Manipulations in Multi-Agent Systems

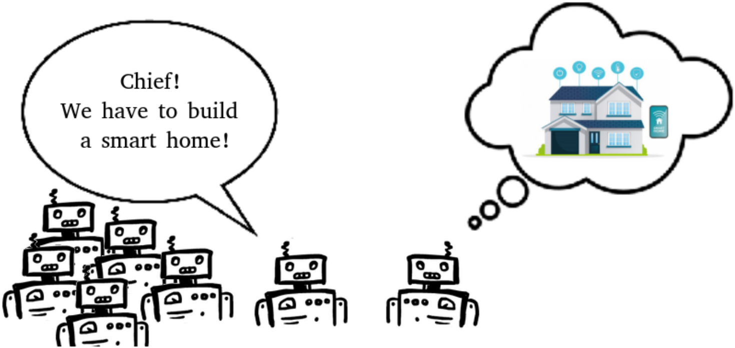
16 octobre 2020

Christopher Leturc

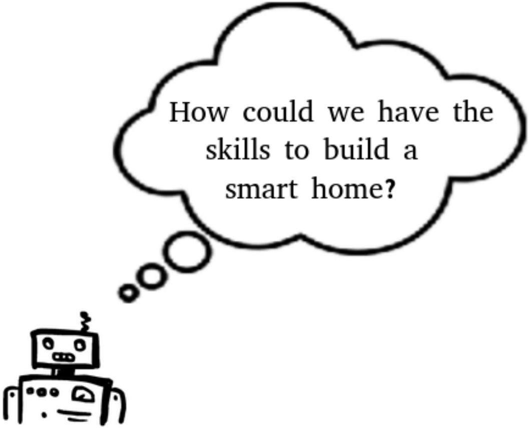
christopher.leturc@emse.fr

leturc.users.greyc.fr

The story of the robotic builders



The story of the robotic builders



The story of the robotic builders

The story of the robotic builders

The story of the robotic builders

What do robots need ?

Having **full trust** in another agent (here a company) means :

- the agent proposes a fair price (trust in its sincerity)

What do robots need ?

Having **full trust** in another agent (here a company) means :

- the agent proposes a fair price (trust in its sincerity)
- the agent makes the job (trust in its disposition)

What do robots need ?

Having **full trust** in another agent (here a company) means :

- the agent proposes a fair price (trust in its sincerity)
- the agent makes the job (trust in its disposition)
- the agent makes a good job (trust in its reliability)

What do robots need ?

Fortunately for the robots there are **reputation systems** !

A system that evaluates a collective trust based on testimonies of other robots.
The robots will therefore be able to make a first sorting.

A complex multi-agent framework

Hypothesis about agents

- **Multi-Agent Systems** (MAS) with **cognitive agents** i.e. capable of reasoning about mental states (beliefs, knowledge, intentions, . . .)
- Agents are assumed to be **rational** and **perfect reasoners**
- Some agents may be **malevolent** and **manipulative**

Objectives of this course

Understand the concept of manipulation through different MAS types :

1. Reputation systems : how to evalute trust ?
2. Normative systems : how to define laws for MAS ?
3. Voting systems : how to make a collective decision ?

Questions asked in this course

1. What is a manipulation in Multi-Agent Systems ?
2. How manipulation strategies can be constructed ?

Table of Contents

1 Manipulate agents with reputation

Presentation of reputation systems

Manipulation strategies and defense strategies

2 Study incentive mechanisms with GT

Normative systems (NS)

Introduction to Game Theory

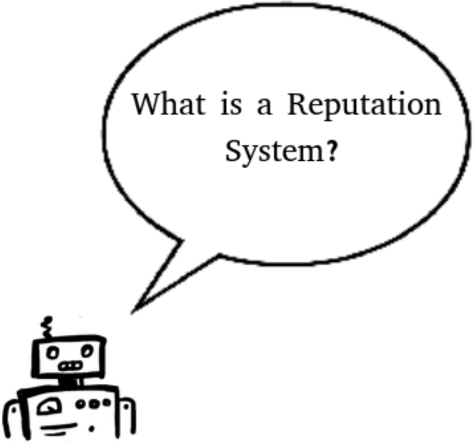
3 Influence a collective decision

Introduction to Social Choice Theory

Examples of manipulations in voting systems

4 What is a manipulation ?

Definition and examples of Reputation Systems



Definition and examples of Reputation Systems

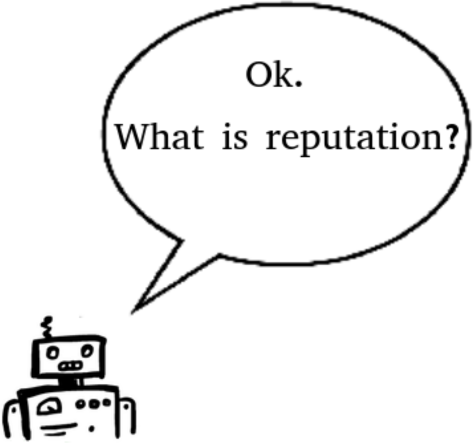
Definition [Sabater et Sierra, 2005]

Reputation Systems (RS) are systems where agents **interact**, **collect**, **share**, and **aggregate the results of their past interactions** to decide where they should be agents they can trust for future interactions.

