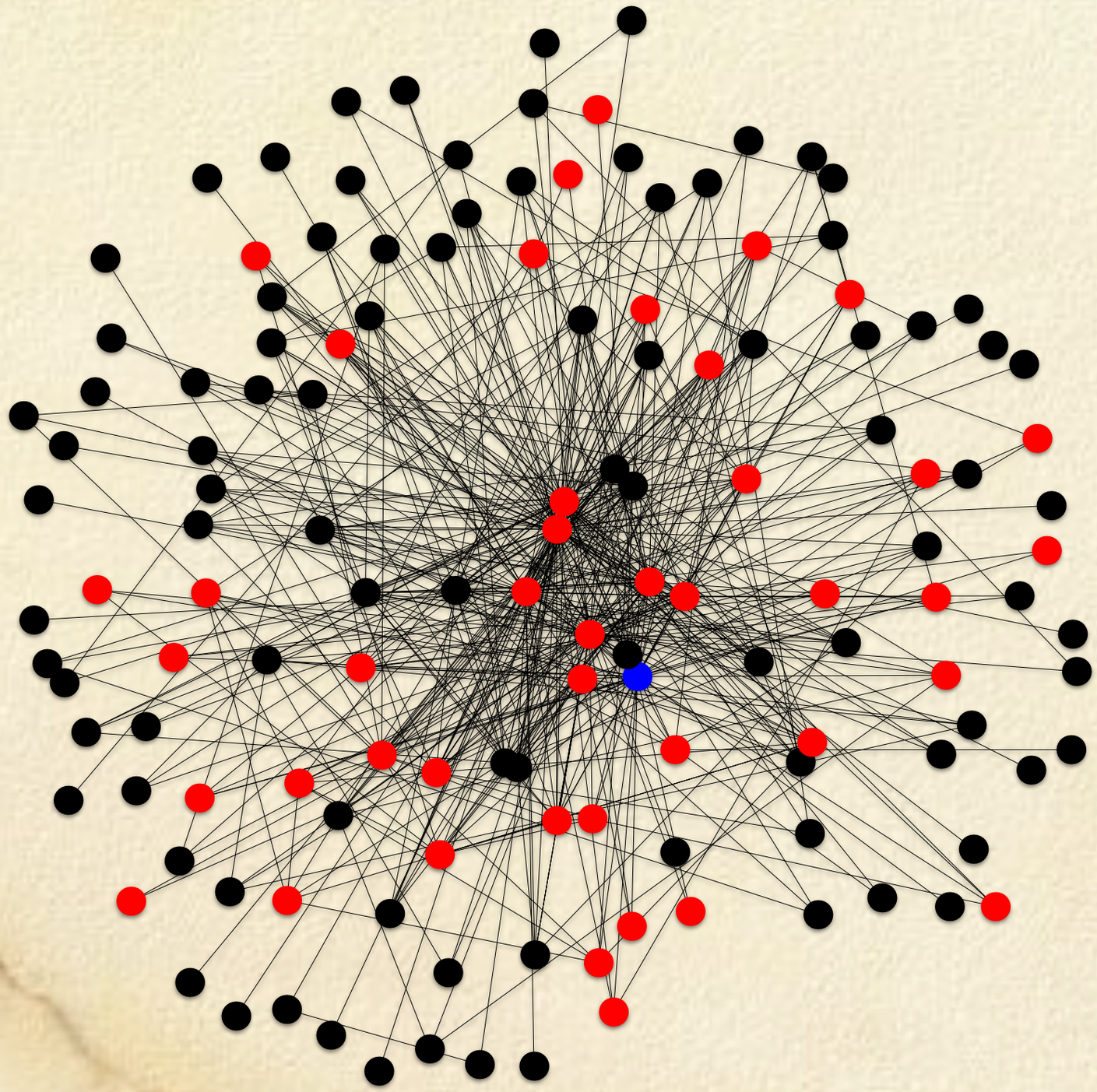


Social network for *Anna Karenina*

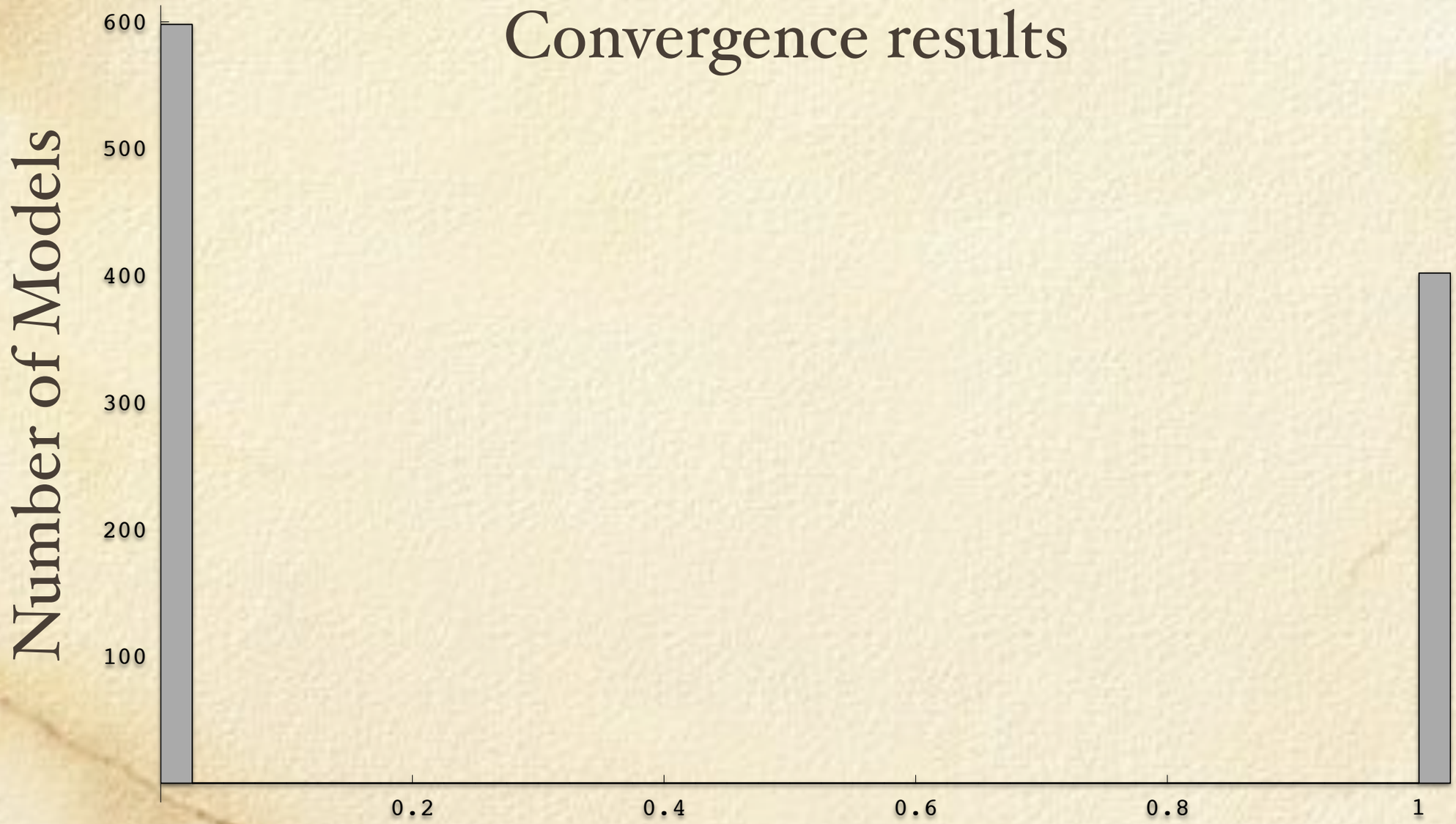
- Edges represent encounters between characters in novel.

