# Indirect Reciprocity and the Evolution of "Moral Signals"

Rory Smead

Department of Logic and Philosophy of Science
University of California, Irvine

Evolution of Psychological Categories - IMBS
March 16, 2008

# Moral Signals

Classifying other individuals, or individual actions, as “Good” or “Bad” is an essential feature of human moral systems.

## “Moral Signal”

A signal that both carries information about an individual (or an individual’s actions) and influences the actions of others toward that individual.

# Moral Signals

Classifying other individuals, or individual actions, as “Good” or “Bad” is an essential feature of human moral systems.

**“Moral Signal”**

A signal that both carries information about an individual (or an individual’s actions) and influences the actions of others toward that individual.

To investigate the evolution of moral signals, models of indirect reciprocity may provide a starting point (Harms and Skyrms 2008).

## Direct Reciprocity

If you scratch my back, I will scratch yours.

Mark and Juliette McLean (mark-ju.net)

Helena Goscilo and Petre Petrov (1999) (pitt.edu)

## Indirect Reciprocity

If you scratch my back, someone else will scratch your back.

(kleptography.com)

# The Evolution of Reciprocity

**The Prisoner's Dilemma**

| | $c$ | $d$ |
|---|---|---|
| $c$ | $10, 10$ | $0, 11$ |
| $d$ | $11, 0$ | $1, 1$ |

Evolving cooperation by reciprocation requires that individuals have a method of detecting defectors so that agents can defect on defectors and cooperate with cooperators.

Direct Reciprocity

- Trivers (1971)
- Axelrod and Hamilton (1981)

Indirect Reciprocity

- Alexander (1987)
- Kandori (1992)

# Outline

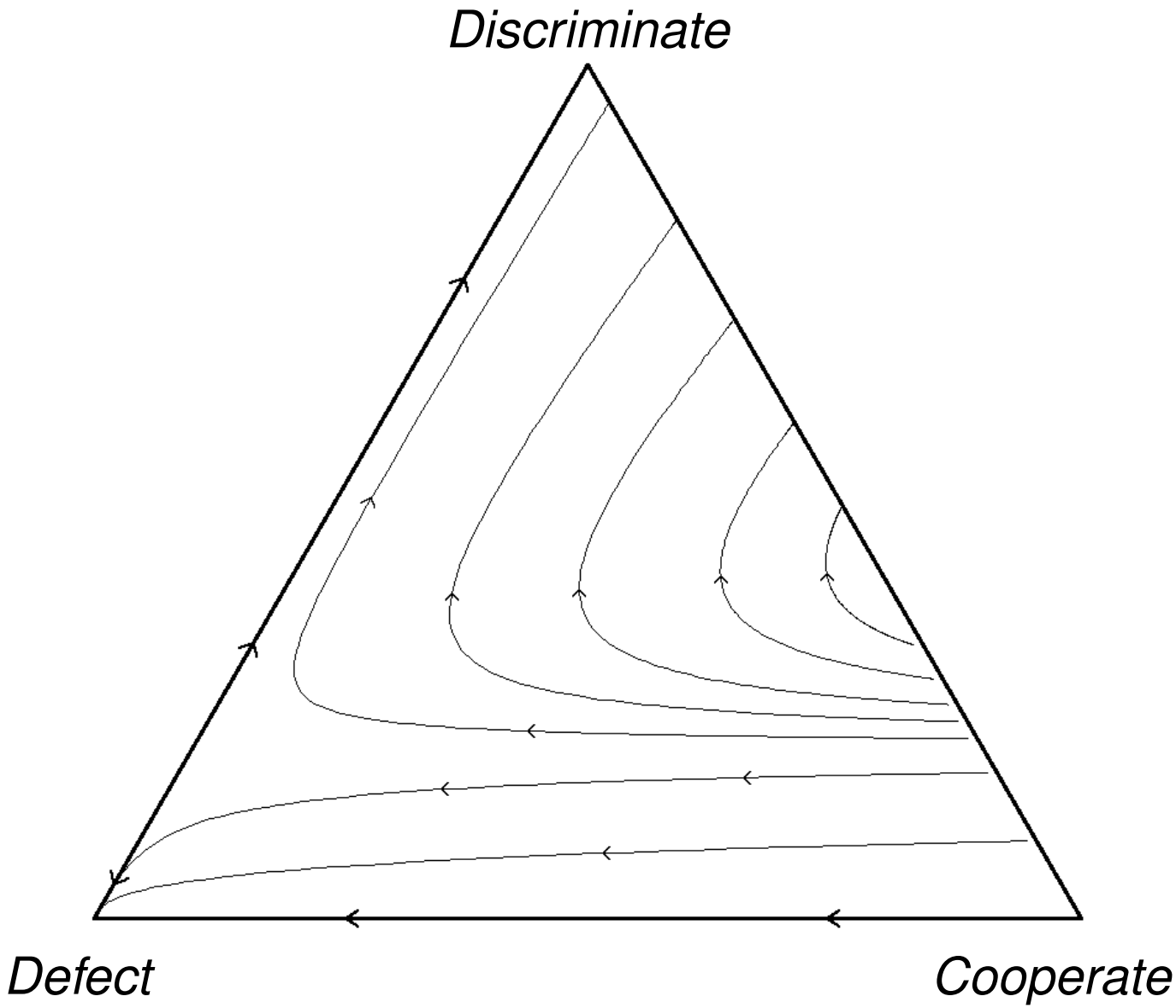
1. A simple model of the evolution of indirect reciprocity (Nowak and Sigmund 1998)
2. Introduce signaling into this simple setting
3. The prospects for co-evolving “Moral Signals” and cooperation
4. Extensions of the model

