

Dynamic Sanctioning for Robust and Cost-Efficient Norm Compliance

Daniel Villatoro, Giulia Andrichetto, Jordi Sabater-Mir and Rosaria Conte

Why study Social Norms?

- ▶ “The social norms I am talking about are not the formal, prescriptive or proscriptive rules designed, imposed, and enforced by an exogenous authority through the administration of selective incentives. I rather discuss **informal norms that emerge through the decentralized interaction of agents within a collective and are not imposed or designed by an authority.**”

[Bicchieri 2006]

- ▶ Self-organization in heterogeneous soc.: virtual & humans



12 M users (Population of Portugal)



22 M users (Population of Australia)



950 M users (Population of Europe, including Russia)

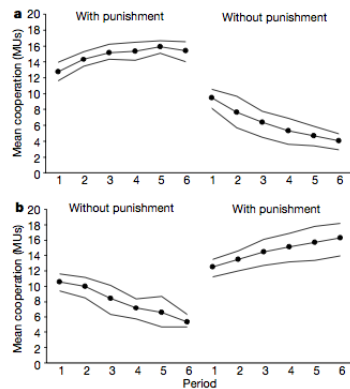
Punishment and Social Order

- ▶ “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.

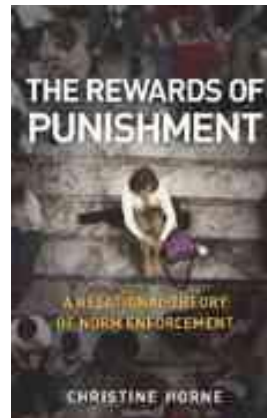
Norm obedience sometimes *vanishes* quickly *in the absence* of a credible *punishment threat*”

[Spitzer et al. 2007]

Related Work



Fehr and Gächter, 2000



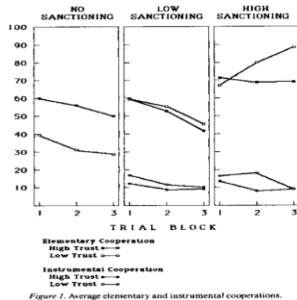
Horne, 2009



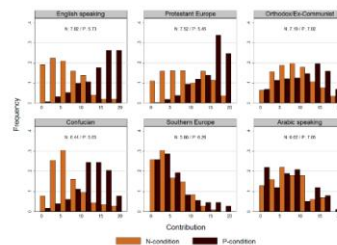
REVIEW ARTICLE

Punishment in animal societies

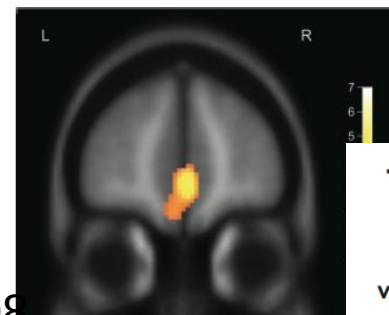
T. H. Clutton-Brock & G. A. Parker



Yamagishi, 1986



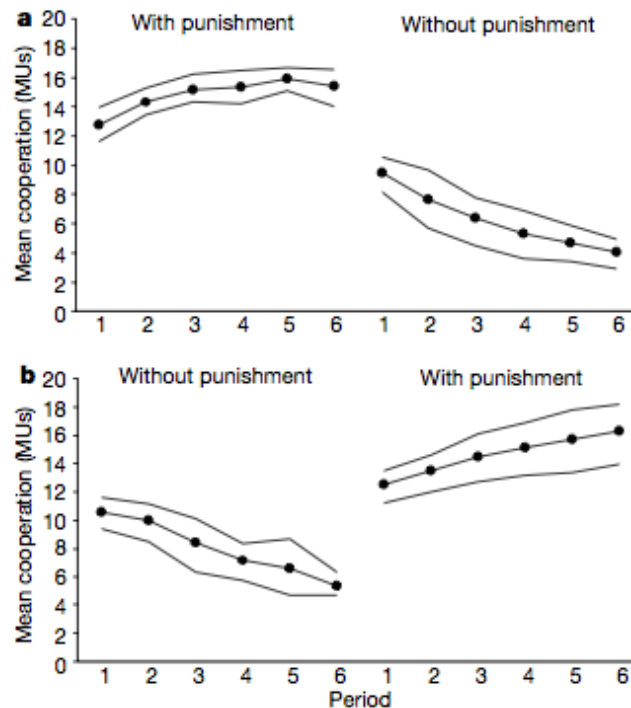
Herrmann et al., 2008



The Neural Basis of Altruistic Punishment

Dominique J.-F. de Quervain,^{1*†} Urs Fischbacher,^{2*} Valerie Treyer,³ Melanie Schellhammer,² Ulrich Schnyder,⁴ Alfred Buck,³ Ernst Fehr^{2,5†}

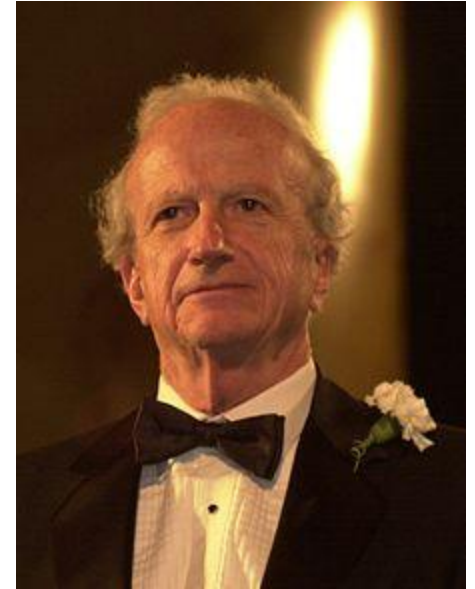
Cooperation and Punishment in Public Goods Experiments



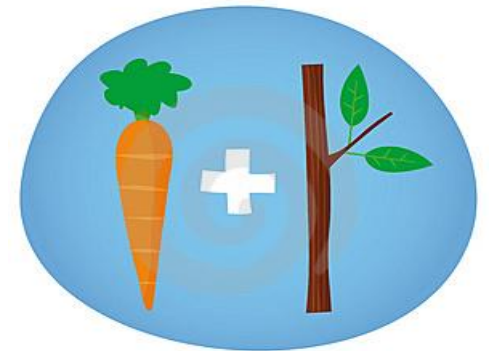
[Fehr and Gächter 2000]

Carrot and Stick Approach to Punishment

Individuals obey or break the norm depending on the price of violation, i.e. the severity of punishment discounted by the probability that it will be imposed [Becker, 1968].