(Source: Stanford GraphBase)

The *Anna Karenina* social network

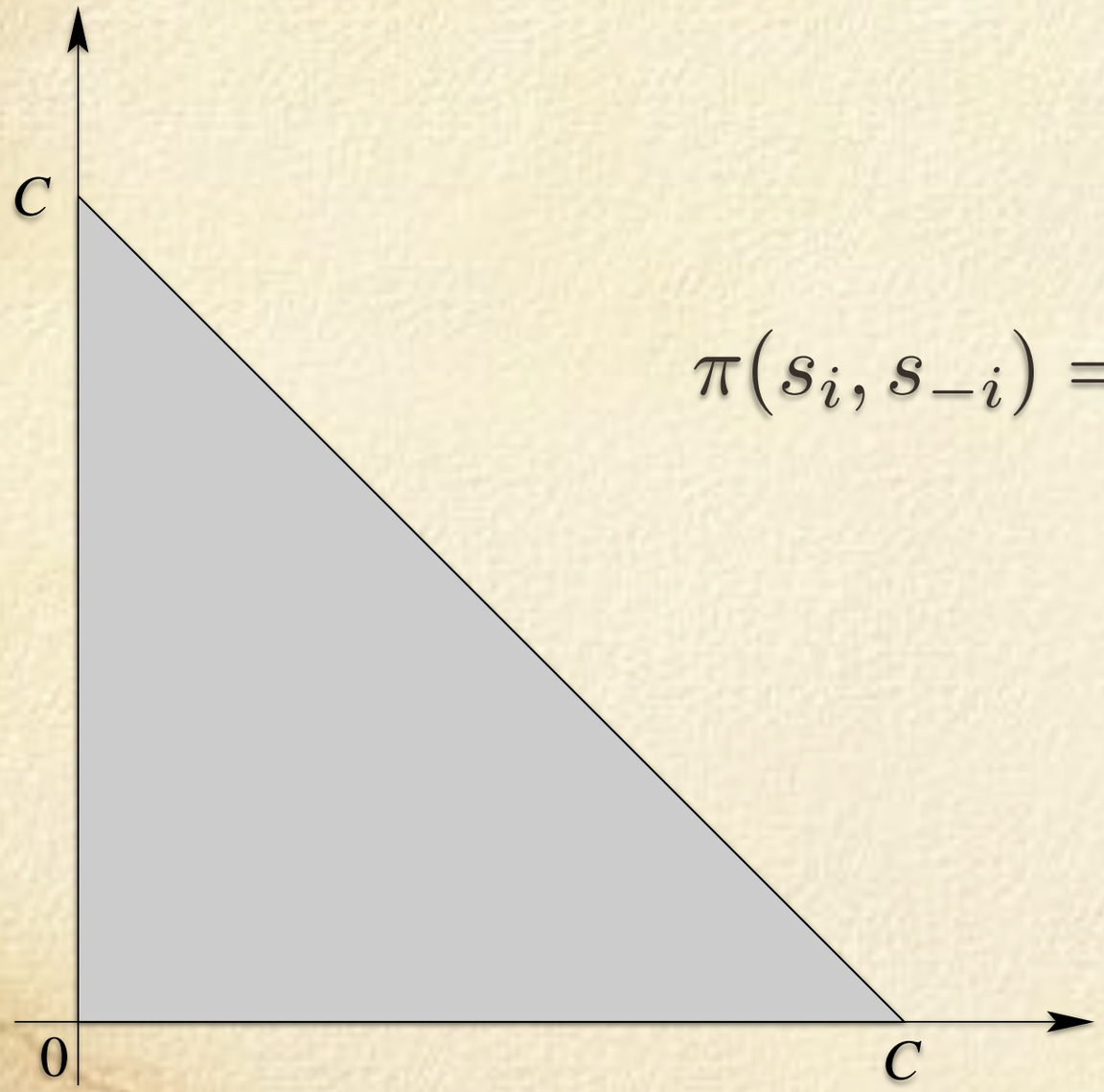


Stag Hunt played on the *Anna Karenina* social network



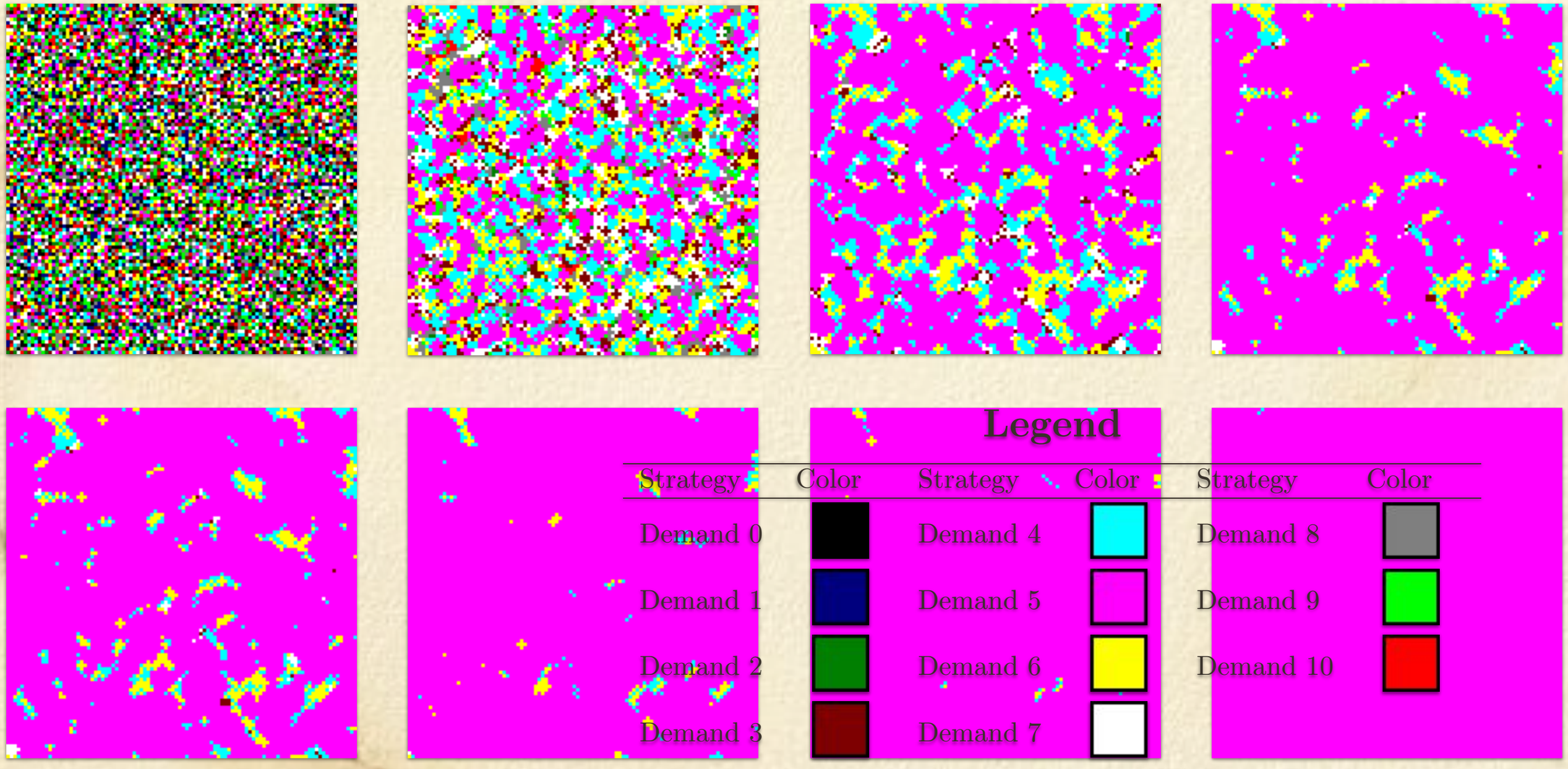