# The Evolution of Indirect Reciprocity by Image Scoring

Nowak and Sigmund (1998, 2005)

- A large, randomly mixing population playing games of altruism against different opponents for a fixed number of rounds.
- Each player has a binary “image score” which is based on their most recent action.
- There are 3 strategies in the simplest setting: always cooperate, always defect, and discriminate (cooperate with those who have a “good” image score and defect on anyone who does not).
- Evolution occurs according to the replicator dynamic.

# The Dynamics of Indirect Reciprocity



# Image Scoring and “Standing” Strategies

Possible problems with image scoring strategies...

“Standing strategies” are considered by Sugden (1986), Leimar and Hammerstein (2001), Panchanathan and Boyd (2003), and Ohtsuki and Iwasa (2004)

Milinski, Semmann, Bakker and Krambeck (2001)

# The Role of Language in Indirect Reciprocity

Language may provide the mechanism necessary for the reputation tracking that occurs in image scoring models of indirect reciprocity:

> *"The overriding idea, relevant to human societies, is that information about another player does not require a direct interaction but can be obtained indirectly either by observing the player or by talking to others. The evolution of human language as a means of such information transfer has certainly helped in the emergence of cooperation based on indirect reciprocity" (Nowak and Sigmund, 1998).*

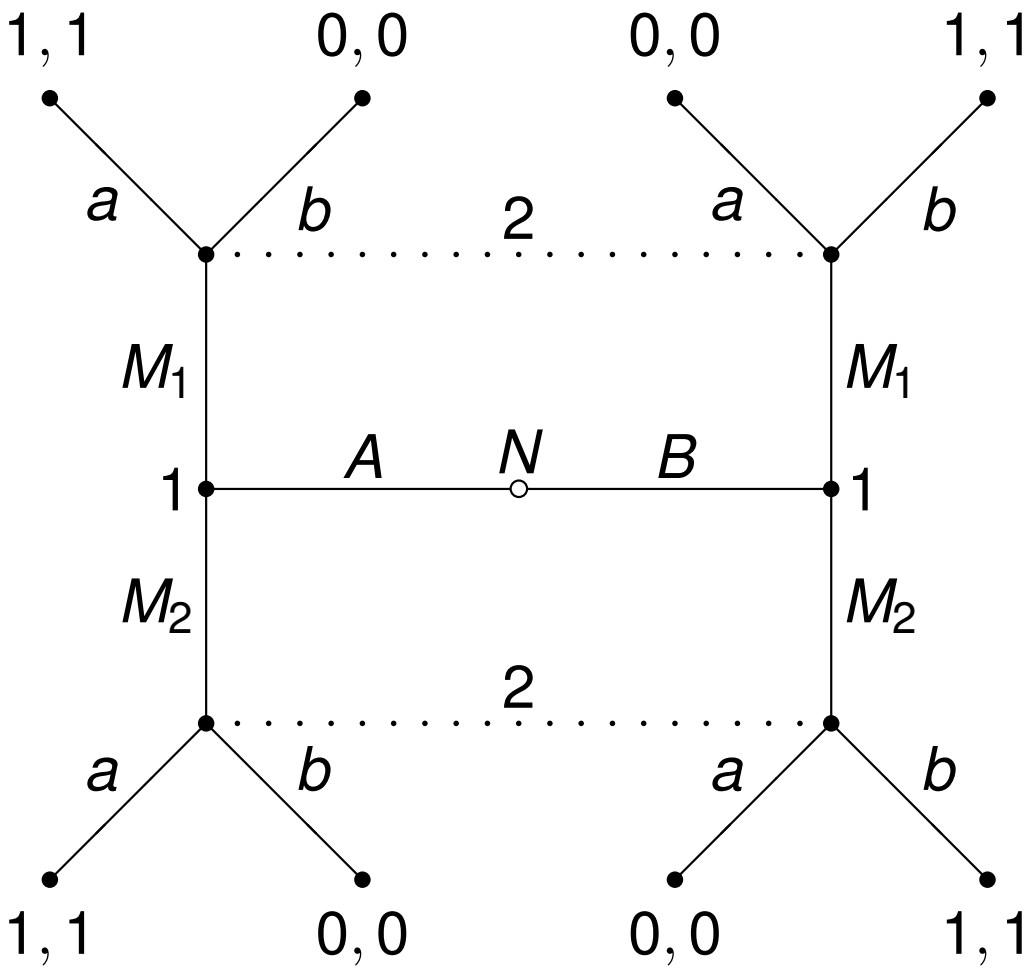