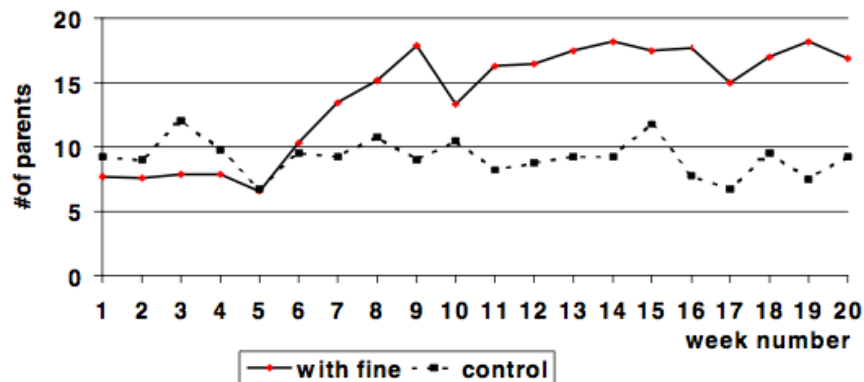


Utilitarian framework:
human action is **instrumentally oriented and utility-maximizing**



The Carrot and Stick Approach to Punishment Under Attack

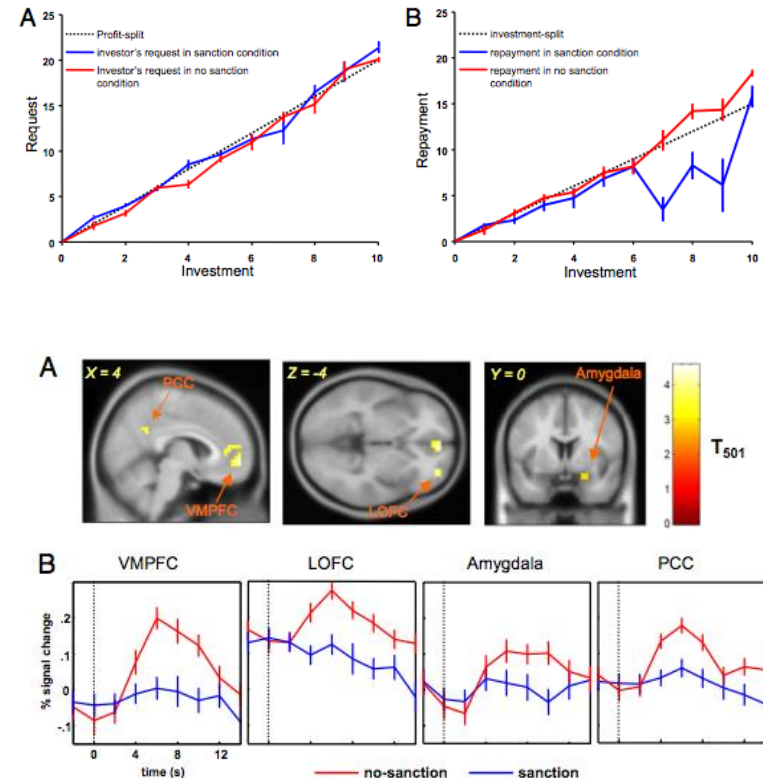
Experimental and real world evidence show that in some circumstances, punishment has no effect or even worst a **detrimental effect** [Gneezy & Rustichini, 2000; Fehr and Rockenbach, 2003; Li et al. 2008]



Gneezy &
Rustichini, 2000

Neuroscientific Evidence

- Trustees reciprocate less when facing sanction
- Presence of punishment:
 - reduces trustee's brain activities involved in social evaluation: ventromedial prefrontal cortex; lateral orbitofrontal cortex and amygdala
 - increases brain activities in the parietal cortex, implicated in rational decision making



Li et al. 2008

The signalling power of punishment

Depending on how it is designed and implemented, punishment conveys distinct messages which frame the decision environment in different ways.

When implemented merely as the imposition of a cost



Utility-driven decisions

Sanction

In addition to imposing a cost for the wrongdoing, sanction intentionally **signals** that this conduct is not approved of because it violates a social norm

Scolding, indignation, blame have the effect of focusing people's attention on a) the existence and violation of a norm; b) the high rate of surveillance; c) causal link between violation and sanction; d) sanctioner is a norm defender; etc.

Utility-
driven
decisions



BOYS ALWAYS SHAKE HANDS when introduced to other boys. They shake hands with a girl, too—if she offers her hand first. But if she makes no move to offer it, a boy should just bow slightly.

Social norm -
driven
decisions

Sanction and norm salience

- To be effective, social norms have to be **activated** and **made salient** [Bicchieri, 2006; Cialdini et al. 1990]
- Norm salience is a measure indicating how operative and relevant a norm is in a specific context [Andrighetto, Villatoro, Conte, 2010]
- Normative behaviors - compliance, violation, enforcement - are, intentionally or functionally, communicative acts on the basis of which individuals can infer how salient a norm is
 - “Look at me! Norms must be respected”;
 - “I do not care about the norm”
 - “I want you to comply with the norm”

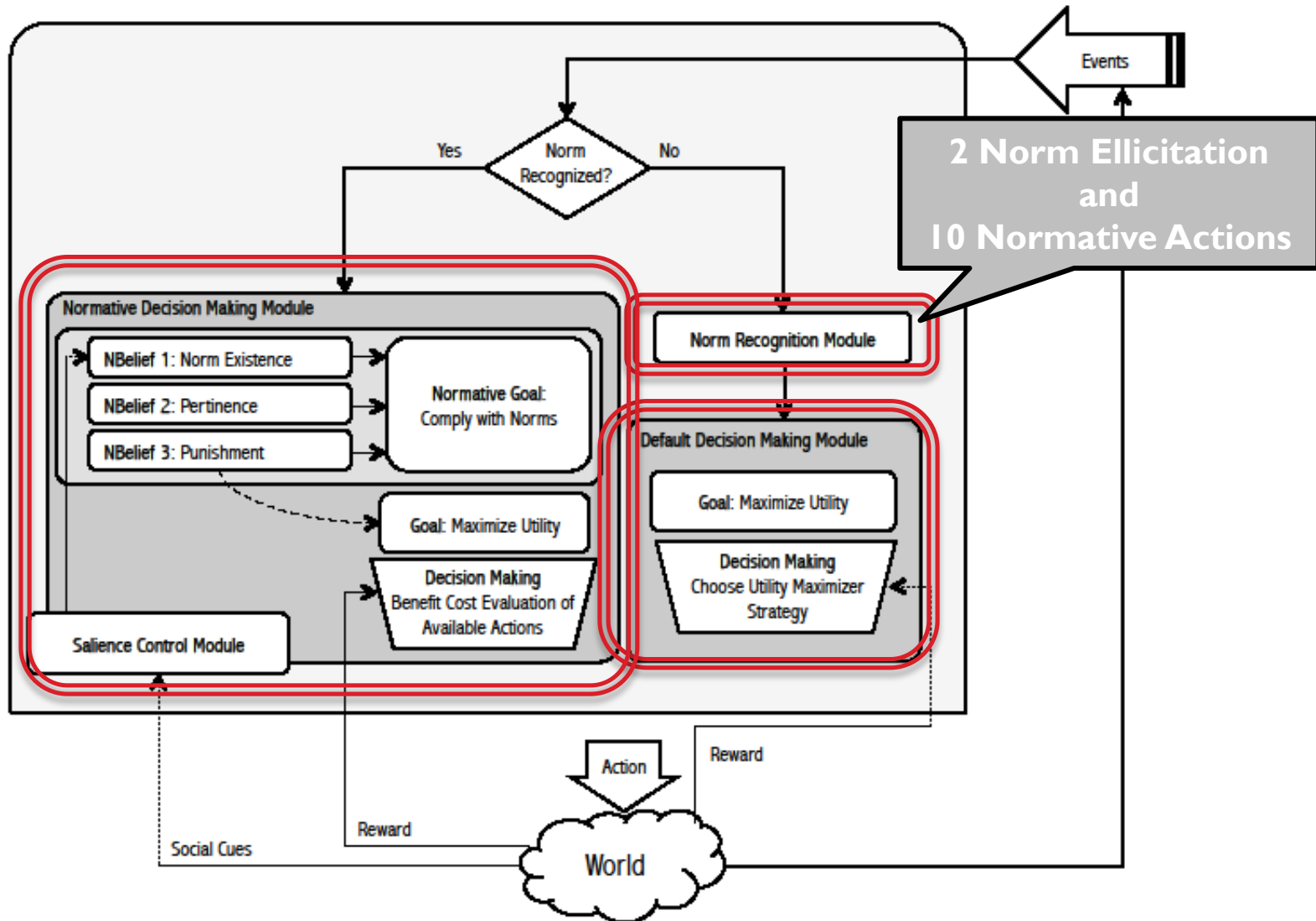
Updating Norm Salience

The actions of others provide information about how important a norm is within that social group:

- the amount of compliance and the cost people are willing to spend to obey;
- the surveillance rate, the frequency and intensity of punishment,
- the enforcement typology (private or public, 2nd and 3rd party, punishment or sanction, etc.)
- the efforts and costs expended to educate the population to a certain norm, e.g. publicity campaigns;
- the credibility and legitimacy of the normative source.