Concrete systems

- The Amazon's product rating system
- The driver evaluation system of Blablacar
- The referencing of google pages

Definition and examples of Reputation Systems



Definition and examples of Reputation Systems

Reputation [Wang et Vassileva, 2003]

An agent's reputation is an agent's belief in the ability, honesty and reliability of another agent based on testimonials the agent has received.

Building trust based on testimonials needs two mechanisms :

- **Representation** is the method for representing testimonials
- **Dissemination** is the method used by agents to share testimonials

Dissemination and representation

- **Representation of testimonials** corresponds to the fact that the agents directly share the result of each interaction or an aggregation of trusts and testimonies already received [Schafer et al., 1999, Jøsang et Ismail, 2002]
- **Dissemination** is the method used by agents to share testimonials :
  - **Centralized** : a central authority is in charge of collecting agents' testimonials and then sharing them with the system agents [Schafer et al., 1999, Carbo et al., 2002, Srivatsa et al., 2005, Zhou et Hwang, 2007, Jøsang et Haller, 2007]
  - **Decentralized** : agents provide their testimonies directly to other agents in the system [Sabater et Sierra, 2001, Xiong et Liu, 2004, Sabater et al., 2006, Zhou et Hwang, 2007]

Notices

Agents can share their personal observations (direct), but also testimonials that other agents have provided themselves (indirect).

Reputation functions

Definition

The **reputation function** is the algorithm used by agents (or the central authority) to compute reputation values. It can be **global** or **individual** [Sabater et Sierra, 2005,Pinyol et Sabater-Mir, 2013].

Global : the reputation of an agent is defined independently of who is evaluating it [Schafer et al., 1999, Kamvar et al., 2003, Zhou et Hwang, 2007].
⇒ It is therefore the same for any agent.

Individual : the reputation is different from each agent's point of view [Jøsang et Ismail, 2002,Sabater et Sierra, 2001,Cheng et Friedman, 2005, Srivatsa et al., 2005]

Notices

Individual reputation functions model the fact that the evaluation of an interaction is subjective and that what may seem like a good interaction to one agent is not necessarily a good interaction for another agent.

Example of a ReputationSystem : BetaReputation

Representation of Trust

Trust is modeled by a pair $\langle r_{ij}, s_{ij} \rangle \in \mathbb{R}^+ \times \mathbb{R}^+$ corresponding respectively to the **positive** and **negative part** of the evaluation of an agent *i* of the interactions he had with an agent *j*.

Dissemination mechanism

When the agent *i* receives a testimony $\langle r_{jk}, s_{jk} \rangle$ from the agent *j*, about a third agent *k*, he aggregates it with his own observations as follows :

$$r_k^{i:j} = \frac{2r_{ij}r_{jk}}{(s_{ij}+2)(r_{jk}+s_{jk}+2)+2r_{ij}}$$

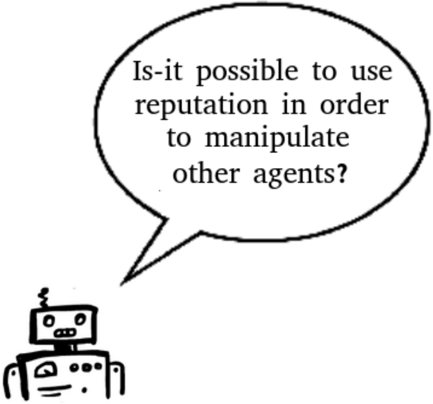
$$s_k^{i:j} = \frac{2r_{ij}s_{jk}}{(s_{ij}+2)(r_{jk}+s_{jk}+2)+2r_{ij}}$$

Reputation function

BetaRep is an **individual reputation function** where the reputation of *k* (noted $Rep(r_k, s_k)$) is computed from the aggregation of all the testimonials received :

$$Rep(r_k, s_k) = \frac{r_k - s_k}{r_k + s_k + 2}$$

Weaknesses of reputation systems


Is-it possible to use reputation in order to manipulate other agents?

Weaknesses of reputation systems

If there are many reputation systems, one of the main issues beyond formalization of multi-agent systems is that some agent may be **malevolent**, **dishonest** or **manipulative**.

Indeed, an agent can use a certain strategy to take advantage of the system he belongs to. We will talk about **manipulation strategy**. Another part of literature is devoted to the **study of the robustness** of these systems against manipulations.

