Tit for Tattling: Cooperation, communication, and how each could stabilize the other

by

Victor Vikram Odouard and Michael Holton Price
Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501

Presentation by Luizo :)

https://ncase.me/trust/

The evolution of altruism

- Altruism → Helping others at a personal cost
- All explanation for the evolution of altruism ensure that altruists receive a second-order benefit that compensates the initial cost
- Indirect reciprocity → individuals have their favors returned even by third parties
- **Communication** (in indirect reciprocity) is required to disseminate reputation

Under what conditions does the interaction between signaling and cooperation stabilize both high levels of (altruistic) cooperation and truthful, informative, and effective communication?

Cooperation, communication, stability

- The agents evolve rules to
    1. Act in a prisoner’s dilemma
    2. Communicate about the actions of others
- Under the conditions that allow for a stable cooperative-communicative state, agents
    1. act and signal according to an aligned *norm*
    2. occasionally deviate from their strategy
    3. exert normative pressure on each other's signals → favors truthfulness

Background, Cooperation

- **Strategy** → set of rules for acting and signaling
- **Norm** → strategy that encodes a set of rules that is followed by the large majority of a social group
- **Standing** → defined recursively: some-one is in good standing if they cooperated, or if they defected against someone in bad standing, and to cooperate with agents iff they are in good standing
- **Stern judging** → similar to standing, additionally puts those who cooperate with those in bad standing into bad standing.

Background, Communication

- The stability of honest communication systems usually rely on some kind of pressure on the signaler to be truthful
- In its basic form, communicating the reputations of others places no such pressure on the signaler
- Truthfulness is not the only requirement: agents also need to be forthcoming (actually share the information they have)
- The stable strategy deals with it by treating failures to signal in the same way it treats lying

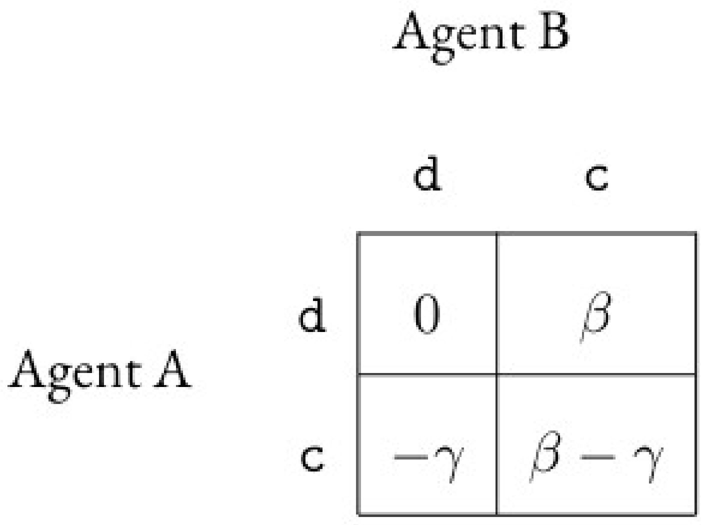
Communication

- **Communication system** → set of mappings from meanings to symbols
- A communication system is **effective** when it is→
  - **Uniform** → everyone abides by the same mapping from meanings to symbols (corresponds to truthfulness)
  - **Forthcoming** → agents never fail to signal when they possess relevant information
  - **Informative** →everyone's mapping distinguishes between at least two meanings (guarantees that some information gets transmitted)

Cooperation and Stability recalls

- Cooperation is paying a cost, *γ*, for the benefit, *β*, of someone else

- Agent A's payoff in the prisoner's dilemma. **d** stands for defect and **c** stands for cooperate.



		Agent B	
		d	c
Agent A	d	0	β
	c	$-\gamma$	$\beta - \gamma$

- A strategy is **stable** if no other strategy can invade it
- That is, no other strategy can proliferate in a population of 100% A-type agents (when A is said to 'predominate').

Model 1: A first pass

- Infinite number of rounds, agents interact in prisoner’s dilemma
- Agents “tag” each other, each agent has one tag at a time
- Each round everyone pairs up at random and makes a choice (cooperate/deflect). An agent’s choice may depend on their partner’s current tag.
- Everyone signals a new “tag” for their partner (This process can be imagined as your partner writing a 0 or 1 on your forehead for your next partner to see, as the signal need only be known to one’s next partner)
- Each agent can use whatever mapping they desire

Model 1: Strategies

- We want to find the **focal strategy**, that is, a strategy that leads to a *stable, highly cooperative and effective communicating population*.