A specific social norm is more or less salient depending on how it fits with the social norms and values previously accepted by the individual.

EMIL-I-A



Salience Control Module

$$Sal_t^N = Sal_{t-1}^N + \frac{1}{\alpha \times \phi} (w_C + O \cdot w_O + NPD \cdot w_{NPD} + P \cdot w_P + S \cdot w_S + E \cdot w_E)$$

Social Cue	Weight
Norm Compliance/Violation (C)	$w_C = (+/-) 0.99$
Observed Norm Compliance (O)	$w_O = (+) 0.33$
Non Punished Violators (NPD)	$w_{NPD} = (-) 0.66$
Punishment Observed/Applied/Received (P)	$w_P = (+) 0.33$
Sanctioning Observed/Applied/Received (S)	$w_S = (+) 0.99$
Explicit Norm Invocation Observed/Received (E)	$w_E = (+) 0.99$

- ▶ Empirical values obtained from [Cialdini, 90] and experiments with human subjects.

Model

Agents **randomly** paired

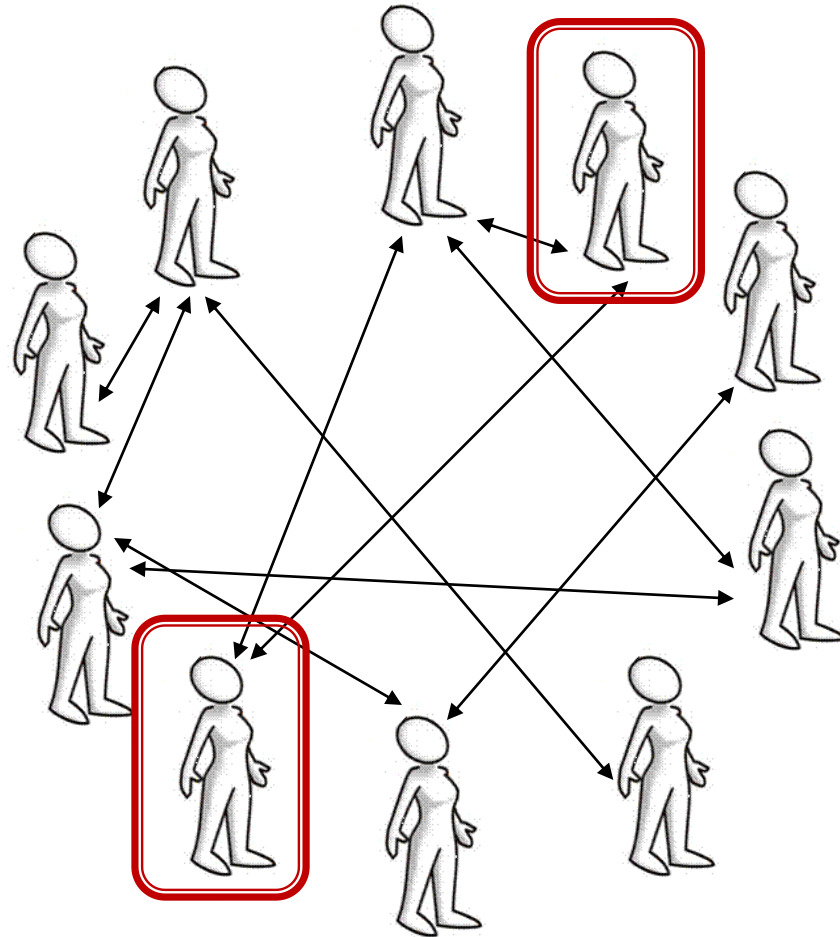
1st Stage: Each pair plays a PD game

2nd Stage: Each agent can punish his opponent

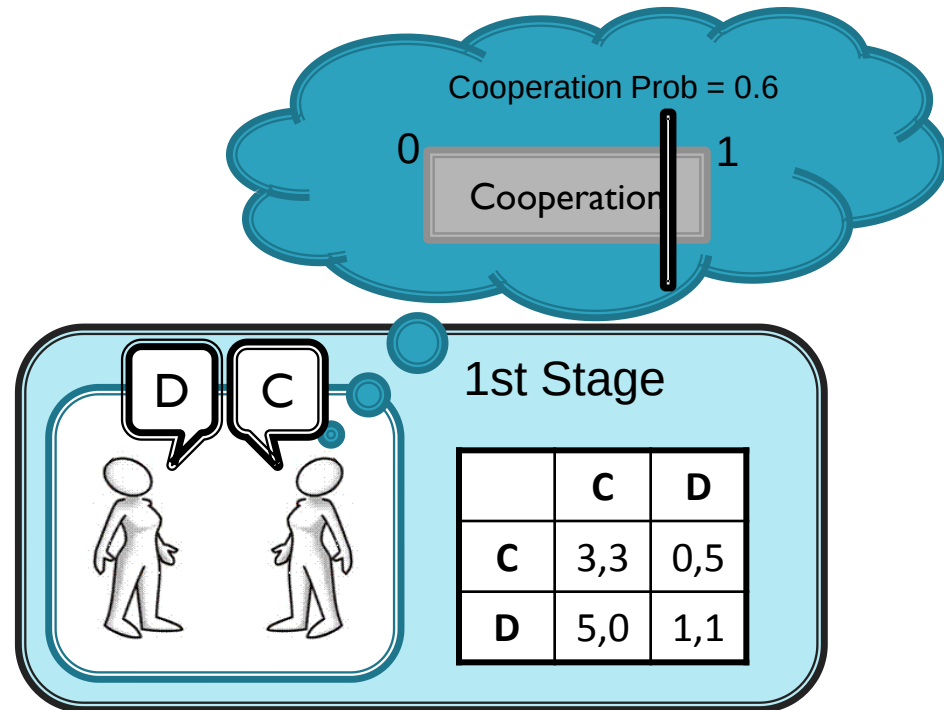
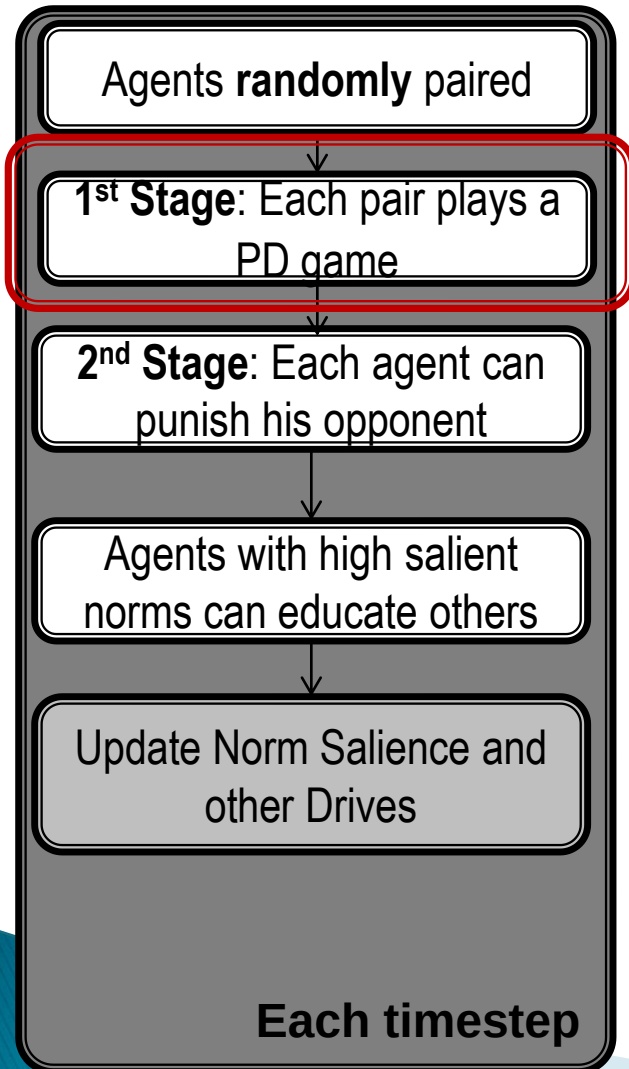
Agents with high salient norms can educate others