Thus, in the sequel we will study :

- Some manipulation strategies in the case of Reputation Systems
- Present some defense mechanisms

List of manipulation strategies

C1 : Agents must be persistent over time

Sybil attack creates false identities in the system to affect the reputation of other agents [Douceur, 2002]

Whitewashing involves leaving the system when a reputation value is too low in order to re-enter the system under a new identity with the same reputation as a new entrant agent [Feldman et al., 2006]

Discrimination consists in targeting interactions : always interacting well with agents that are persistent over time (such as trusted agents) and always interacting poorly with agents identified as not very persistent.

List of manipulation strategies

C2 : The results of interactions must be communicated to other agents and are accessible in the future

Three manipulation strategies attack this axiom :

- **promotion**
- **self-promotion**
- **defamation**

They consist in sharing false testimony with other agents in the system so that the calculation of reputation values is to the advantage of the dishonest agent [Jin et al., 2007].

We distinguish here between :

- promotion carried out by agents in collusion
- and self-promotion that consists in using Sybil agents.

List of manipulation strategies

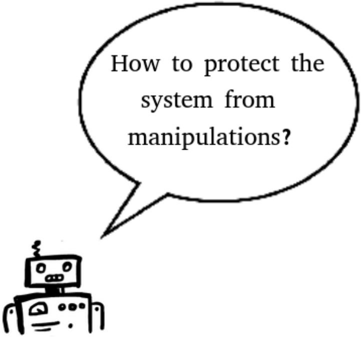
C3 : The decision-making process must be guided by these interaction results [Jøsang and Golbeck, 2009]

We can give three kind of possible manipulation strategies :

- Betrayal
- Planned attack

These manipulation strategies consist in increasing the reputation value of the malicious agent through a succession of good interactions before deliberately carrying out a bad interaction at a given moment.

Defense against manipulation strategies



Defense against manipulation strategies

Three possible approaches

1. Design robust system by defining a robustness axiomatics
2. Detect manipulation by reasoning or by statistical measure when a manipulation occurred
3. Play on the fact that designing a manipulation strategy is too complex
4. Design incentive mechanisms (punish bad behaviour and reward good behaviour)

1. Defining a robustness axiomatics

The idea of this defensive approach

To defend against those manipulation strategies, wouldn't it be best to build systems that make it impossible to make such strategies operable ? This approach consists in giving certain properties to our system and **showing that with these properties, manipulations are impossible to perform**.

Examples of defense

- **Forgetfulness** consists in avoiding information persistence and quickly detect treachery (although this makes the system more sensitive to oscillating attacks, since only a small number of good interactions are needed to regain a good reputation value) [Jøsang et Ismail, 2002]
- **Certificate of origin** allows to certify the origin and content of messages (although this provides no protection against promotions where multiple malicious agents agree to fake interactions)
- **No identify persists over time** consists by forcing the use of short-term disposable identifiers to avoid discrimination [Dellarocas, 2000, Singh and Liu, 2003]

2. Detect manipulation strategies

The idea of this defensive approach

If we are able to detect manipulation strategies, then we can remove agents from the system or prevent the effects of such manipulation strategies.

Examples of defense

- **Using statistical measures** consists in detecting suspicious behaviors in the system by learning or statistical inference [Santos and Johnson, 2004]
- **Detect by automatic reasoning** consists in detecting deception by logical deductions [Muller and Vercouter, 2004]

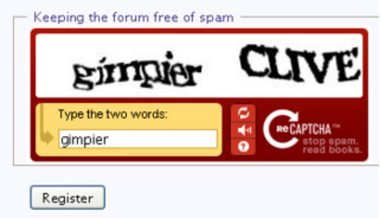
3. Build complex systems

The idea of this defensive approach

Build complex systems that make the manipulations hard to perform

Example of defense

A **CAPTCHA mechanism** consists in defining a **complex authentication mechanism** so as to defend the system against sybil attacks [Douceur, 2002, Borisov, 2006]



4. Design incentive mechanisms

The idea of this defensive approach

Some techniques **make a rationality hypothesis** about the behavior of malicious agents to force them to change strategy.

Example of defense

- Propose **payment mechanisms** that induce agents to provide true testimonials [Miller et al., 2005, Friedman et al., 2007]
- Use a testimonial aggregation function that **forces malicious agents to adopt a stochastic strategy to maximize their gain** [Bonnet, 2012]
- **Punish bad behaviors and reward good behaviors** with a Normative System [Boella *et al.*, 2008]

Normative systems

Normative systems

The concept of Normative System

Normative Systems define laws (or norms) in the system that agent must respect.

Example of application

In those systems a malevolent or dishonest agent that does not respect rules will be punished and a good agent will be rewarded.