# The Role of Language in Indirect Reciprocity

*"Once something like language begins to evolve, it can be incorporated within the [evolutionary] feedback loop as an instrument of both monitoring and control, for gossip serves both functions. Language is superbly adapted for social monitoring" (Sterelny, 2003).*

*"Indirect reciprocity requires information storage and transfer as well as strategic thinking and has a pivotal role in the evolution of collaboration and communication" (Nowak and Sigmund, 2005).*

Nakamaru and Kawata (2004)

# The Lewis Sender-Receiver Game



# The Model

Members a randomly mixing population will play a series of 1-shot PD's. This will be a fixed numer of rounds $N$.
A strategy is an ordered pair of functions $(R, S)$.

- $R$ maps signals to actions: $\{0, 1\} \rightarrow \{c, d\}$
- $S$ maps actions to signals: $\{c, d\} \rightarrow \{0, 1\}$
- We can represent strategies as ordered quadruples: $(c, d, 0, 1)$ means do $c$ if opponent has a 0 image, do $d$ otherwise, send 0 if opponent plays $c$ and send 1 otherwise.

Each player $i$ in a round $n$ also has an image score $k$ which is determined by the signal of their previous opponent $j$:

$$k_i{}^n = S_j(R_i(k_j{}^{n-1}))$$

# The Model

When each pair meets, each responds to the image of their opponent and receives a payoff $u$ and then, after interaction, gives their opponent a new image. For simplicity, we will assume that all players have an initial image of $k = 0$.