Update Norm Salience and other Drives

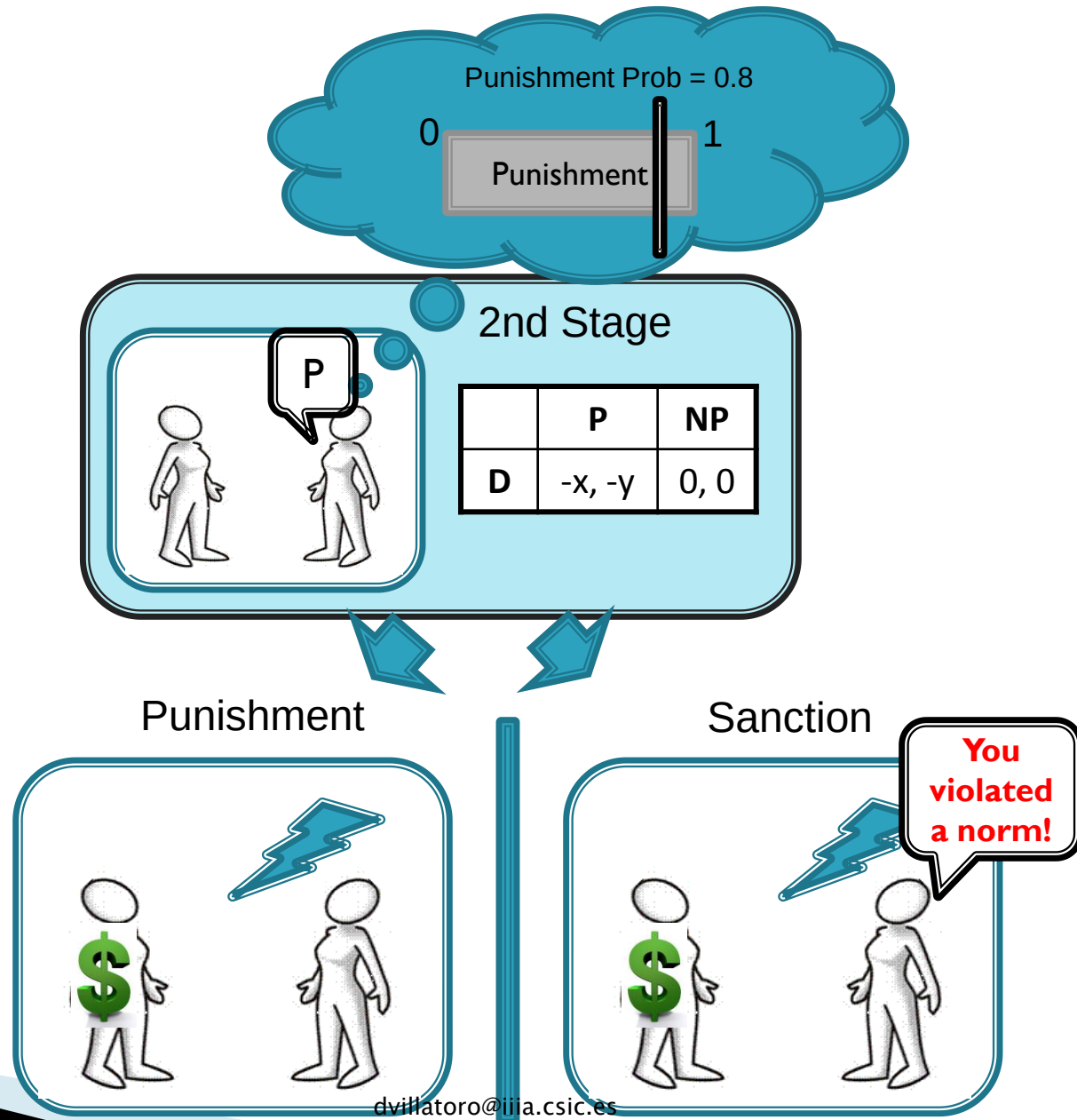
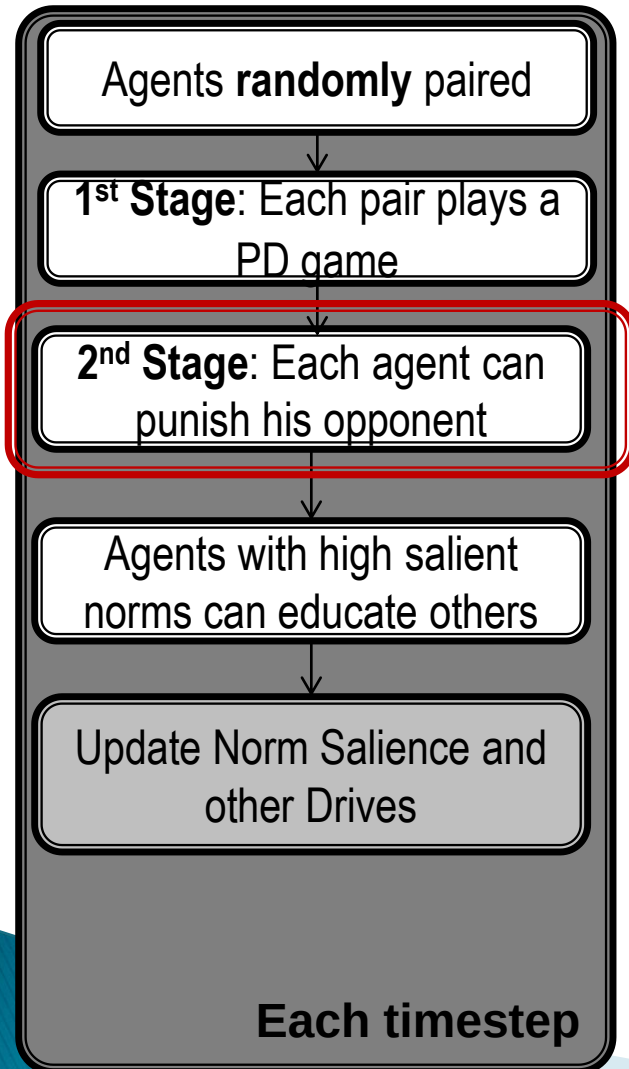
Each timestep