The Game Theory (GT) allows us to study this kind of incentive mechanism.

Table of Contents

1 Manipulate agents with reputation
Presentation of reputation systems
Manipulation strategies and defense strategies

2 Study incentive mechanisms with GT
Normative systems (NS)
Introduction to Game Theory

3 Influence a collective decision

4 What is a manipulation ?

What is a Normative Multi-Agent System ?

Definition [Boella *et al.*, 2008]

A **Normative Multi-Agent System** (NMAS) is a Multi-Agent System (MAS) organized by means of mechanisms to **represent**, **communicate**, **distribute**, **detect**, **create**, **modify**, and **enforce norms**, and **mechanisms to deliberate about norms** and **detect norm violation and fulfilment**.

Examples of Normative Systems (NS)

- National legal systems (e.g. enforce laws in society)
- A group of people can be considered as a NS. NS can represent the emergence of collective behaviors (e.g. conformism, reciprocity, . . .)

⇒ Normative Systems make it possible to define incentive mechanisms so as to influence the agents' behaviors.

What is an incentive mechanism ?

Technical issues with Normative Systems

Detecting violations of norms can be very hard and costly in ressources (time,energy,etc.).

The problematic of NS

How can we ensure that agents will have interests to adhere to these norms ?

An answer with Game Theory

This can work only if **agents are assumed to be rational**. By rational we mean that an **agent will always try to maximize its interests (utility)**.

Game theory is a mathematical theory which is interested in modeling the interactions of agents when they are rational.

What is the use of Game Theory ?

Modeling agents’ interactions in

- Board games
- Card games
- Nature
- Resource Sharing
- Contract negotiations
- Military and economic strategies
- **Norm systems**
- . . .

Taxonomy of games

Main characteristics of games : players = agents

n-player games : how many players ?

Cooperative / **non-cooperative** : do the agents can form collations ?

Simultaneous / sequential : do the agents play simultaneously ?

Perfect information / imperfect information : do the agents know about :

- its possibilities of action ?
- the possibilities of action of the other players ?
- the gains resulting from these actions ?
- the motivations of other players ?

Zero-sum / **non-zero-sum** : does the agents' interest is strictly opposed to the interest of other ? (yes = zero-sum)

...

Here we will only study the **prisoner's dilemma**.

Definition of a Game

A Game G is N-uplet $G = (\mathcal{N}, \{S_i\}_{i\in\mathcal{N}}, \{\mu_i\}_{i\in\mathcal{N}})$ where :

- $\mathcal{N} = \{1, \ldots, n\}$ is a no-empty set of agents
- $\{S_i\}_{i\in\mathcal{N}}$ is a set of strategies for each agents (here we consider a strategy as an action that one agent can play)
- $\forall i \in \mathcal{N}, \mu_i : S_1 \times \ldots \times S_n \rightarrow \mathbb{R}$ is a function called *utility function* that associates for each agent i and each strategy applied by all agents (joint action) an utility (a real number)

Notices about utility

- A **joint action** is just the combination of all played actions (e.g. $(a_1, \ldots, a_n) \in S_1 \times \ldots \times S_n$ means that agent 1 chooses to play the action a_1, agent 2 chooses a_2, . . ., agent n chooses a_n)
- Utility μ_i is a **subjective measure** which represents the gains (if positive) or losses (if negative) associated with the joint action for agent i.

Prisoner's dilemma

Two prisoners are being interrogated. They can betray or remain silent.

The rules of the game

1. if agent 1 **betrays** while 2 **stays silent**, then agent 1 stays **one year** in prison the while 2 takes for **10 years**
2. if agent 2 **betrays** while 1 **stays silent**, then agent 2 stays **one year** in prison the while 1 takes for **10 years**
3. if each agent **stays silent**, then each agent takes for **2 years**
4. if each agent **betrays**, then each agent takes for **5 years**

Prisoner’s dilemma

A bit of formalization

The formal game

A set of agents : $\mathcal{N} = \{1, 2\}$

A set of strategies : $\forall i \in \mathcal{N}, \mathcal{S}_i = \{Betrays(B), \text{Stays silent}(C)\}$

Possible joint actions : $\mathcal{J} = \{(B, B); (B, C); (C, B); (C, C)\}$

For instance : (B, C) means that 1 betrays while 2 stays silent.

