Retaliation, Punishment and Sanction.
Cognitive Modelling and Experimental Data*

Rosaria Conte

LABSS
(Laboratory for Agent-Based Social Simulation)
Institute of Cognitive Sciences and Technologies
Italy

SINTELNET Workshop – May 31- June 01, 2012, Toulouse

*With the contribution of the FuturICT Coordination Action, and
The SEMIRA FET-funded project under FP7
Outline

The puzzle of cooperation and social order
  – Strong reciprocity: open questions
  – Punishment is far from a homogeneous phenomenon

A social and cognitive model of:
  – retaliation
  – punishment
  – sanction

Cross methodological experiments
  – natural
  – Artificial

Concluding remarks

Future work
Punishment and Social Order

“Ethnographic evidence, evolutionary theory, and laboratory studies indicate that the maintenance of social norms typically requires a punishment threat, as there are almost always some individuals whose self-interest tempts them to violate the norm. [Spitzer et al. 2007]

• What is this predisposition? How did it appear and evolve?

• How does this theory account for
  – proximate mechanisms of punishment?
  – Punishment is usually considered homogeneous [see Elster, 1989; Durkheim, 1893; Foucault, 1975; Ostrom et al., 1992; Gintis, 2000; Fehr, Gachter, 2000],
  – Poor attention on the cognition behind different kinds of reactions [Carlsmith, Darley, Robinson, 2002; Falk, Fehr, Fischbacher, 2001]

“Cooperation is maintained because many humans have a predisposition to punish those who violate group-beneficial norms, even when this reduces their fitness relative to other group members” (Bowles & Gintis, 2003)

• A cognitive model of reaction to damage is needed in order to
  – identify and model the cognitive underpinnings of different reactions to aggression
  – draw evolutionary trajectory from retaliation to punishment and sanctioning
  – demonstrate that high level cognitive systems are pivotal to the evolution of enforcement institutions
Related Work

Fehr and Gachter, 2000

Yamagishi, 1986

Horne, 2009

Herrmann et al., 2008
The whole history of punishment and its adaptation to the most various uses has finally crystallized into a kind of complex which is difficult to break down and quite impossible to define. [Nietzsche, *The genealogy of morals*, 1956, p. 212]

Punishment is far from homogeneous...

Our aim is to break down this complex into three specific behaviors:

- Revenge
- Punishment
- Sanction
Cognitive premises

Social behaviour is often based on

- **Mindreading** - $B_x(MS_y)$ - (ToM, Premack & Woodruff, 1978, etc.)

Sometimes aimed at

- **Mindchanging** - $G_x(MS_y)$ - modify mindstates (beliefs, goals, emotions). In particular,

  - what leads to further structures: $G_x((B_y) \rightarrow (G_y))$
Reactions to damage

• Different mental configurations at different levels of complexity can be paired with distinct reactions, e.g. revenge, punishment and sanction
  – What is the difference, if any?
  – What about preference order?
  Why is revenge more deprecated than any other?
  – What are specific effects?
Cognitive and social dimensions
(Giardini, Andrighetto, Conte, CogSci, 2010)

- **Time perspective** (whether backward Vs forward-looking)
- **Cognitive configuration** (change Other’s mind Vs restore One’s power conditions)
- **Social relationship** (Dominance Vs equality)
- **Social structure** (dyadic Vs triangular)

Defining the specific mental configurations behind each reaction allows us to discriminate among actions that are only apparently similar.
Time perspective

- Revenge is **backward-looking**: avenger wants to restore power balance by injuring the victim.

  “You have killed my son, so I killed yours; I have taken revenge for that, so I now sit **peacefully** in my chair” [old tribesman from Montenegro, Boehm, 1984]

- Punishment and Sanction are **forward-looking**: aimed at deter further attacks from target

  Neuroscientific evidence [deQuervain et al., 2004; Knutson, 204] shows that punishing a violator activates brain regions related to the anticipation of a reward
Mental configuration

Cognitive influencing drove (exaptation?) towards enforcing
- *future* behaviour (deterrent systems), based on
  - *respectful,* even
  - *impersonal* will (norm).

<table>
<thead>
<tr>
<th></th>
<th>$G_x$ (Damaged $y$)</th>
<th>$G_x((B_y) \rightarrow (G_y))$</th>
<th>$G_x((N_B_y) \rightarrow (N_G_y))$</th>
</tr>
</thead>
<tbody>
<tr>
<td>R</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>P</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>S</td>
<td>X</td>
<td></td>
<td>X</td>
</tr>
</tbody>
</table>

- *P and S are aimed at modifying the mental states of the target to influence her actions*  
- In P and S
  - The goal of *deterring* T (and possibly O) from further hostility  
  - Hence, *cognitive influencing* [Posner, 1980; Becker, 1968; Bandura, 1991]
    - Belief: “I will sustain a cost, if I again”  
    - Goal: “Abstaining from further attacks”
Social dominance

- R does not imply dominance
- P and S do
  - Punishment: showing and maintaining dominance over the target [Clutton-Brock and Parker, 1995; Dreber et al. 2008]
Social structure