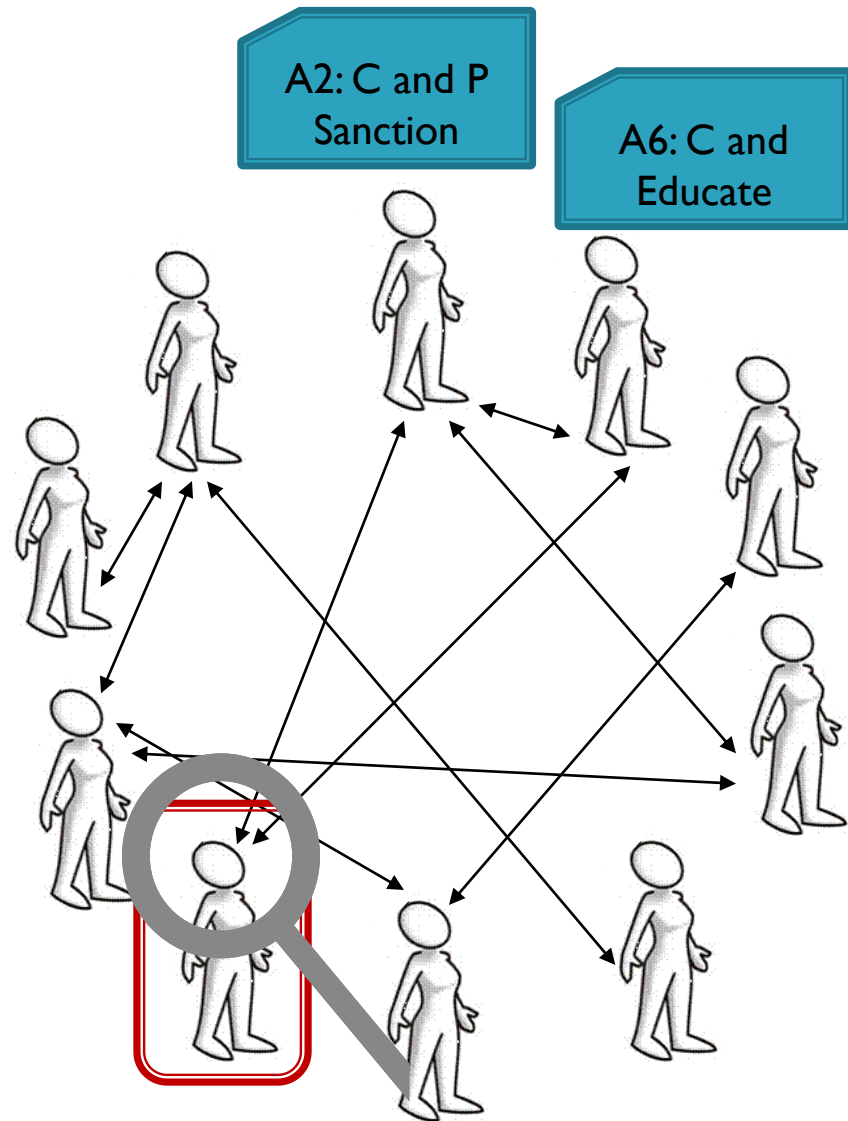
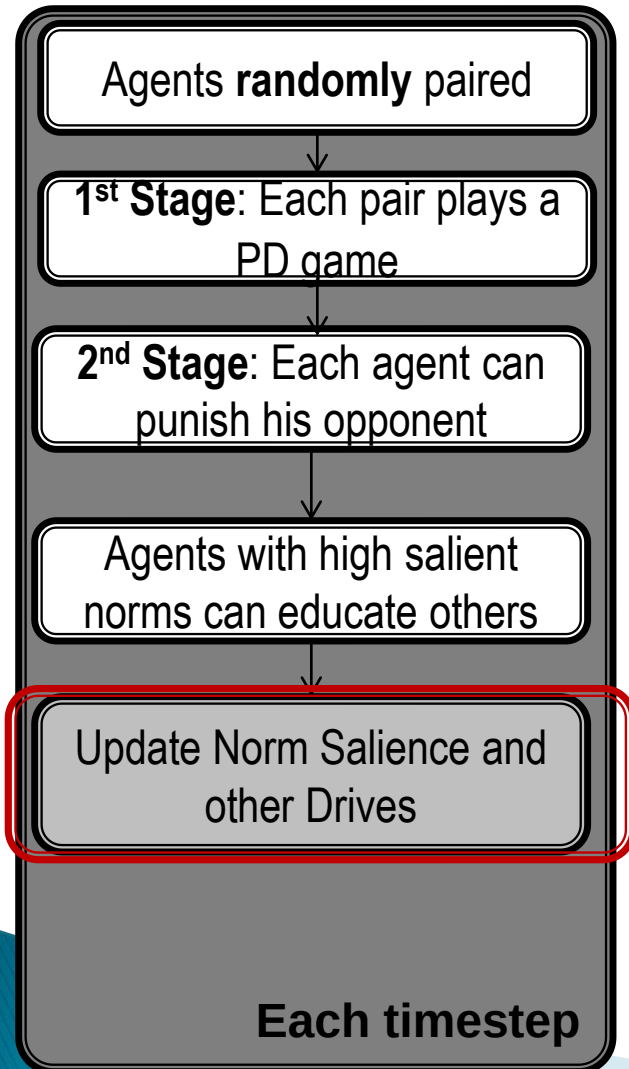
Model