Utility functions :

- $\mu_1((B, B)) = -5$; $\mu_2((B, B)) = -5$
- $\mu_1((B, C)) = -1$; $\mu_2((B, C)) = -10$
- $\mu_1((C, B)) = -10$; $\mu_2((C, B)) = -1$
- $\mu_1((C, C)) = -2$; $\mu_2((C, C)) = -2$

Classical example of the Prisoner's Dilemma

		Prisoner 2	
		Stays silent	Betrays
Prisoner 1	Stays silent	-2 -2	-1 -10
	Betrays	-10 -1	-5 -5

TABLE – A payoff matrix for the prisoner's dilemma

Classical example of the Prisoner's Dilemma

		Prisoner 2	
		Stays silent	Betrays
Prisoner 1	Stays silent	-2 -2	-1 -10
	Betrays	-10 -1	-5 -5

TABLE – A payoff matrix for the prisoner's dilemma

Question

If you are agent 1, then what would you do ?

Classical resolution by considering Nash equilibrium

		Prisoner 2	
		Stays silent	Betrays
Prisoner 1	Stays silent	-2 -2	-1 -10
	Betrays	-10 -1	-5 -5

TABLE – A payoff matrix for the prisoner's dilemma

Classical resolution by considering Nash equilibrium

		Prisoner 2	
		Stays silent	Betrays
Prisoner 1	Stays silent	-2 -2	-1 -10
	Betrays	-10 -1	-5 -5

TABLE – A payoff matrix for the prisoner's dilemma

Classical resolution by considering Nash equilibrium

		Prisoner 2	
		Stays silent	Betrays
Prisoner 1	Stays silent	-2 -2	-1 -10
	Betrays	-10 -1	-5 -5

TABLE – A payoff matrix for the prisoner's dilemma

Classical resolution by considering Nash equilibrium

		Prisoner 2	
		Stays silent	Betrays
Prisoner 1	Stays silent	-2 -2	-1 -10
	Betrays	-10 -1	-5 -5

TABLE – A payoff matrix for the prisoner's dilemma

Classical resolution by considering Nash equilibrium

		Prisoner 2	
		Stays silent	Betrays
Prisoner 1	Stays silent	-2 -2	-1 -10
	Betrays	-10 -1	-5 -5

TABLE – A payoff matrix for the prisoner's dilemma

Classical resolution by considering Nash equilibrium

Answer

If all agent are assumed to be **rational** (i.e. they maximize their utility), then the best rational choice for me is to choose to **betray**.
$\Rightarrow$ It is what we call a **Nash Equilibrium** (NE).

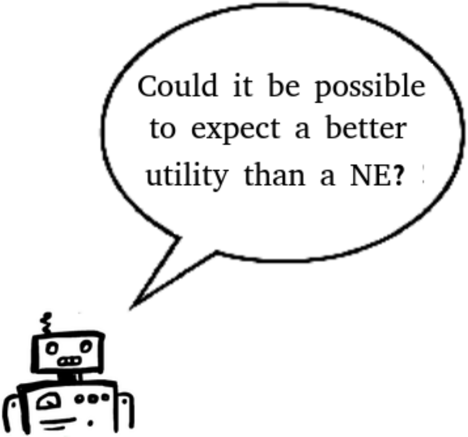
Examples where NE is used

- **Laws are made in such a way as to encourage you to side with the law !**
- Economical systems follows NE (e.g. supply and demand)

Interesting remark for NE

If you try to deviate from a NE, you will be a loser.

Classical resolution by considering Nash equilibrium



Classical resolution by considering Nash equilibrium

For sure !

You will need two things :

- you will need to **coordinate with the other agent**
- you will need to **trust your partner that he will keep his strategy**

Is it really worth it ?

Because if he betrays you, then you will lose everything...

Let’s go back to our Normative System (NS)

How could we model our incentive mechanism to push agent to be honest ?

Let's go back to our Normative System (NS)

How could we model our incentive mechanism to push agent to be honest ?

Exercice on modeling an incentive mechanism with a NS

- Let a_c be an agent (**company**) of the MAS and a_s be a **supervisory agent** whose role is to monitor other agents
- The agent a_s may or may not choose to put one agent on surveillance to check its behaviour. But putting another agent on surveillance is going to **cost him resources** c. But if it detects a malevolent agent, then this agent will get a **bonus** b and impose a **penalty** p_c on the other agent. If agent a_s misses a malevolent agent, then he will have also **a penalty** p_s.
- We assume that a **company** a_c has only two strategies : be honest or not (be malevolent). Being malevolent can bring to the company lot of **money** m if it is not detected. Being honest can give her a **small reward** r if the agent a_c is detected and nothing otherwise.

Questions

Model with a prisoner's dilemma this situation. What can you deduce from this ? Is it possible to have a NE on the joint action $(\overline{S}, H)$?

Let’s go back to our Normative System (NS)

How could we model our incentive mechanism to push agent to be honest ?

A model of this situation

A set of agents : $\mathcal{N} = \{a_s, a_c\}$

A set of strategies : $\mathcal{S}_{a_s} = \{\text{Supervise}(S), \text{Do not supervise}(\overline{S})\}$
$\mathcal{S}_{a_c} = \{\text{Be honest}(H), \text{Be malevolent}(M)\}$

Possible joint actions : $\mathcal{J} = \{(\overline{S}, M); (\overline{S}, H); (S, M); (S, H)\}$

Let’s go back to our Normative System (NS)

How could we model our incentive mechanism to push agent to be honest ?