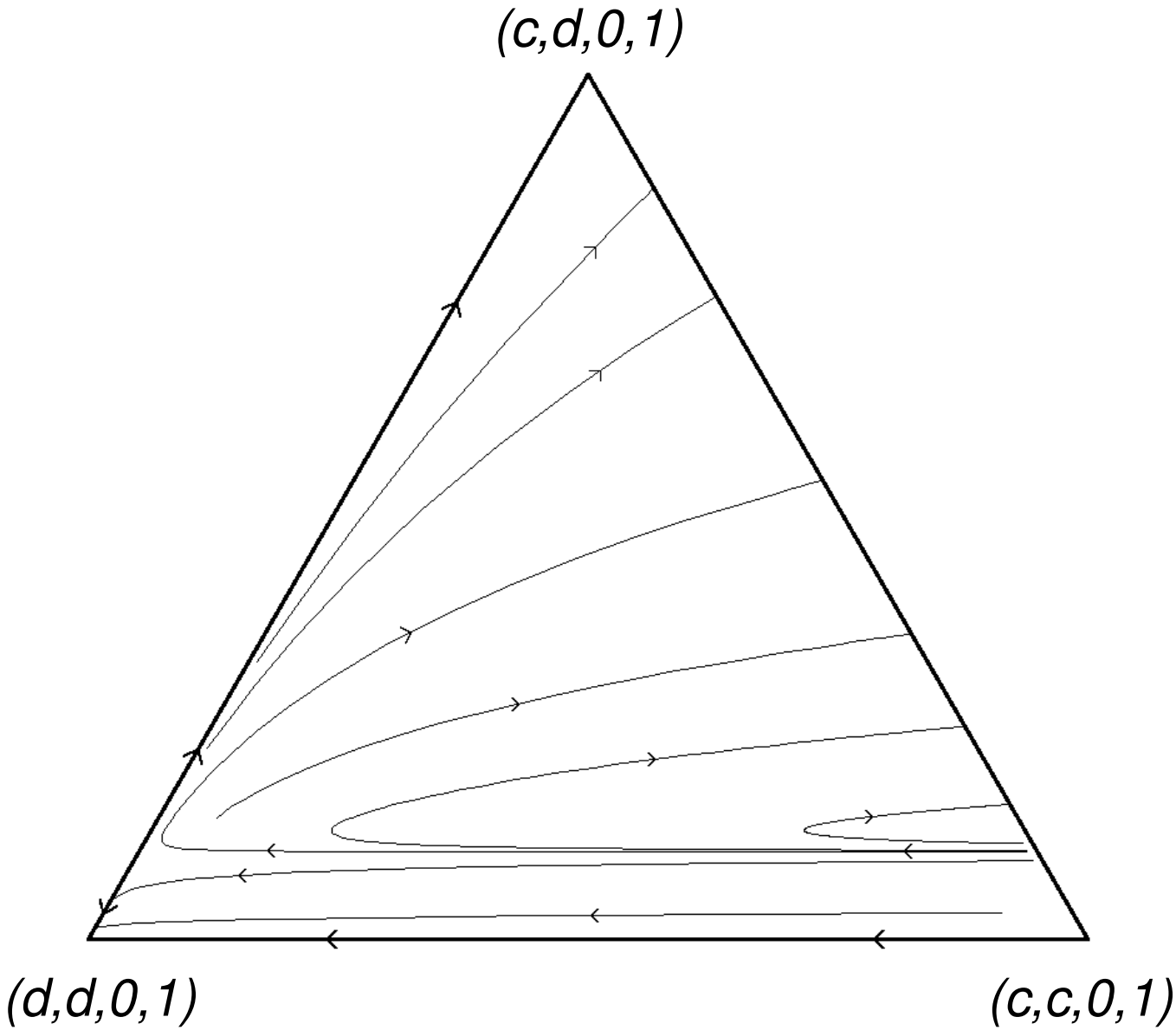
The fitness score for $i$ in round $n$ is based on payoffs against the population for that round.

$$f_i^n(X) = \sum_{j \in Strat} u_i(R_i(k_j^{n-1}), R_j(k_i^{n-1}))x_j$$

The total fitness for the game $f_i(X)$ is the sum of the fitness across all rounds of play. In simulations, evolution occurs according to the discrete time replicator dynamic:

$$x_i' = x_i \left( \frac{f_i(X)}{\theta(X)} \right)$$

# The Benchmark Case



# Simulation Results

| Strategies | % Cooperative |
|---|---|
| $(c, c, 0, 1)$, $(c, d, 0, 1)$, $(d, d, 0, 1)$ | 79% |
| $(c, c, 0, 1)$, $(c, d, 0, 1)$, $(d, d, 0, 1)$, $(d, c, 0, 1)$ | 39% |
| $(c, c, 0, 1)$, $(c, d, 0, 1)$, $(d, d, 1, 0)$ | 48% |
| All Strategies | $< 1\%$ |
| $(c, c, 1, 1)$, $(c, d, 0, 1)$, $(d, d, 0, 1)$ | 0% |