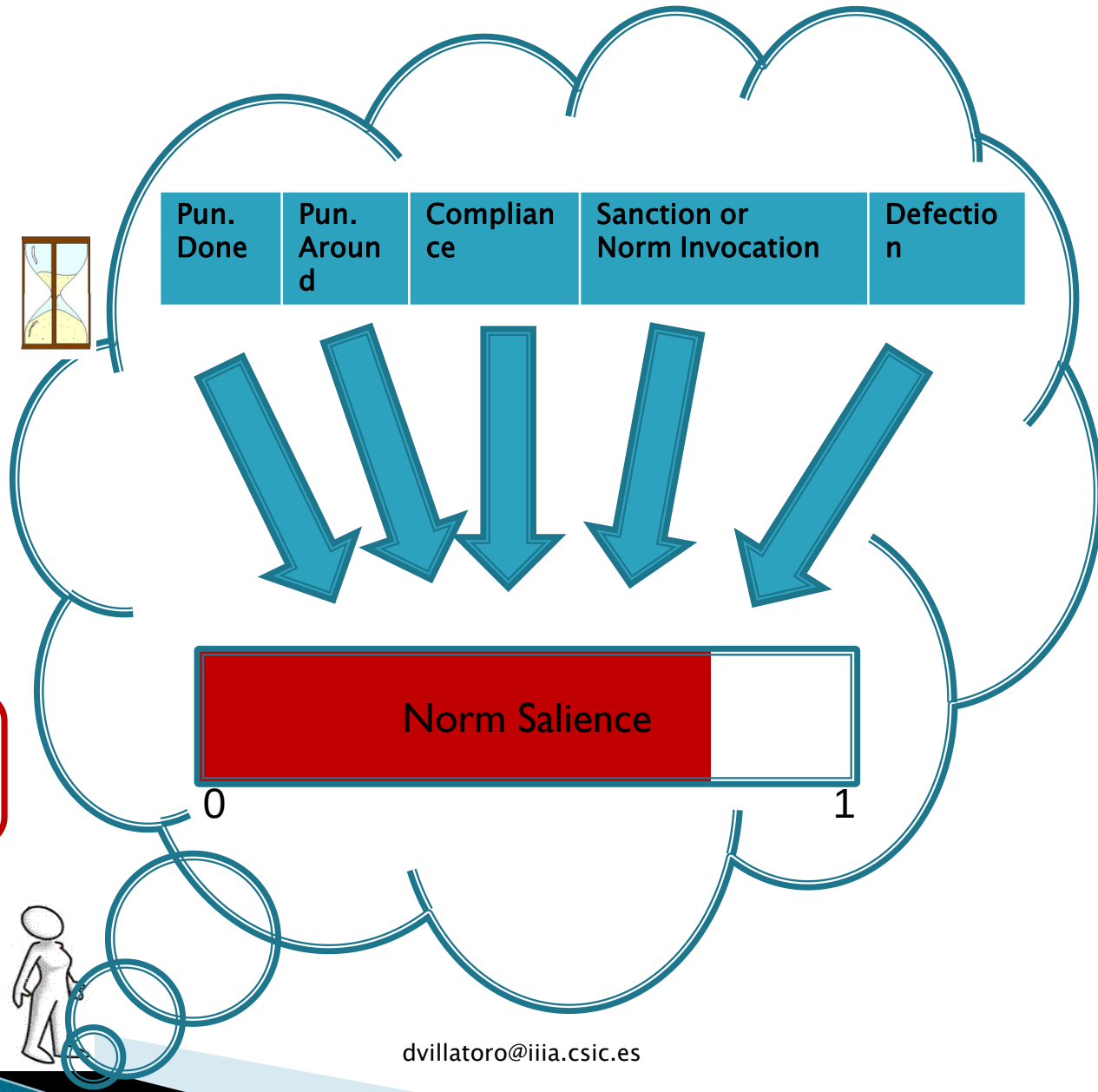
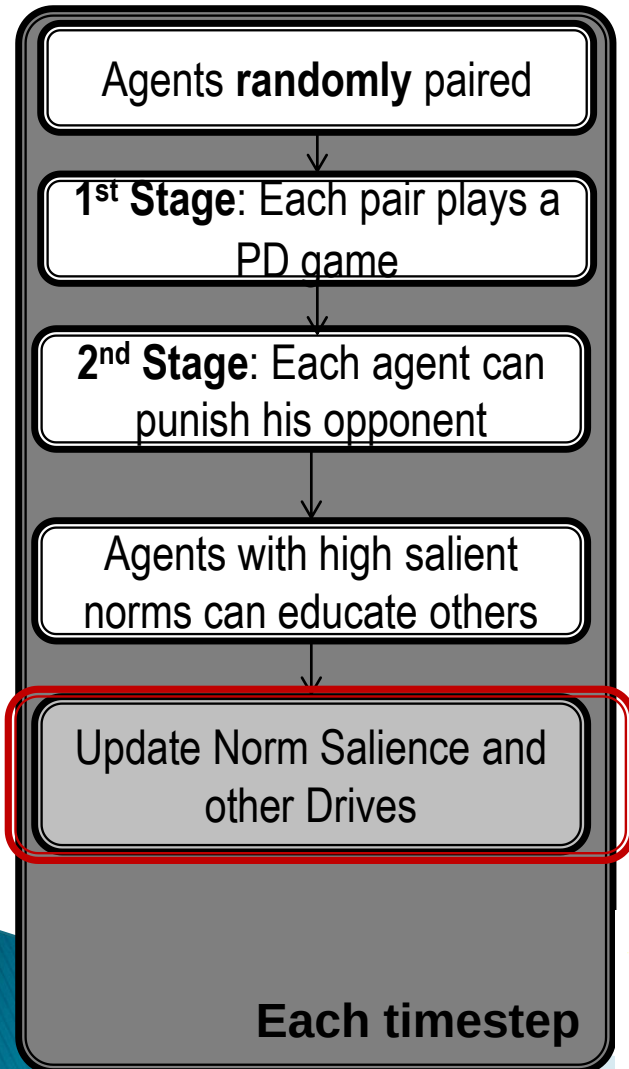
Model