A model of this situation

A set of agents : $\mathcal{N} = \{a_s, a_c\}$

A set of strategies : $\mathcal{S}_{a_s} = \{\text{Supervise}(S), \text{Do not supervise}(\overline{S})\}$
$\mathcal{S}_{a_c} = \{\text{Be honest}(H), \text{Be malevolent}(M)\}$

Possible joint actions : $\mathcal{J} = \{(\overline{S}, M); (\overline{S}, H); (S, M); (S, H)\}$

Utility functions :

- $\mu_{a_s}((\overline{S}, M)) = p_s$; $\mu_{a_c}((\overline{S}, M)) = m$

Let’s go back to our Normative System (NS)

How could we model our incentive mechanism to push agent to be honest ?

A model of this situation

A set of agents : $\mathcal{N} = \{a_s, a_c\}$

A set of strategies : $\mathcal{S}_{a_s} = \{\text{Supervise}(S), \text{Do not supervise}(\overline{S})\}$
$\mathcal{S}_{a_c} = \{\text{Be honest}(H), \text{Be malevolent}(M)\}$

Possible joint actions : $\mathcal{J} = \{(\overline{S}, M); (\overline{S}, H); (S, M); (S, H)\}$

Utility functions :

- $\mu_{a_s}((\overline{S}, M)) = p_s$; $\mu_{a_c}((\overline{S}, M)) = m$
- $\mu_{a_s}((\overline{S}, H)) = 0$; $\mu_{a_c}((\overline{S}, H)) = 0$

Let’s go back to our Normative System (NS)

How could we model our incentive mechanism to push agent to be honest ?

A model of this situation

A set of agents : $\mathcal{N} = \{a_s, a_c\}$

A set of strategies : $\mathcal{S}_{a_s} = \{\text{Supervise}(S), \text{Do not supervise}(\overline{S})\}$
$\mathcal{S}_{a_c} = \{\text{Be honest}(H), \text{Be malevolent}(M)\}$

Possible joint actions : $\mathcal{J} = \{(\overline{S}, M); (\overline{S}, H); (S, M); (S, H)\}$

Utility functions :

- $\mu_{a_s}((\overline{S}, M)) = p_s$; $\mu_{a_c}((\overline{S}, M)) = m$
- $\mu_{a_s}((\overline{S}, H)) = 0$; $\mu_{a_c}((\overline{S}, H)) = 0$
- $\mu_{a_s}((S, M)) = c + b$; $\mu_{a_c}((S, M)) = p_c$

Let's go back to our Normative System (NS)

How could we model our incentive mechanism to push agent to be honest ?

A model of this situation

A set of agents : $\mathcal{N} = \{a_s, a_c\}$

A set of strategies : $\mathcal{S}_{a_s} = \{\text{Supervise}(S), \text{Do not supervise}(\overline{S})\}$
$\mathcal{S}_{a_c} = \{\text{Be honest}(H), \text{Be malevolent}(M)\}$

Possible joint actions : $\mathcal{J} = \{(\overline{S}, M); (\overline{S}, H); (S, M); (S, H)\}$

Utility functions :

- $\mu_{a_s}((\overline{S}, M)) = p_s \,;\, \mu_{a_c}((\overline{S}, M)) = m$
- $\mu_{a_s}((\overline{S}, H)) = 0 \,;\, \mu_{a_c}((\overline{S}, H)) = 0$
- $\mu_{a_s}((S, M)) = c + b \,;\, \mu_{a_c}((S, M)) = p_c$
- $\mu_{a_s}((S, H)) = c \,;\, \mu_{a_c}((S, H)) = r$

Let’s go back to our Normative System (NS)

How could we model our incentive mechanism to push agent to be honest ?

A model of this situation

A set of agents : $\mathcal{N} = \{a_s, a_c\}$

A set of strategies : $\mathcal{S}_{a_s} = \{\text{Supervise}(S), \text{Do not supervise}(\overline{S})\}$
$\mathcal{S}_{a_c} = \{\text{Be honest}(H), \text{Be malevolent}(M)\}$

Possible joint actions : $\mathcal{J} = \{(\overline{S}, M); (\overline{S}, H); (S, M); (S, H)\}$

Utility functions :

- $\mu_{a_s}((\overline{S}, M)) = p_s$; $\mu_{a_c}((\overline{S}, M)) = m$
- $\mu_{a_s}((\overline{S}, H)) = 0$; $\mu_{a_c}((\overline{S}, H)) = 0$
- $\mu_{a_s}((S, M)) = c + b$; $\mu_{a_c}((S, M)) = p_c$
- $\mu_{a_s}((S, H)) = c$; $\mu_{a_c}((S, H)) = r$

Let’s go back to our Normative System (NS)

How could we model our incentive mechanism to push agent to be honest ?