Action Strategies

Name	ID	Act for 0s	Act for 1s
Deflector	dd	d	d
Discriminator	dc	d	c
Reverse-discriminator	cd	c	d
Cooperator	cc	c	c

Signal of actor's new tag

Partner's tag	Actor's action	Image scoring	Stan ding	Stern Judging
0	d	0	1	1
0	c	1	1	0
1	d	0	0	0
1	c	1	1	1

Model 1: Focal strategy

- Stern discriminators as candidates to analyze for this model
  1. Action: coop with 1s, defects with 0s
  2. Signaling: tags cooperation with 1-agents and defection with 0-agents with a 1, and cooperation with 0s and defection with 1s with 0.
- Analysis
  - Is the state *cooperative?*
  - Does the state *effectively communicate?*
  - Is it *stable*?

Model 1: Analysis

- Is the state *cooperative?*

  All who follow the stern discriminator strategy receive a tag of 1, so everyone will receive a 1-tag to start the second round, and thus, everyone will cooperate in the second round and so on. All rounds except the first are fully cooperative

- Does the state *effectively communicate?*

  all agents abide by a separating, uniform mapping (as specified by the stern judging norm), and never withhold information

- Is it *stable*?
  - No payoff difference between individuals with the same action strategy but **different languages** (a ‘pushover’ discriminator that signals 1 about everyone does just as well as a stern discriminator)
  - **Some traits are unexpressed** (eg. Everyone cooperates, everyone receives a tag of 1)
  - A different strategy that exploit either of the above can invade

Model 2: Next!

- Addressing instabilities
  - Unpunishability of language → ***Meta-signaling*** tags agents based on their signal
  - Unexpressed traits → ***Error*** creates diversity
- Each round agents are either *actor* or *observer*
- Observers are assigned randomly (to actors or another observer)
- Actors interact in a prisoner’s dilemma, using their strategies with error rate $\boldsymbol{\epsilon}$ and signal by tagging each others with error rate $\boldsymbol{\delta}$
- Observers ‘meta-signal’ whether they agree or disagree with the obervee’s signal, and tag the accordingly to their meta-signaling strategy with error rate δ, overwriting the observee’s tag.

Model 2: Analysis

- Focal strategy → Stern discriminator, meta-signaling strategy: 0 if they disagree, 1 if they agree.
- Is the state *cooperative?*
- Does the state *effectively communicate?*

  Because of errors there are a nonzero portion of defections, but that portion can be made arbitrarily small by shrinking ϵ and δ.

- Is it *stable*?

  We need to ensure that:

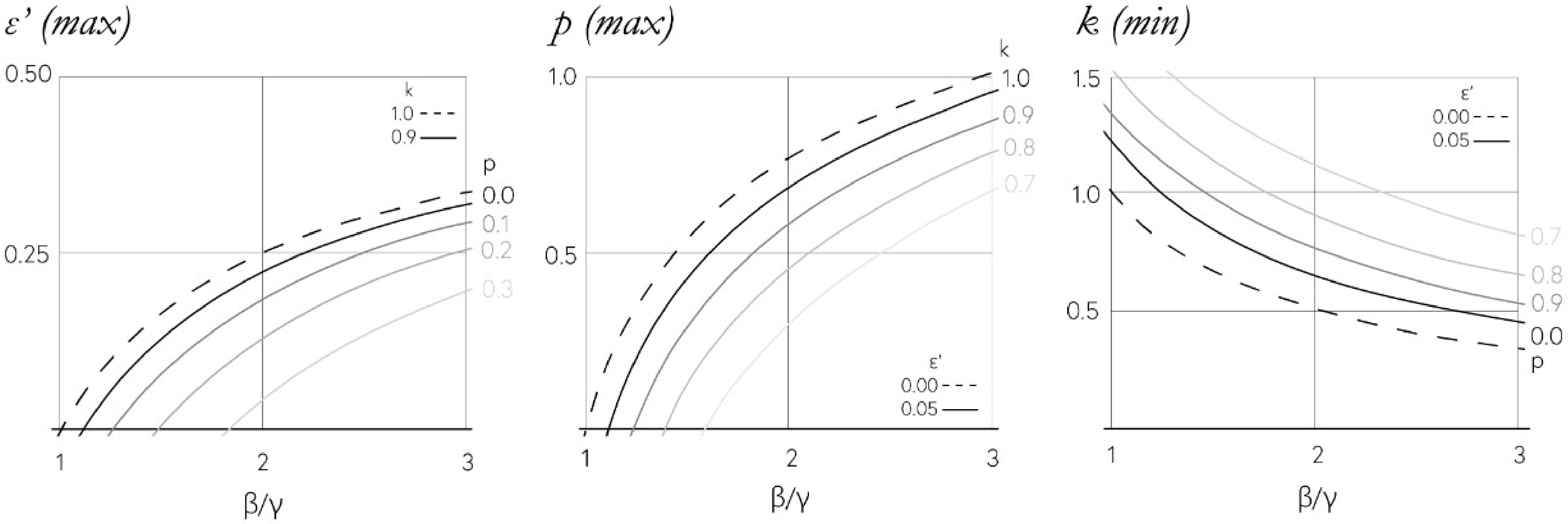
  - Any strategy that deviates from the stern discriminator does strictly worse
  - No latent traits exist that will cause invading strategies to be indistinguishable from stern discriminators.