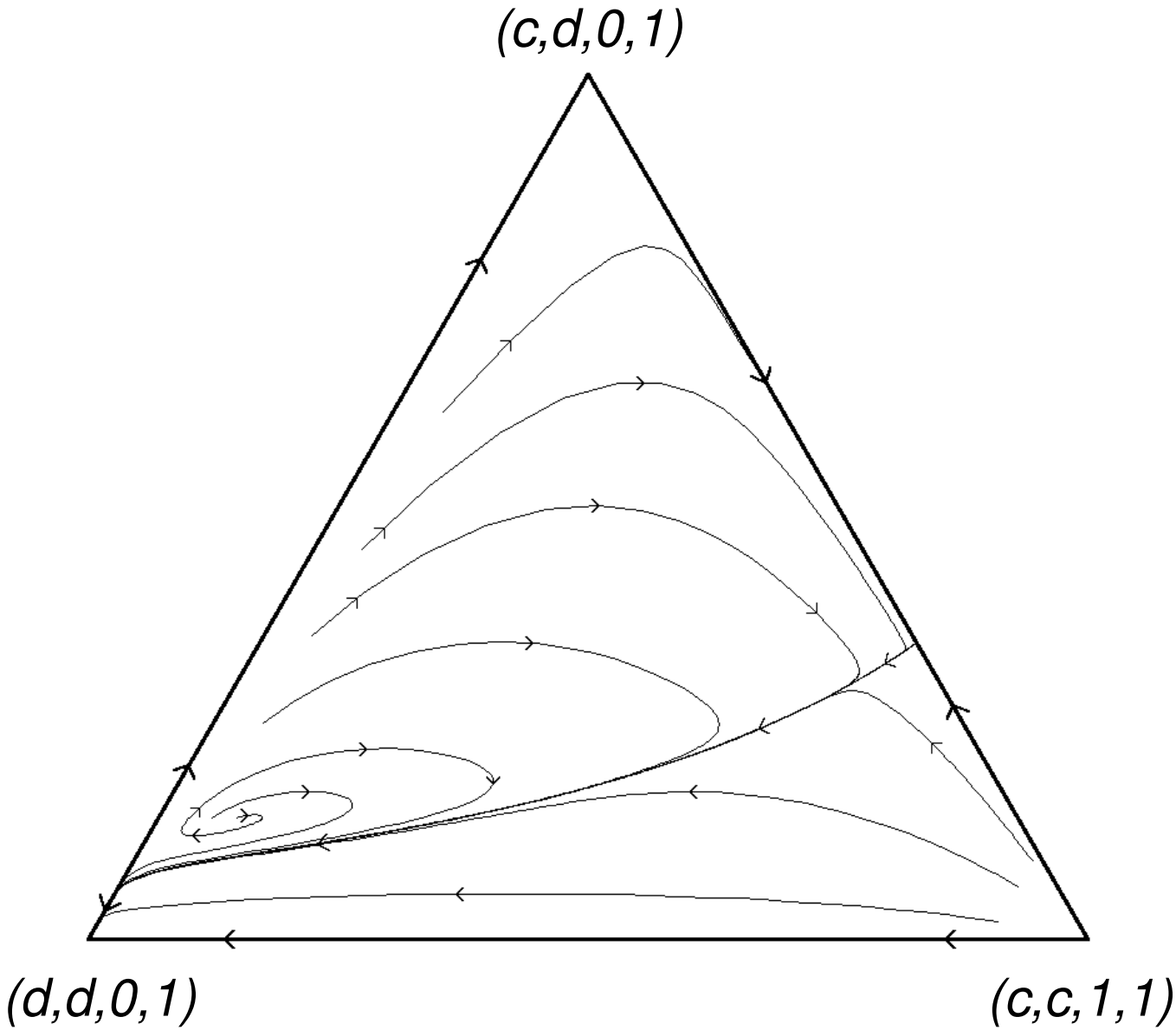
# Simulation Results

| Strategies | % Cooperative |
|---|---|
| $(c, c, 0, 1)$, $(c, d, 0, 1)$, $(d, d, 0, 1)$ | 79% |
| $(c, c, 0, 1)$, $(c, d, 0, 1)$, $(d, d, 0, 1)$, $(d, c, 0, 1)$ | 39% |
| $(c, c, 0, 1)$, $(c, d, 0, 1)$, $(d, d, 1, 0)$ | 48% |
| All Strategies | $< 1\%$ |
| $(c, c, 1, 1)$, $(c, d, 0, 1)$, $(d, d, 0, 1)$ | 0% |

It appears that the prospects for the co-evolution of indirect reciprocity and "moral signals" are dim.