		Company	
		M	H
Supervisor	$\overline{S}$	m p_s	0 0
	S	p_c $c + b$	r c

TABLE – Prisoner’s dilemma

What to say about variables ?

Let’s go back to our Normative System (NS)

How could we model our incentive mechanism to push agent to be honest ?

		Company	
		M	H
Supervisor	$\overline{S}$	m p_s	0 0
	S	p_c $c + b$	r c

TABLE – Prisoner’s dilemma

Obvious hypothesis

- $0 < m$

Let’s go back to our Normative System (NS)

How could we model our incentive mechanism to push agent to be honest ?

		Company	
		M	H
Supervisor	$\overline{S}$	m p_s	0 0
	S	p_c $c + b$	r c

TABLE – Prisoner’s dilemma

Obvious hypothesis

- $0 < m$
- $p_c < 0$ and $0 < r$

Let’s go back to our Normative System (NS)

How could we model our incentive mechanism to push agent to be honest ?

		Company	
		M	H
Supervisor	$\overline{S}$	m p_s	0 0
	S	p_c $c + b$	r c

TABLE – Prisoner’s dilemma

Obvious hypothesis

- $0 < m$
- $p_c < 0$ and $0 < r$

Deductions

- $p_s < c < 0$ if we want that a_s makes its job (chooses action S sometimes)
- A first NE is (S, H)

What about the other agent’s point of view ?

Let’s go back to our Normative System (NS)

How could we model our incentive mechanism to push agent to be honest ?

		Company	
		M	H
Supervisor	$\overline{S}$	m p_s	0 0
	S	p_c $c + b$	r c

TABLE – Prisoner’s dilemma

Obvious hypothesis

- $0 < m$
- $p_c < 0$ and $0 < r$
- $0 < b$ (**we notice that** b **is not relevant since** $p_s < c < 0$ **are sufficient**)

Deductions

- $p_s < c < 0$ if we want that a_s makes its job (chooses action S sometimes)

Let's go back to our Normative System (NS)

How could we model our incentive mechanism to push agent to be honest ?

		Company	
		M	H
Supervisor	$\overline{S}$	m p_s	0 0
	S	p_c $c + b$	r c

TABLE – Prisoner's dilemma

Obvious hypothesis

- $0 < m$
- $p_c < 0$ and $0 < r$
- $0 < b$

Deductions

- $p_s < c < 0$ if we want that a_s makes its job (chooses action S sometimes)

Let’s go back to our Normative System (NS)

How could we model our incentive mechanism to push agent to be honest ?

		Company	
		M	H
Supervisor	$\overline{S}$	m p_s	0 0
	S	p_c $c + b$	r c

TABLE – Prisoner’s dilemma

Obvious hypothesis

- $0 < m$
- $p_c < 0$ and $0 < r$
- $0 < b$ (we notice that b is not relevant since $p_s < c < 0$)

Deductions

- $p_s < c < 0$ if we want that a_s makes its job (chooses action S sometimes)

Conclusion

There are two Nash Equilibriums : $(\overline{S}, H)$ and (S, H).

Other applications with Normative Systems

Study other incentive mechanisms [Boella *et al.*, 2007]

Violation games interacting with normative systems, obligation mechanism, with applications in trust, fraud and deception.

Institutionalized games counts-as mechanism, with applications in distributed systems, grid, p2p, virtual communities.

Negotiation games MAS interaction in a normative system, norm creation action mechanism, with applications in electronic commerce and contracting.

Norm creation games multiagent system structure of a normative system, permission mechanism, with applications in legal theory.

Control games interaction among normative systems, nested norms mechanism, with applications in security and secure knowledge management systems.

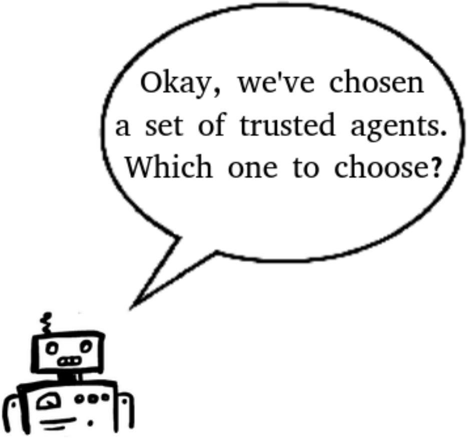

Okay, we've chosen a set of trusted agents. Which one to choose?