Fairness

Game 1: Divide-the-dollar



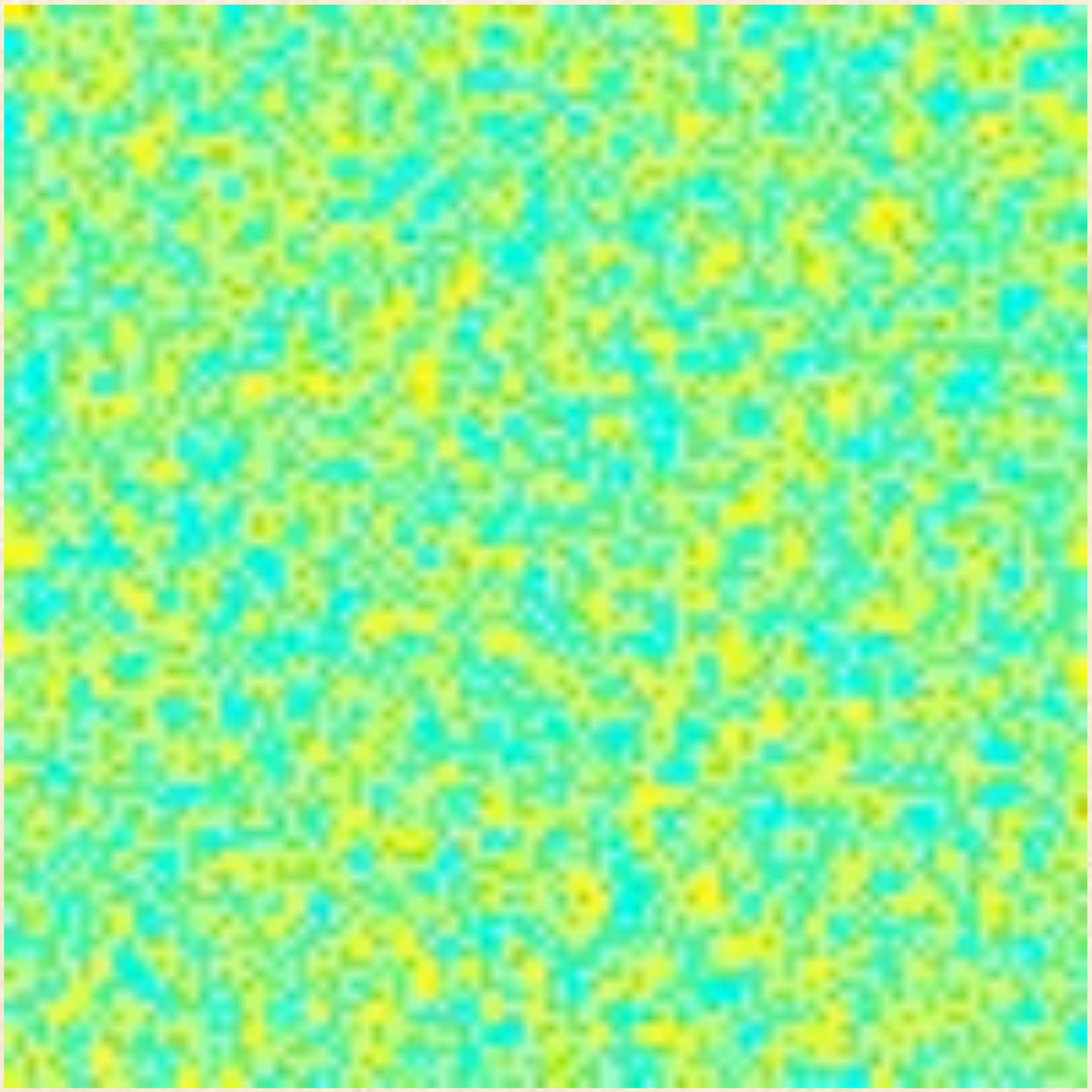
$$\pi(s_i, s_{-i}) = \begin{cases} s_i & \text{if } s_i + s_{-i} \leq C \\ 0 & \text{otherwise} \end{cases}$$