The "two-faced" strategy $(c, c, 1, 1)$ is particularly problematic.

# The Two-Faced Strategy



# Adding Pressure on the Signaling Strategies

For the co-evolution of "moral signals" and indirect reciprocity there will need to be some kind of pressure on the signaling strategies.

*Exogenous Benefit:*

- Individuals receive a bonus $x$ to signaling "correctly" (using 0 for cooperators and 1 for defectors).
- In addition to playing cooperation games, players engage in a Lewis sender-receiver game and receive an additional payoff $y$.

# Adding Pressure on the Signaling Strategies

For the co-evolution of "moral signals" and indirect reciprocity there will need to be some kind of pressure on the signaling strategies.

*Exogenous Benefit:*

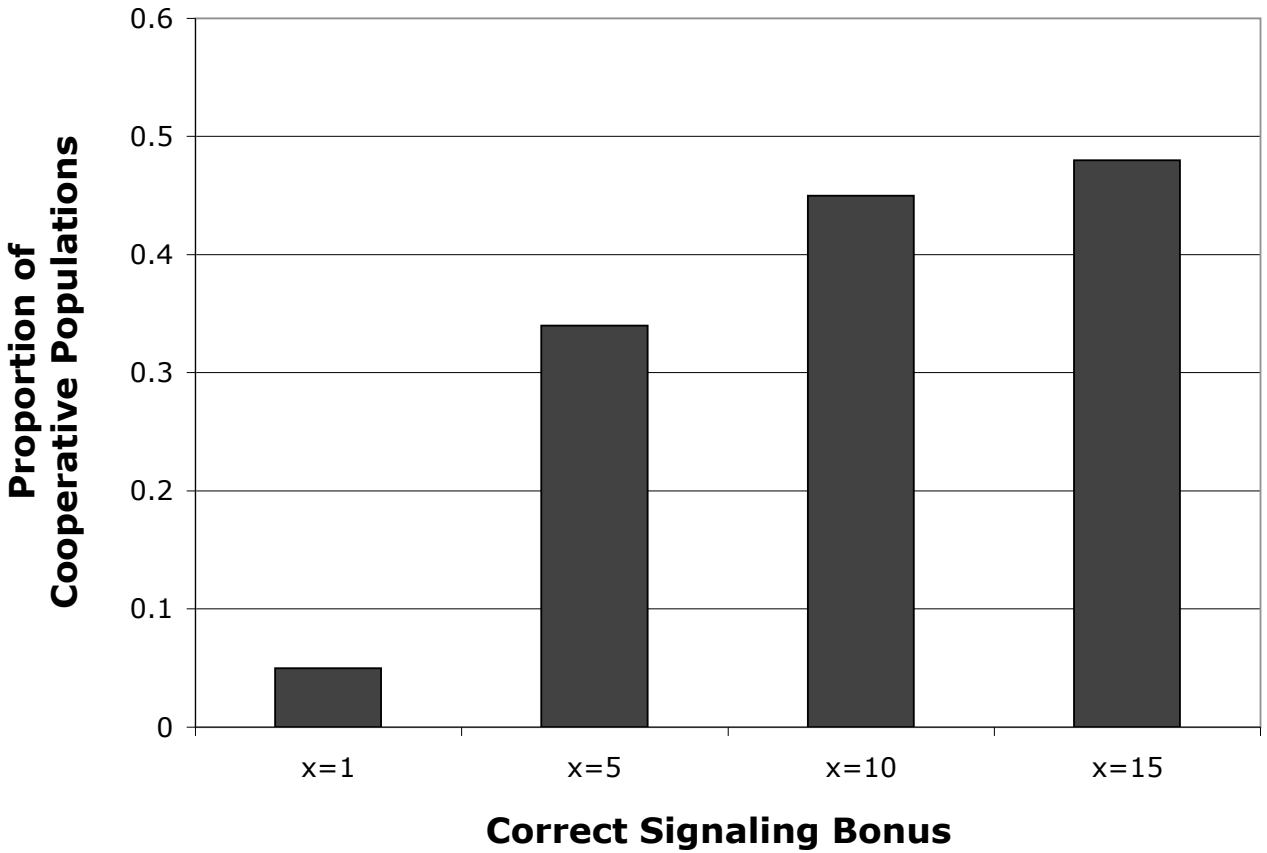
- Individuals receive a bonus $x$ to signaling "correctly" (using 0 for cooperators and 1 for defectors).
- In addition to playing cooperation games, players engage in a Lewis sender-receiver game and receive an additional payoff $y$.