Model

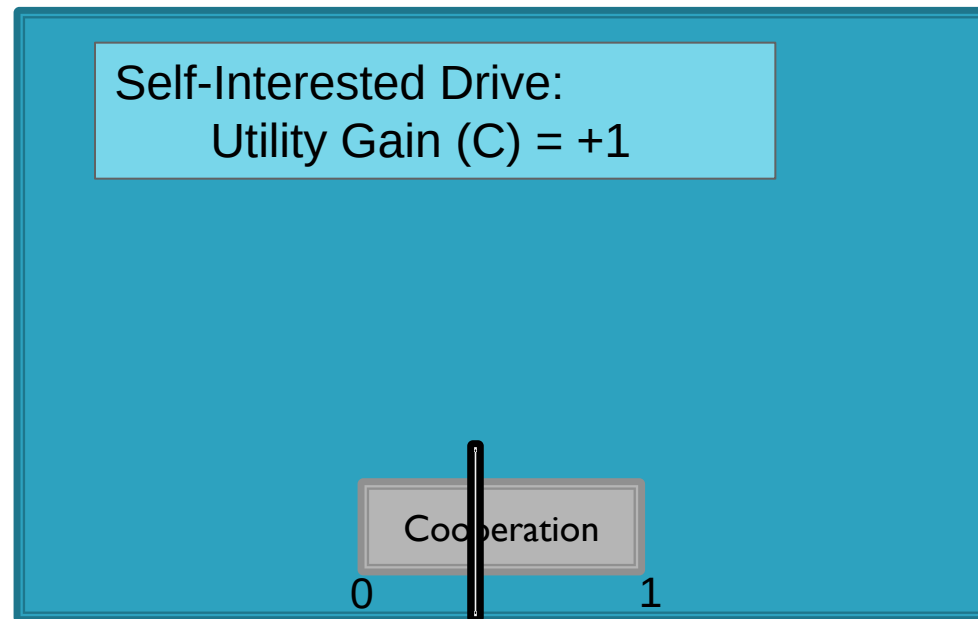


Model



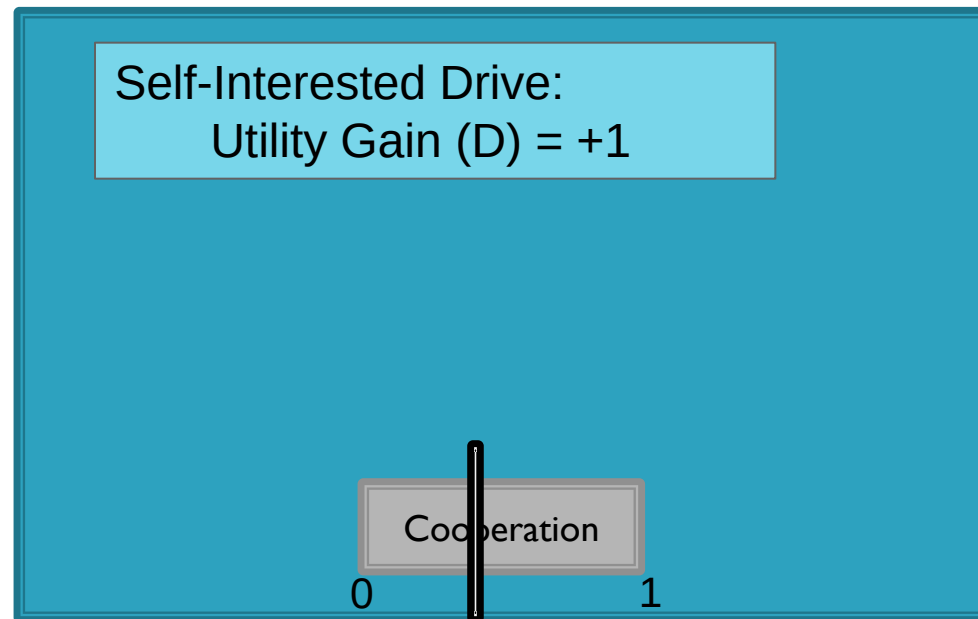
Agents' motivations

- **Cooperate or Defect?**



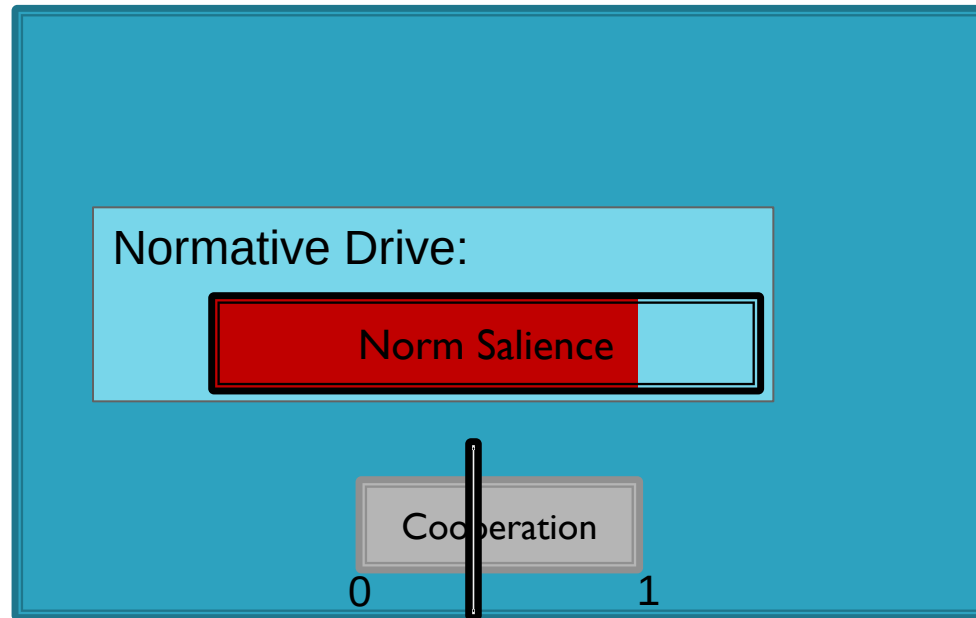
Agents' motivations

- **Cooperate or Defect?**



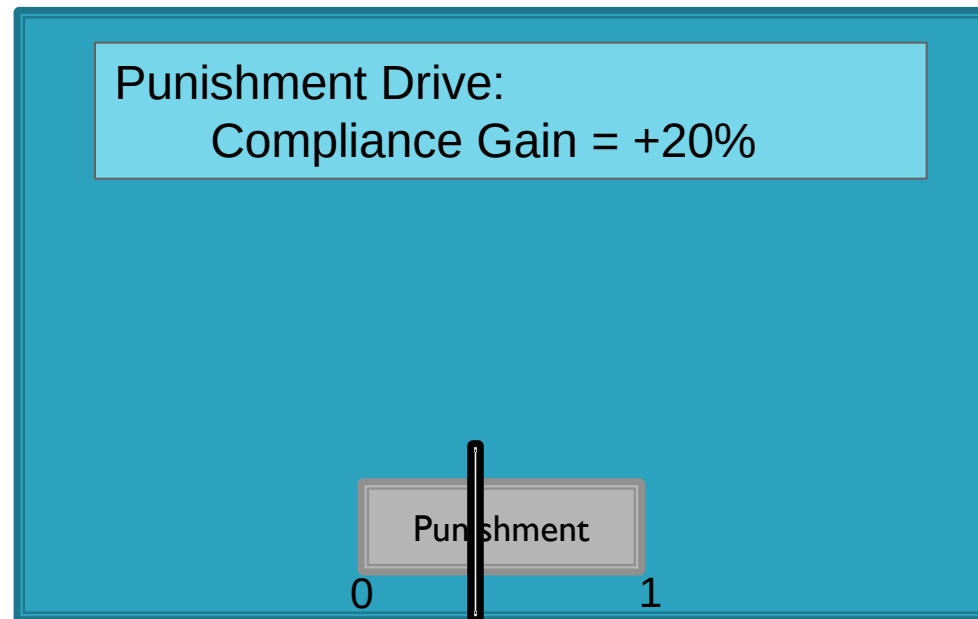
Agents' motivations

- **Cooperate or Defect?**



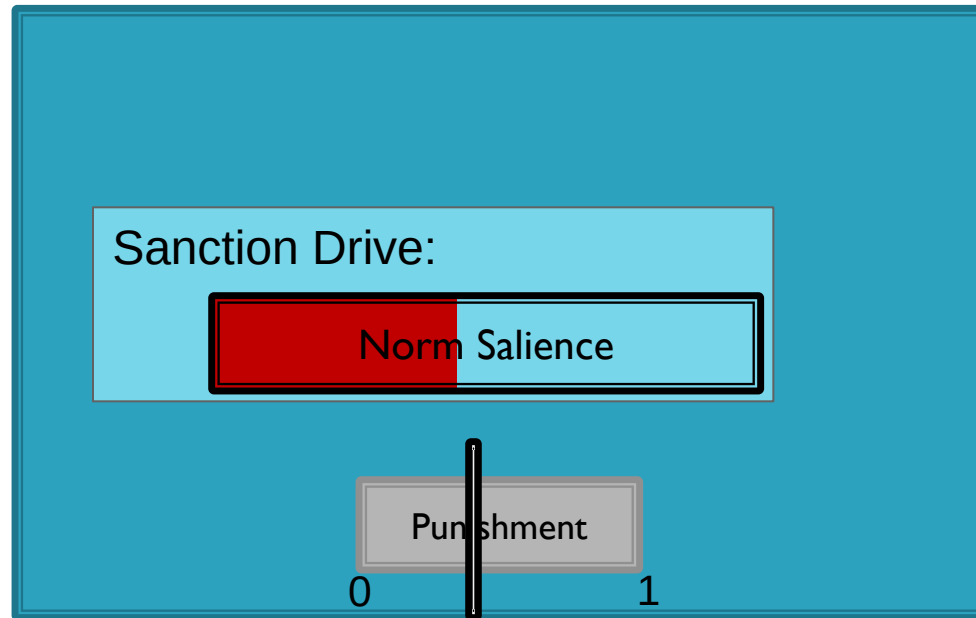
Agents' motivations

- **Punish or Not Punish?**



Agents' motivations

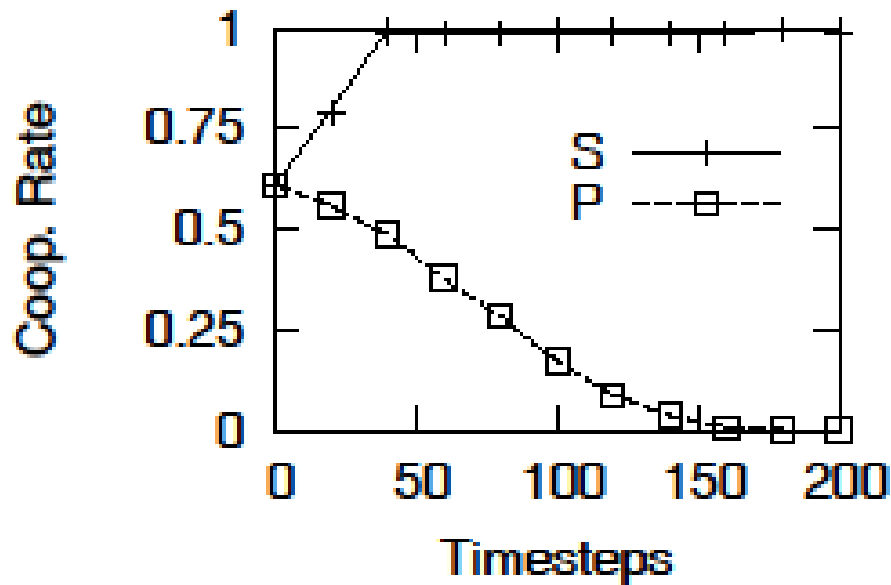
- **Punish or Not Punish?**



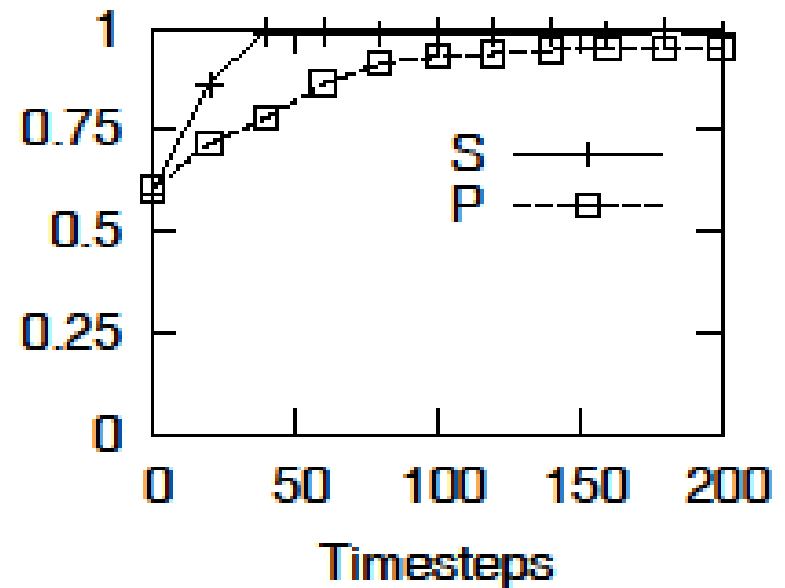
Scientific Questions

- ▶ Can sanction achieve higher levels of cooperation than punishment?
- ▶ Is sanction a cheaper method for society to self-regulate?
- ▶ What is the right punishment to be sent to norm violators?
- ▶ How stable is the society against environmental changes?

Effects of Punishment and Sanction



a) Damage 3



b) Damage 5

Relative Costs of Punishment and Sanction

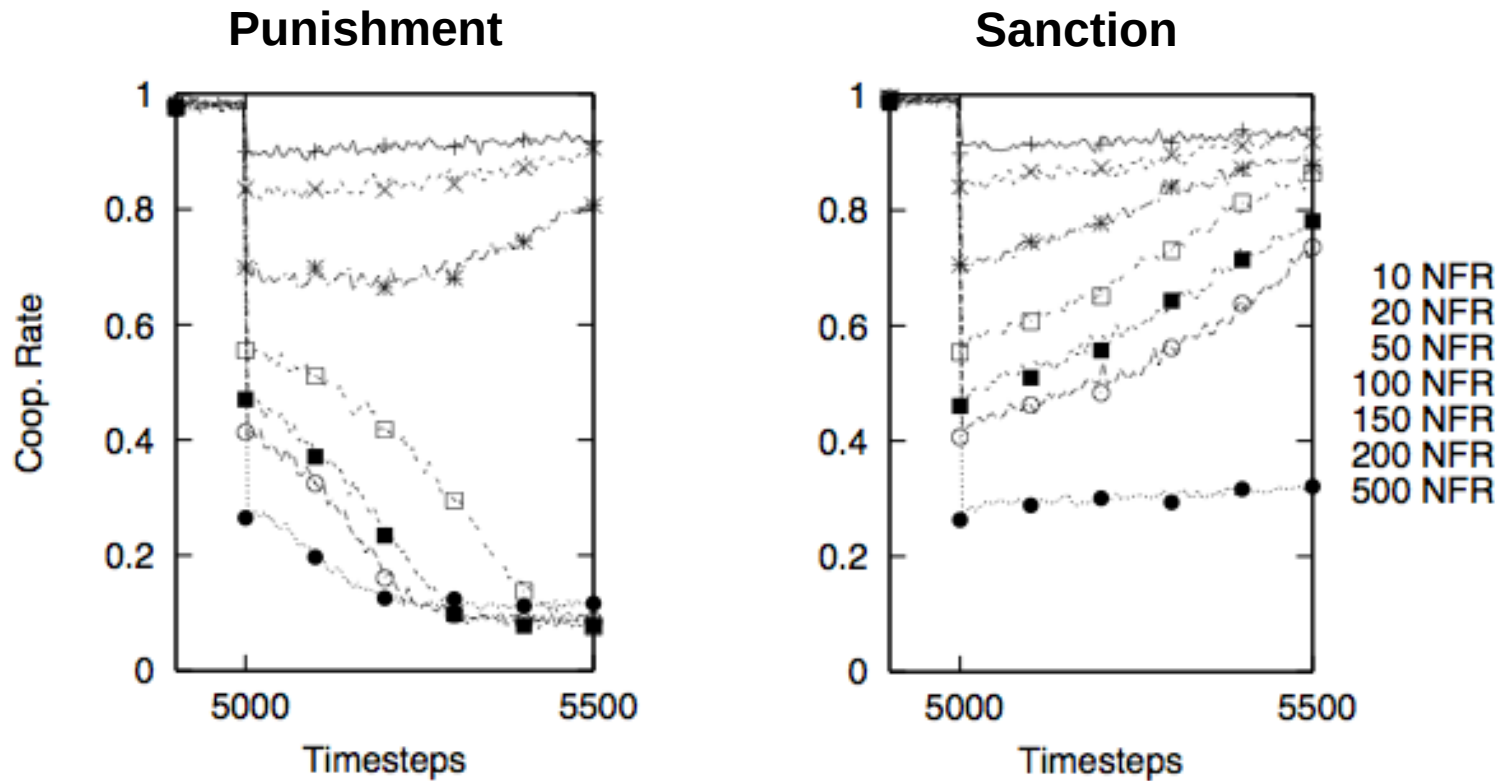
	Number of acts	Global Costs
Sanction	31.221	51.515
Punishment	37.757	62.300

Dynamic Adaptation Heuristic

If (Defectors_{t-1} ≤ Defectors_t)
then