von Neumann neighbourhood



Every 2nd generation shown

The robustness of demand half

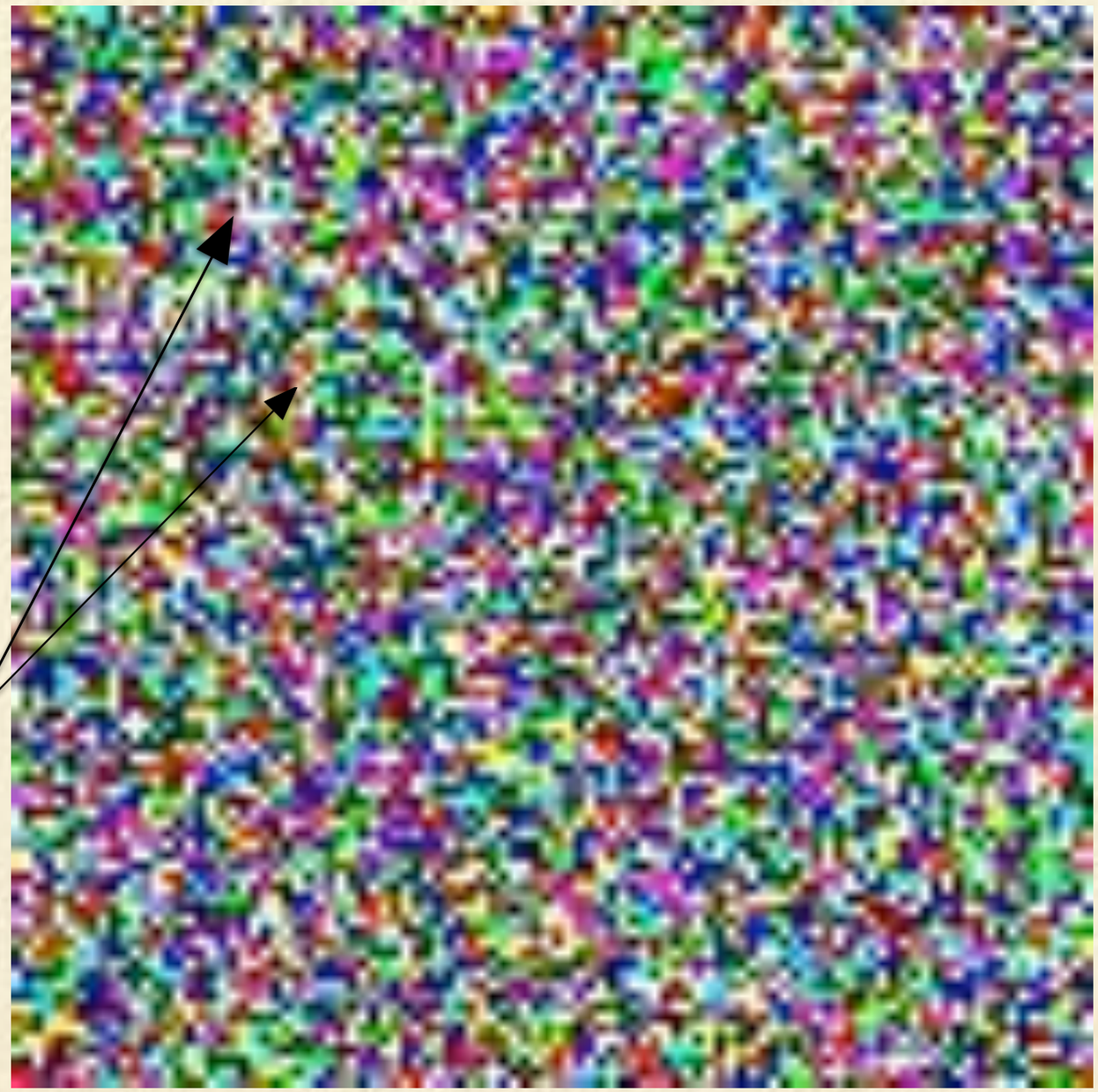


$\mu = 0.0001$

Legend

Strategy	Color	Strategy	Color	Strategy	Color
Demand 0		Demand 4		Demand 8	
Demand 1		Demand 5		Demand 9	
Demand 2		Demand 6		Demand 10	
Demand 3		Demand 7			

Inferiority of best response



Regions of suboptimal performance

Transient regions of fair division

$\mu = 0.01$

Part II.

Evolutionary Game Theory as a Tool for the Moral Philosopher

The central problem

- There's more to explaining morality (even the origins of morality) than accounting for mere behaviour.

- "There is a significant conceptual difference between apparent altruism in general, and the kinds of human motivational patterns that are thought to be morally significant. Furthermore, there's a difference between demonstrating that evolution by natural selection is compatible with morality and demonstrating that the former helps to explain the latter."

D'Arms (2000)

- “…it’s important to demonstrate that the *forms of behaviour* that accord with our sense of justice and morality can originate and be maintained under natural selection. Yet we should also be aware that the demonstration doesn’t necessarily account for the superstructure of concepts and principles in terms of which we appraise those forms of behaviour.”