*Endogenous Benefit:*

- Incorporate signaling into the same cooperative game setting, allowing "punishment" for deviant signals.

# Bonus to “Correct” Signaling

Including pressure $x$ (independent of the indirect reciprocity setting) to conform to an already established signaling system.



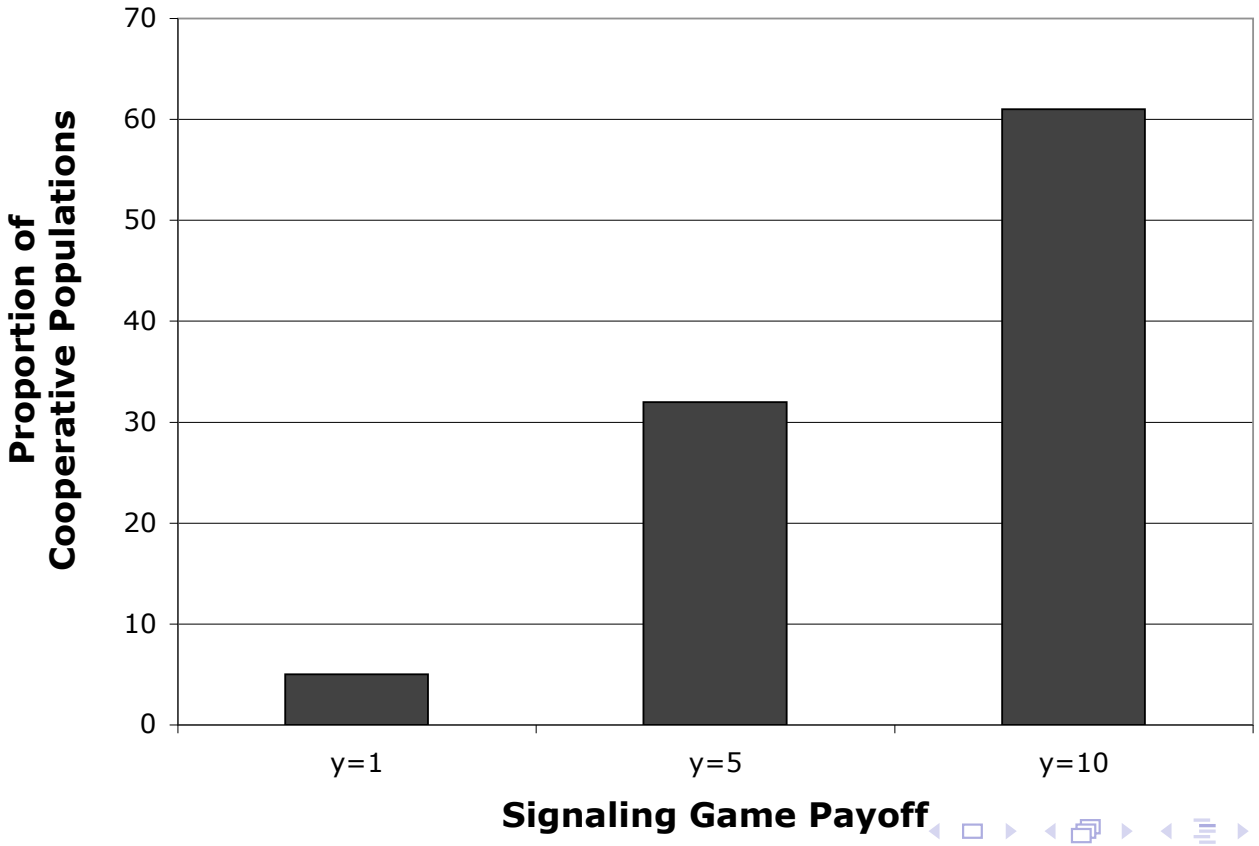
# An “Entangled” Signaling Game

Agents may be unable to distinguish between image signaling and another more direct signaling. Thus, each individual’s strategy in the indirect reciprocity setting doubles as a strategy for a simple signaling game.