Model 2: Any alternative strategy does worse

- Whenever an agent makes any decision there are three possible consequences of that decision:

1. **first-order costs** → immediate costs paid for a decision (eg. paying γ for cooperating)
2. **second-order costs** → costs borne for a decision in the subsequent action round (eg, their subsequent partner defects with them for having a 0-tag)
3. **effects on future scenarios** → the present decisions of agents may affect the conditions surrounding their future decisions, potentially affecting payoffs.

Model 2: Values for strategy stability



- Effective error rate **ϵ′** → combination of ϵ and δ
- Probability of being an observer **p**
- Discounting factor **k** → determines the present value of a series of future cash flows

Model 2: All traits are expressed with error

- Any strategy that deviates from the stern discriminator does strictly worse
- No latent traits exist that will cause invading strategies to be indistinguishable from stern discriminators. (   )
- There are three safeguards against the problem of unexpressed traits.

1. Any bounded world state comes to pass (assuming infinite time), due to **error**
2. Minimal relevant information is available to an agent in any given round, paring down the plausible world states to a set whose members are all reasonably likely to occur.
3. Rare world states are extremely difficult to exploit because they rely on the confluence of deviations of many agents, which is very unlikely to occur.

Discussion: Meta-signaling

- The concept of meta-signaling makes it clear why it makes sense to study the evolution of communication in the context of cooperation.
- It’s what makes altruism conventionally stable
- Meta-signaling co-opts the indirect reciprocity mechanism for the purposes of maintaining the communication system
- The core innovation was repurposing a technology for the analogous purpose of maintaining itself, saving the effort of maintaining a separate system.
- it sets up a positive feedback loop: enhancements to the system lead to further enhancements

Discussion: Norms & Equilibrium

- Communicating moral judgments (good/bad) simplifies complex behavior by collapsing detailed chains of actions into simple moral categories.
- Instead of tracking who cooperated or defected through complex histories, agents only need to know whether someone’s action aligns with a “good” or “bad” tag.
- Stern discriminators outperform all other strategies, thus making the equilibrium robust
- Although there could be stable but defect-oriented states, group selection would favor cooperative groups due to low defection tendencies.

Discussion: Social enforcing & Reputation

- Social enforcement of truthful signaling has two main advantages
    - it can flexibly set costs to maintain truthfulness
    - it’s resistant to destabilization by selection pressures
- Reputation is crucial across social interactions, helping agents decide whom to cooperate with and how to administer punishment.
- This reputation system evolved to facilitate cooperation but has since been adapted for third-party punishment, enabling social order in complex groups.

Conclusions

- This study identifies three stability conditions for a basic communication system:
  - Meta-signaling
  - error tolerance
  - stern judging norms
- The logical next research step would be about Alternative equilibria, exploring alternative pathways or Complex communication, investigating more elaborate language systems (reinforced learning, cognitive limits).