(Kitcher, 1999)

Descriptions of action

- Thick and thin action descriptions:
  - An agent thinly conforms to a moral principle (or theory) if her behaviour conforms to what is required by the principle (or theory).
  - An agent thickly conforms to a moral principle (or theory) if her behaviour conforms to what is required by the principle (or theory); she holds "enough" of the associated moral attitudes, intentions, preferences, and desires; and she performs that action for the "right reasons in the right way."
- Can evolutionary game theory account for thick descriptions of moral action?

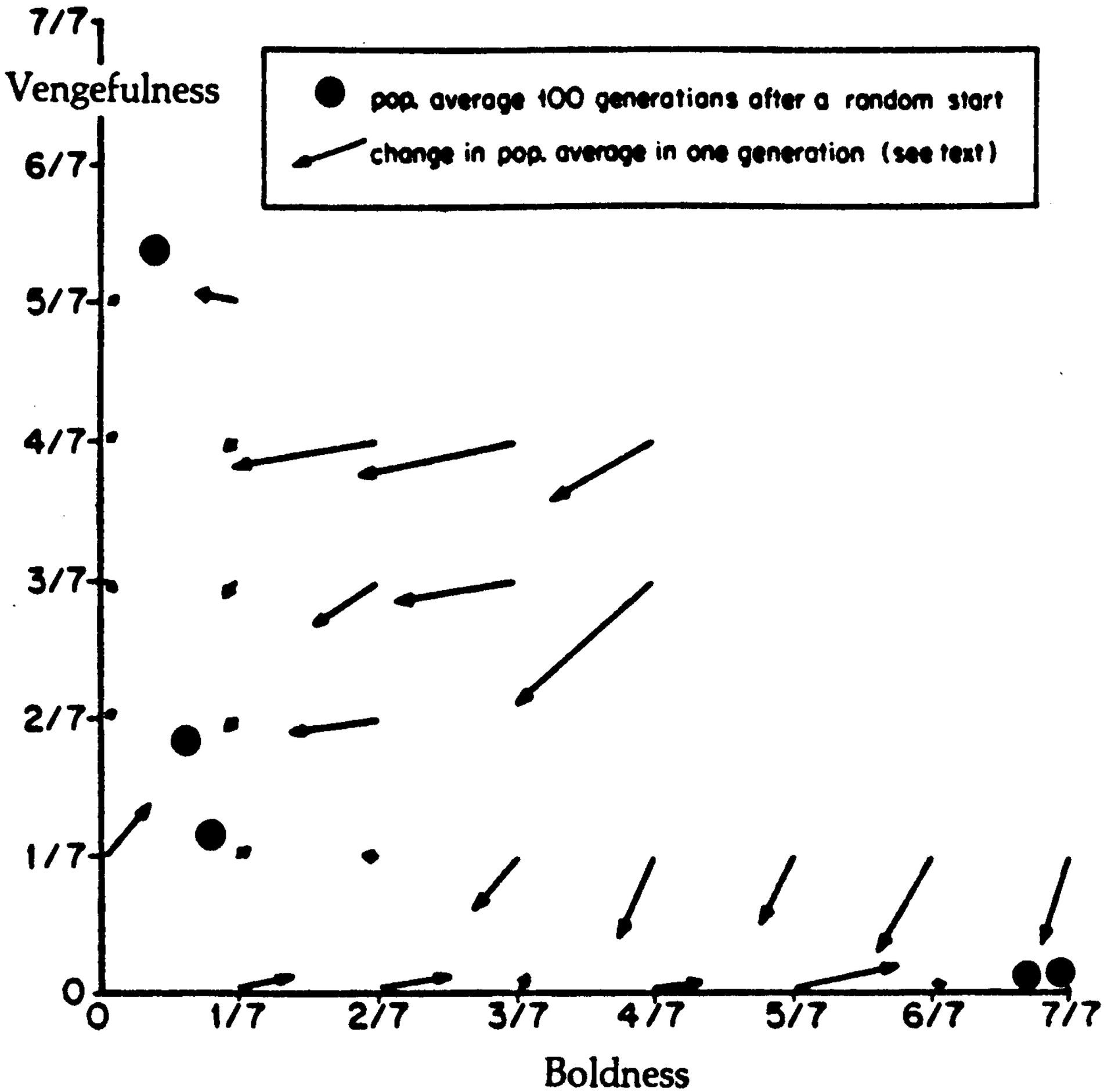
Two possible solutions

1. Expand the set of games to include ones with richer strategy sets.
    - Consider games whose strategies include the adoption of certain attitudes by the agent. (Axelrod 1986, Gintis 2000, many others).
2. Enrich the conception of boundedly rational agents to included nonstrategic, psychological elements.
    - Fast and frugal heuristics (Gigerenzer, etc.)
    - Evolutionary psychology

Solution approach 1: Axelrod's Norm game

The Rules:

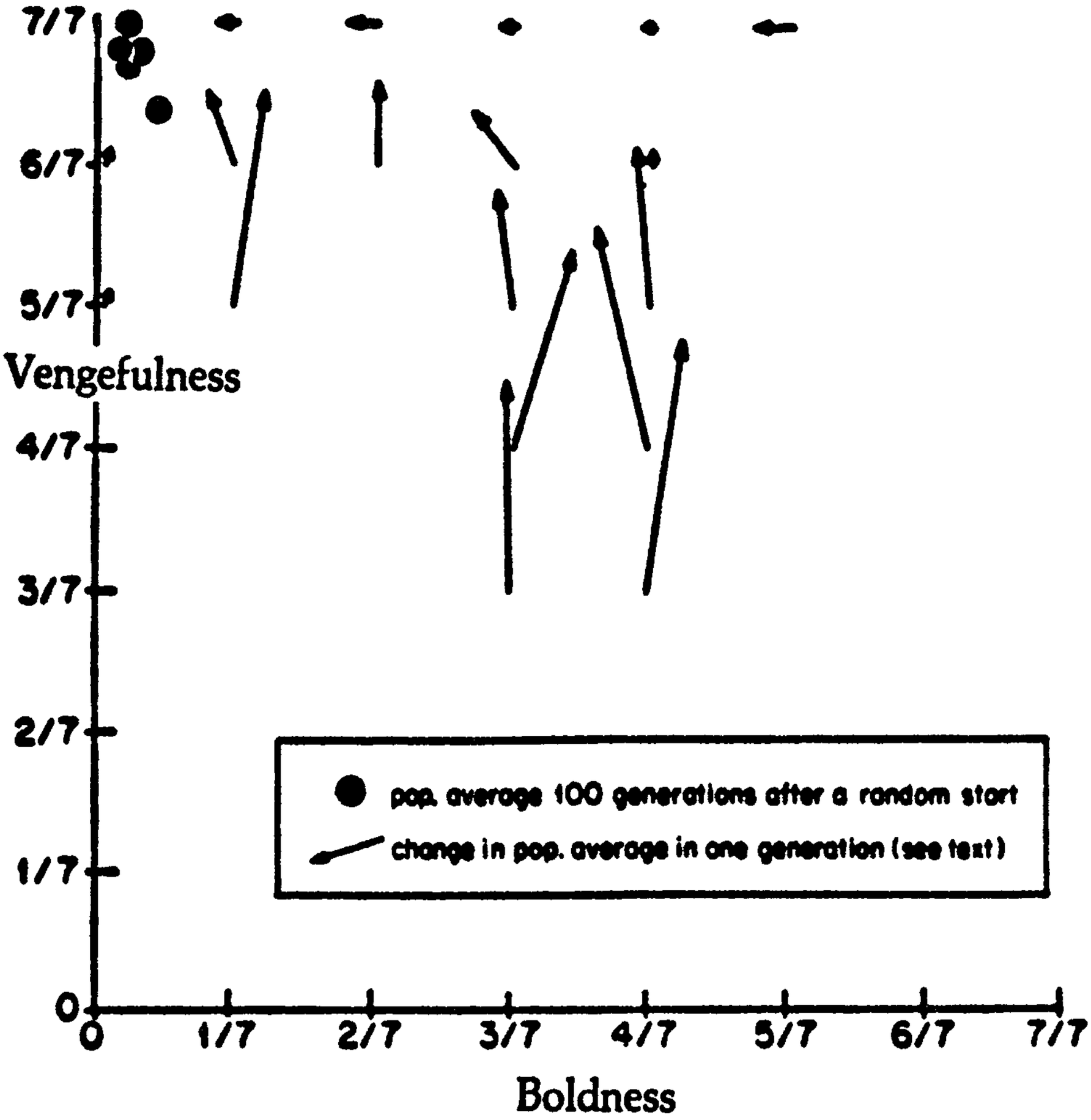
- A population of individuals plays a game. Each individual has a chance to cheat.
- If a player cheats, a payoff of $T=3$ is received. All other players are hurt, receiving a payoff of -1.
- A cheater has a known chance of being seen.
- If seen, another player may punish. Punishment gives the cheater a payoff of -9, with the punisher incurring a cost of -2.


7/7
Vengefulness
6/7
5/7
4/7
3/7
2/7
1/7
0
pop. average 100 generations after a random start
change in pop. average in one generation (see text)
0
1/7
2/7
3/7
4/7
5/7
6/7
7/7
Boldness

Solution approach 1: Axelrod's Metanorm game

The Rules:

- A population of individuals plays the norm game. At the end of each norm game, we add the following:
- If a player (say Jim) sees a cheater and chooses not to punish, then every player (except the cheater) has a chance to punish Jim.
- The parameter which determines punishing non-punishers is the same as that which determines punishing cheaters.


7/7
6/7
5/7
Vengefulness
4/7
3/7
2/7
1/7
0
pop. average 100 generations after a random start
change in pop. average in one generation (see text)
0
1/7
2/7
3/7
4/7
5/7
6/7
7/7
Boldness

Why this approach fails

- It still suffers from the problem of thin descriptions.
- It doesn't matter whether we call the strategies "punish" and "enforce a norm," or whether we have a model of cultural evolution instead of biological evolution. The model still admits a *purely behavioural* interpretation.
- We don't want an account of how evolutionary pressures make people act *as if* they are punishing defectors, we want an account of why they *really* punish – with all the associated mental machinery.

Solution approach 2: Enriching the boundedly rational agent

- Consider enriching our model of the boundedly rational agent to include psychological, nonstrategic elements.

  - Fast and frugal heuristics (Gigerenzer *et al.*)

  - Emotions as fast and frugal heuristics.

Recognition of lost returns

- Regarding fair division, D'Arms claims that when we don't ask for half of the cake, "recognition of the lost returns should suffice to bring us back on track."

- Yes, recognition of lost returns *should* suffice, but it *may* not.

- Agents may realise they are not doing as well as they'd like, but yet be able to correlate any particular act in any particular context as *the* (or *a*) reason why.

- The game of life does not come with a given payoff matrix.

Recognition of rule-governed relations

- We are capable of identifying rule-governed relations and patterns, even when we cannot articulate the rule which governs them.

- Shock experiment (cited by Hardin).

- Given that we can recognise such relations and patterns, yet not be capable of identifying the true underlying rule or process generating those relations, it makes sense that boundedly rational agents will set out rules for themselves to follow - if the choice recommended by the rule correlates well enough with successful payoffs.

Moral rules as heuristics

- Moral rules are heuristics for effective search.

- "There are at least three important types of building blocks of which simple heuristics are composed… (a) there are building blocks to guide information search; (b) different heuristic building blocks determine how to stop search; (c) other building blocks are used to make a decision based on the information gathered. All of these building blocks can be influenced or implemented by processes involving emotions… individual learning… and social learning."

(Sadrieh et al., pp. 93–94)

Evolutionary game theory as a tool for the moral philosopher

- Evolutionary game theory allows us to identify courses of action which maximise long-run expected utility of persons.

- Moral rules are fast and frugal heuristics which individuals rely on to make choices in interdependent decision problems whose complexity precludes full deliberation

- These heuristics are partially embedded in the "morally relevant" emotions, moods, interpersonal affective stances, attitudes, etc., as these play important roles in *producing* behaviour.

What evidence exists?

- Yaari and Bar-Hillel, "On Dividing Justly"
  - Subjects were given mathematically identical distribution problems phrased according to the following categories
    1. differences in needs
    2. differences in tastes
    3. differences in beliefs
- In category (1), subjects favoured Rawls' maximin principle.
- In category (2), subjects increased use of the utilitarian distribution principle.