- Two possible states of the world $A$ and $B$ that occur with equal probability.
- One player observes and sends a signal 0 or 1.
- The other player observes and acts accordingly with $c$ or $d$.
- If $c$ is chosen in state $A$ each get a payoff $y$ and if $d$ is chosen in state $B$ each get a payoff $y$, otherwise each receive no payoff.

# An “Entangled” Signaling Game

There are two types of cooperative populations that can evolve: all $(c, d, 0, 1)$ and polymorphic mixes involving $(c, c, 1, 0)$ and $(d, c, 1, 0)$.



# Endogenous Benefit for Signals

With the aim of providing an *endogenous* benefit to accurate communication we can look at a variation on the previous model with a new complication:

- With probability $q$ an interaction (between two agents $A$ and $B$) and sent signal (by $B$) are observed by another member of the population ($C$).
- $C$ may then signal regarding $B$, changing the image of that agent.
- If $B$'s signal matches what $C$ would have sent regarding $A$, $C$ labels $B$ whatever corresponds to a cooperative action. And, if $B$'s signal did not match, $C$ labels $B$ whatever corresponds to a defective action.

# Endogenous Benefit – Results

This way of making the benefits to signaling endogenous, works to eliminate troublesome strategies such as the Two-Faced cooperator:

- The cooperative outcomes that do emerge are a mix between $(c, d, 0, 1)$, $(c, c, 0, 1)$, $(c, d, 0, 0)$, and $(c, c, 0, 0)$ (they are always optimal).

# Endogenous Benefit – Results

This way of making the benefits to signaling endogenous, works to eliminate troublesome strategies such as the Two-Faced cooperator:

- The cooperative outcomes that do emerge are a mix between $(c, d, 0, 1)$, $(c, c, 0, 1)$, $(c, d, 0, 0)$, and $(c, c, 0, 0)$ (they are always optimal).
- It is possible for a image-tracking signaling system to evolve in the context of indirect reciprocity and for that same system to provide the mechanism for signaling enforcement.

# Endogenous Benefit – Results

This way of making the benefits to signaling endogenous, works to eliminate troublesome strategies such as the Two-Faced cooperator:

- The cooperative outcomes that do emerge are a mix between $(c, d, 0, 1)$, $(c, c, 0, 1)$, $(c, d, 0, 0)$, and $(c, c, 0, 0)$ (they are always optimal).
- It is possible for a image-tracking signaling system to evolve in the context of indirect reciprocity and for that same system to provide the mechanism for signaling enforcement.
- When all the strategies are included, the proportion of resulting populations acheiving a cooperative outcome is $< 2\%$.
- Mixed Results: *possible* but not likely.

# Conclusions

- The co-evolution of signaling and indirect reciprocity is far from straight-forward and the most direct method of modeling such co-evolution reveals that some method of "policing" the use of signals will be needed.

# Conclusions

- The co-evolution of signaling and indirect reciprocity is far from straight-forward and the most direct method of modeling such co-evolution reveals that some method of "policing" the use of signals will be needed.

- By including an endogenous benefit to signaling it is possible to co-evolve indirect reciprocity and a signaling system serving as a basic reputation-tracking mechanism.

# Conclusions

- The co-evolution of signaling and indirect reciprocity is far from straight-forward and the most direct method of modeling such co-evolution reveals that some method of "policing" the use of signals will be needed.
- By including an endogenous benefit to signaling it is possible to co-evolve indirect reciprocity and a signaling system serving as a basic reputation-tracking mechanism.
- However, without some additional exogenous benefit to signaling correctly, it seems unlikely that such co-evolution would occur.

# Conclusions

- The co-evolution of signaling and indirect reciprocity is far from straight-forward and the most direct method of modeling such co-evolution reveals that some method of "policing" the use of signals will be needed.
- By including an endogenous benefit to signaling it is possible to co-evolve indirect reciprocity and a signaling system serving as a basic reputation-tracking mechanism.
- However, without some additional exogenous benefit to signaling correctly, it seems unlikely that such co-evolution would occur.
- When such exogeneous benefits are included, we can see the evolution of different moral systems with different "moral signals."