 Increase PunCost Δ;
If (Defectors_{t-1} > Defectors_t)
then
 Decrease PunCost Δ;

	Punishments	Sanctions	Ind. Cost	Social Cost
Static Pun	4.8748	0	5	24.374
DAH Pun	2.1142	0	5.1640	10,9177
Static Sanc	0.3364	1.2491	5	7,9275
DAH Sanc	0.2932	0.8669	3.7031	4.2959

Free Riders Invasions



Conclusions

- ▶ **Sanction is Punishment + Norm Elicitation.**
- ▶ **Sanction is more effective** than punishment in:
 - ❑ achieving cooperation;
 - ❑ **reducing** the **costs** for cooperation to be achieved and maintained and
 - ❑ making the **population resilient** to environmental change - e.g. an abrupt interruption of the enforcement mechanism or the introduction on new free-riders.

Future Work

- ▶ Simulative different games:
 - Public Good Games, with Players > 2 .
- ▶ Study Topological Effects:
 - Where should a norm be loaded into the system to have the maximum virulence?
 - Where can free-riders be added to cause the maximum damage?
- ▶ Introduce a real world problem as an scenario:
 - Recycling.
 - Emergence of Cooperation on P2P networks.

Thanks!

Questions?

**Come and See the Poster tomorrow
11:30-12.10 @ Rooms 122-123**