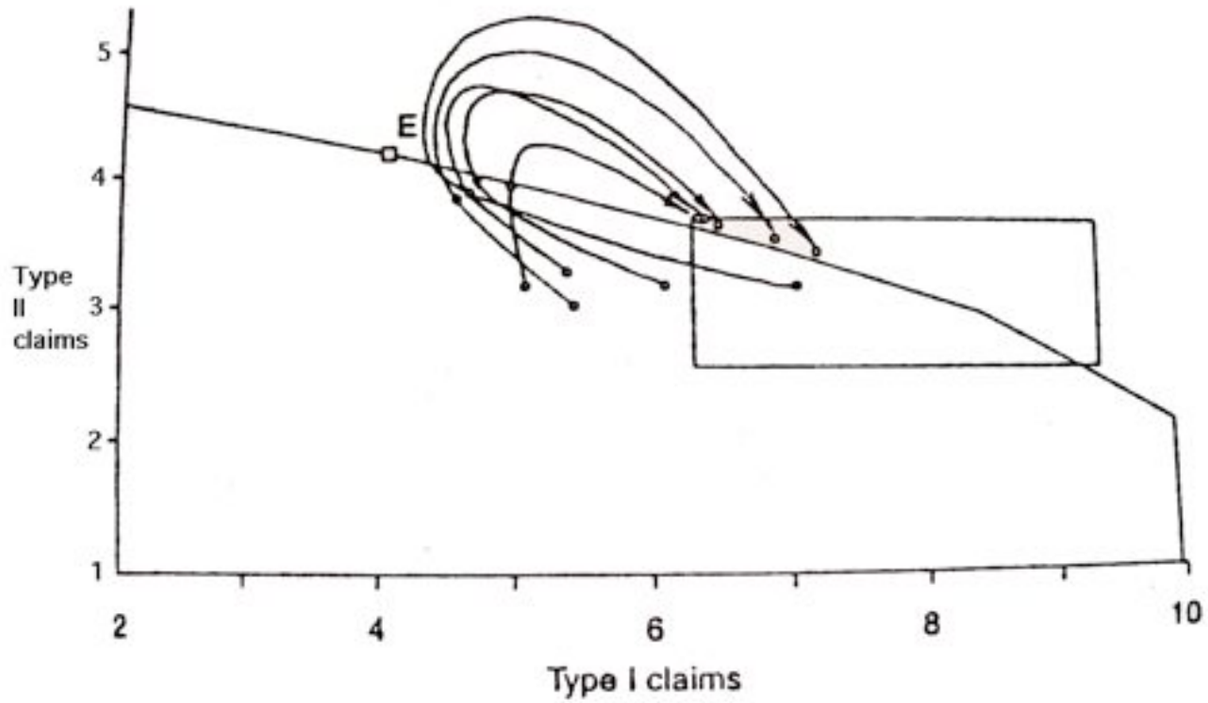
- Binmore *et al.* "Focal Points and Bargaining"

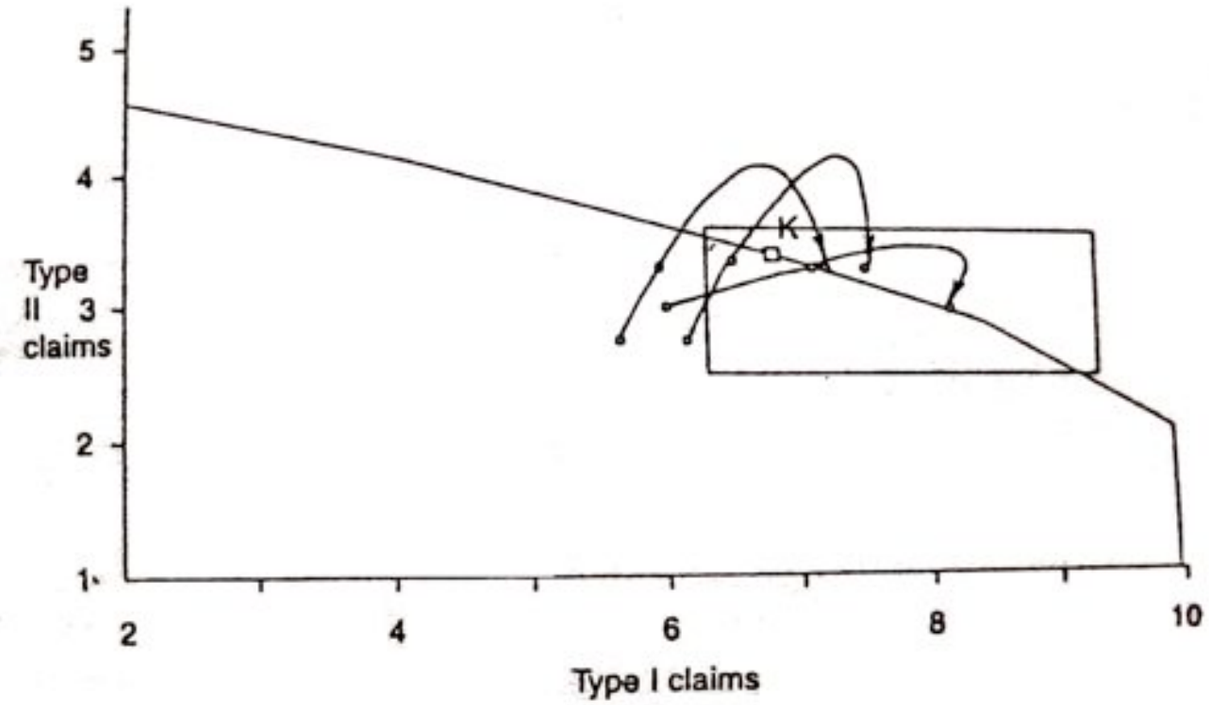
  - Subjects played a repeated asymmetric Nash demand game.