• Whereas R and P are dyadic relationships between 2 parties (symmetrical or not),
• S is a 3-party relationship: the Sanctioner wants the target to
  – believe that
    • She violated a norm, ie.,
      – a behaviour spreading over P to the extent and because the corresponding prescription spreads as well (Ullman-Margalit, 1977)
      – a normative prescription is a command that *pretends* to be adopted for its own sake, because it *ought* to be observed (Conte et al., 2009) (Conte, Andrighetto, Campenni, 2012)
    • Norm violation caused cost imposition
    • Sanctioner acted to defend the norm
  – Want to
    • abstain from future violations in order to
    • comply with the norm
Summing up

<table>
<thead>
<tr>
<th>Perspective</th>
<th>R</th>
<th>P</th>
<th>S</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mindchanging</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Dominance</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Social structure</td>
<td>2-p</td>
<td>2-p</td>
<td>3-p</td>
</tr>
</tbody>
</table>

Punishment between Retaliation and Sanction
Let us check the validity and utility of the model. Here, we check the difference between P and S.

• Is there evidence supporting our model?
• showing their respective effect?
Limits of Punishment

- Punishment as a cost inflicted to the target (Boyd et al., 2010) is
  - not a linear function of its severity (Sonzogni, Cecconi and Conte, 2010; Helbing et al., 2010).
  - What is more punishment may have detrimental effects [Gneezy & Rustichini, 2000; Fehr and Rockenbach, 2003; Li et al. 2008]
Punishment as signalling

- Experimental evidence [e.g. Hauser and Xiao, 2010] shows that drawing people’s attention on a social norm plays a pivotal role in eliciting compliance.

- Ethnographic evidence suggest that punishment is
  - Often accompanied by communication of disapproval
  - Performed by many

- Why? Hypotheses:
  - Punishment is more efficacious when it is norm-signalling (sanction).
  - Distributed punishment is more effective than individual one for the same value of material damage, because a large number of punishers is interpreted as norm-signalling.

- Tested both with human subjects and in a simulated environment.
Norm-signalling and norm salience

Others’ actions signal that a norm exists and how important it is:
• amount of compliance and cost of compliance
• enforcement typology (private or public, 2nd and 3rd party, punishment or sanction, etc.)
• efforts and costs to educate population, e.g. publicity campaigns;
• credibility and legitimacy of normative source
• surveillance rate, frequency and intensity of punishment.

Let's turn to experiments in which the number of punishers is varied for the same individual cost.
First experiment

Villatoro, D, Andrighetto, G., Brandts, Conte, R., Sabater-Mir. AAMAS 2012)

• Three Treatments:
  – No Punishment.
    • Subjects are not allowed to punish others.
  – Uncoordinated Punishment:
    • Subjects spend a **fixed amount** for punishing others.
  – Coordinated Punishment:
    • Subjects **divide the costs of punishment**

• Classical Experimental Economics Conditions:
  – Iterated Public Goods Game for 40 rounds.
  – Partner Treatment.
  – 40 subjects divided in stable groups of 4.
Hybrid laboratory experiment
(Andrighetto et al, AI Communications, 2010; Villatoro et al., IJCAI 2011)

- 80 subjects distributed in groups of 4 agents, 1 real and 3 confederates.
- Public good game. 3 stages
  - 1° decision: whether contribute
  - 2° decision: whether punish
  - update
- Virtual agents contribute 50% of times, punish 25% in first 10 runs
- From 10° round contribute 90% of the times, non-punishers never punish, punishers act 90% of the times and only if they have cooperated at first stage (the confederate agents mimic the behavioral dynamics observed in humans)
- Punishment inflicts a fixed cost for the punished (his payoff = 0) in any treatment
- Four treatments: 0 punisher, 1 punisher, 2 punishers, 3 punishers.
- When punisher is more than 1 costs for the punishers is shared
Findings

- After first 10 rounds, in the 0 punishers treatment, cooperation collapses.
- Distributed punishment matters even at the same cost for the victim.

When punishers increase, probability of being punished increases but ratio of non-punished defectors remains below 5% in the 1 punisher treatment, around 2% in the 2 punishers treatment, and 0.005% in the 3 punishers treatment.
How to test the norm-signalling hypothesis?


- By a simulation experiment in which
  - Normative agents EMIL-As play
  - With pre-programmed agents of the same type as in the hybrid experiment
  - Same design and treatment

Emotional component?

NORM RECOGNITION: N-BELIEF

Epistemic Component with salience update

Pragmatic component

NORM ADOPTION: N-GOAL

NORM DECISION: N-INTENTION

EMIL-Agent (EMIL project: [www.emil.istc.cnr.it](http://www.emil.istc.cnr.it))
Simulation findings
Norms Vs reinforcement learning
Summing up results

Sanction is more effective than punishment in
1. achieving cooperation;
2. reducing the costs for social order to be achieved and maintained and
3. making cooperation stable and resilient to environmental change - e.g. an abrupt interruption of the enforcement mechanism
Future work

• Experimenting in a real-world scenario
• Retaliation
• Emotions
Thanks to Collaborators from the LABSS and elsewhere...

Giulia Andrighetto
Barbara Sonzogni
Federico Cecconi
Francesca Giardini
Daniel Villatoro
Jordi Sabater
Jordi Brandts
Marco Campennì
... and thank you!