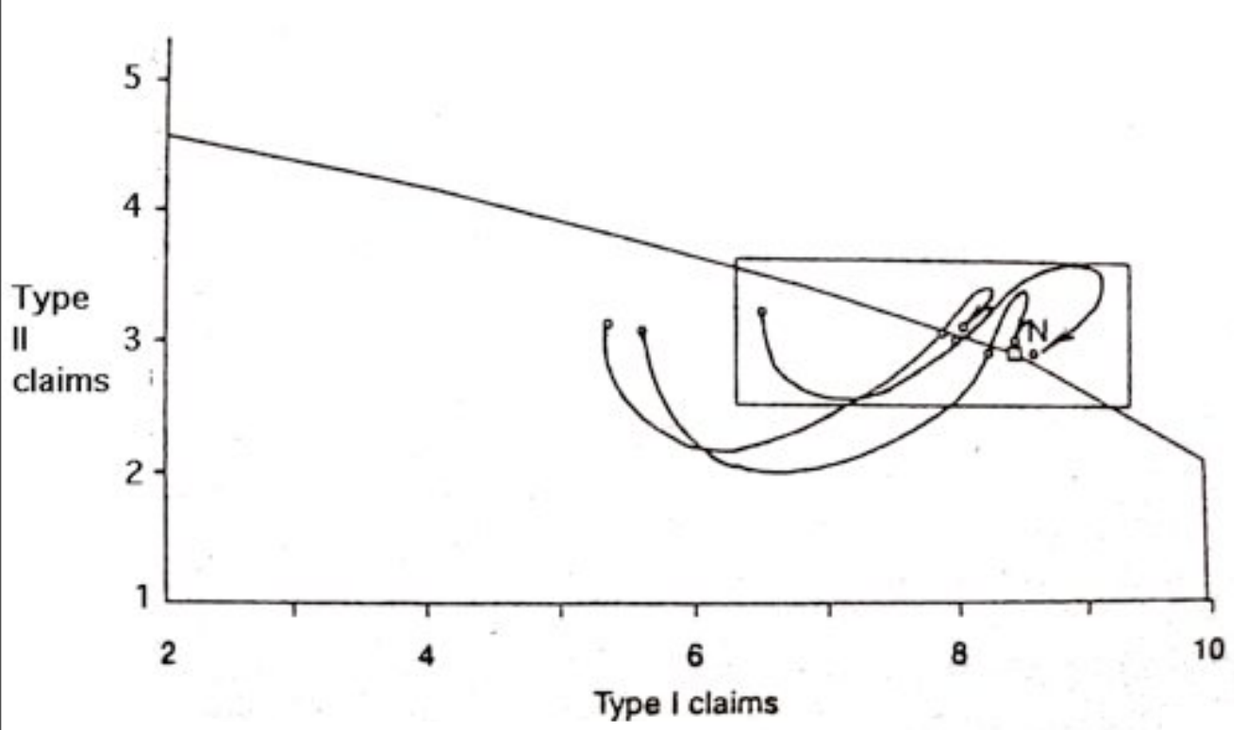
  - Phase 1 was "test runs" against a computer opponent, phase 2 pitted human subjects against human subjects.

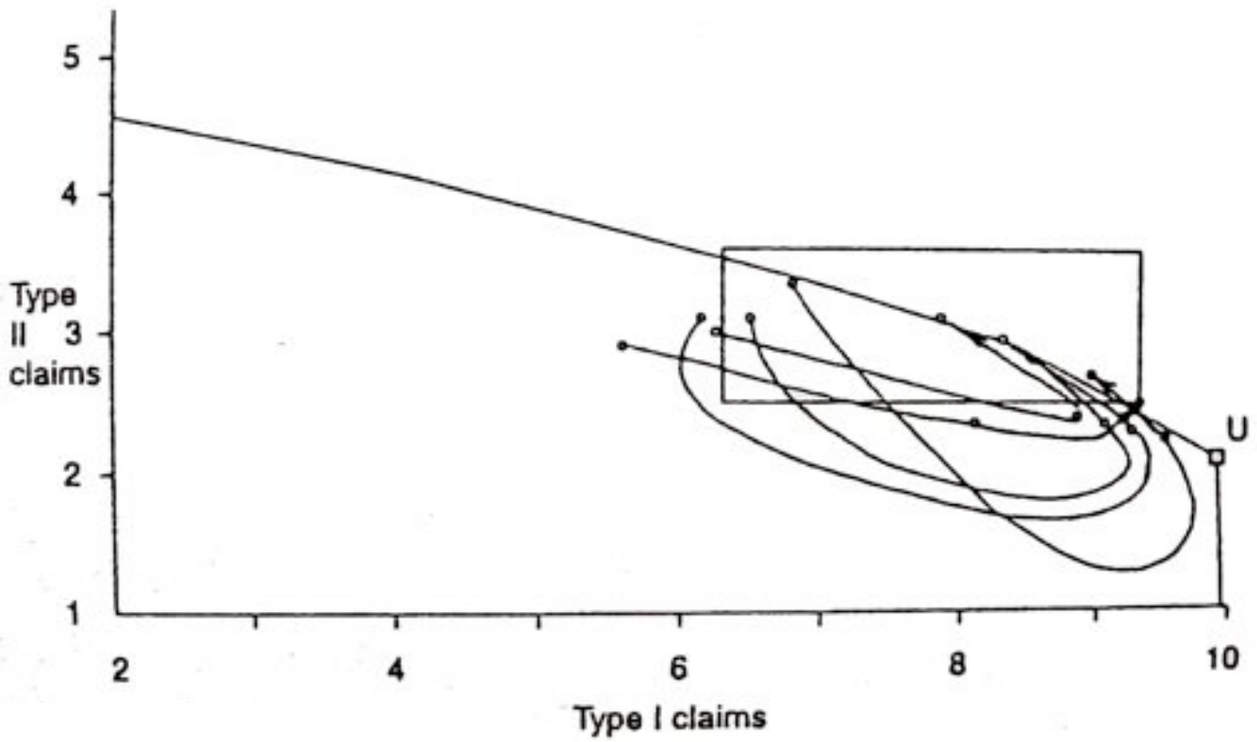
- In phase 1, subjects were conditions to play one of four solutions: Nash, Kalai-Smordinski, Equal increments, and Utilitarian.

Experiments from Binmore *et al.* (1993)









Results

- Subjects could be conditioned to play any solution.
- After conditioning, subjects continued to play as conditioned for some time, but ultimately moved towards the Nash bargaining solution.
- Final outcome depends on conditioning.
- Perceptions of fairness *closely* correlated to the general behaviour of the group.

Perceptions of fairness