Table of Contents

Social Choice Theory (SCT)

Short description

The social choice theory aims at **studying and analyzing** how the combination of individual opinions can lead, at **a collective** level, to a **ranking of possible choices**. It models the decision-making process of agents in a voting system.

Examples of vote systems modeled by SCT

Majoritarian systems is a system in which candidates have to receive a majority of the votes to be elected

Plurality systems is a system in which the candidate(s) with the highest number of votes wins, with no requirement to get a majority of votes.

Proportional systems is a system in which divisions in an electorate are reflected proportionately in the elected body

. . .

A voting system

A formal definition

We call **voting system** a N-uplet $S = (\mathcal{N}, \mathcal{O}, \{\succeq_i\}_{i \in \mathcal{N}}, f)$ where :

- $\mathcal{N}$ is a nonempty set of agents
- $\mathcal{O}$ is a nonempty set of possibles outcomes (possible candidates)
- $\{\succeq_i\}_{i \in \mathcal{N}}$ is a set of binary relations on $\mathcal{O}$
  An element $\succeq_i$ is called a **preference profile** and the **set of all possible preference profiles** is written $\mathbb{P}$.
- $f : \mathbb{P}^{\mathcal{N}} \to \mathbb{P}$ is called the **social choice function**
  It represents the result of a vote in the form of a preference profile

Useful properties to model voting systems

Definition of Total Strict Order

We call $\succ_i$ a **Total Strict Order** (TSO) a binary relation on $\mathcal{O}$ s.t.

1. $\forall o, o', o'' \in \mathcal{O}$: if $o \succ_i o'$ and $o' \succ_i o''$then $o \succ_i o''$ (transitivity)
2. $\forall o \in \mathcal{O} : \neg(o \succ_i o)$ (irreflexivity)
3. $\forall o, o' \in \mathcal{O} : o \succ_i o'$ or $o' \succ_i o$ (total)

Standard properties

- $\succeq_i$ is reflexive iff $\forall o \in \mathcal{O} : o \succeq_i o$
- $\succeq_i$ is symmetrical iff $\forall o, o' \in \mathcal{O} : o \succeq_i o' \Rightarrow o' \succeq_i o$
- $\succeq_i$ is antisymmetric iff $\forall o, o' \in \mathcal{O} : o \succeq_i o' \wedge o' \succeq_i o \Rightarrow o = o'$
- $\succeq_i$ is acyclic iff there is no $(o_1, \ldots, o_n) \in \mathcal{O}^n, o_1 \succeq_i o_2 \wedge \ldots \wedge o_{n-1} \succeq_i o_n \wedge o_1 = o_n$
- $\succeq_i$ is a **pre-order** iff $\succeq_i$ is reflexive and transitive
- $\succeq_i$ is a **order** iff $\succeq_i$ is an antisymmetric **pre-order**

Borda system

Definition [de Borda, 1781]

A voting system $S = (\mathcal{N}, \mathcal{O}, \{\succ_i\}_{i\in\mathcal{N}}, f)$ is called a **Borda system** if for all preference profiles $P = \{\succ'_i\}_{i\in\mathcal{N}}$ defined on $\mathcal{O}$ with $\forall i \in \mathcal{N}, \succ'_i$ is a TSO, the social choice function f is such that :

$$\text{If } \succ = f(P) \text{ then } (\forall o, o' \in \mathcal{O}, o \succ o' \text{ iff } g_{S,P}(o) \geq g_{S,P}(o'))$$

where $g_{S,P} : \mathcal{O} \to \mathbb{N}$ is a counting function of Borda i.e. :

$$\forall o \in \mathcal{O}, g_{S,P}(o) = \sum_{i\in\mathcal{N}} (|\mathcal{O}| - (|\{o' \in \mathcal{O} : o' \succ'_i o\}| + 1))$$

with $|\cdot|$ is the cardinal function

- Each options are associated with points that depends on each agent's preference profile
- A winner of the vote is a candidate $o \in \mathcal{O}$ with the most points i.e. $o \in \underset{\mathcal{O}}{\operatorname{argmax}}\, g_{S,P}$

Example of manipulation strategies in Borda system

A situation

Let consider $\mathcal{N} = \{1, 2, 3\}$ a set of three agents. These agents must choose between four options $\mathcal{O} = \{A, B, C, D\}$. We consider the following genuine preference profiles :

- $1 : A \succ_1 B \succ_1 C \succ_1 D$;
- $2 : C \succ_2 B \succ_2 A \succ_2 D$;
- $3 : B \succ_3 A \succ_3 D \succ_2 C$.

Questions

1. For these 4 options, each agent gives 3 points to the option he prefers, 2 to the second, 1 to the third and 0 to the last. Give a table of this vote.
2. If all agents provide their true preference profile, who is the winner ?
3. Agent 1 knows about others' preferences. How could this agent lie in such a way as to have a favorable outcome ?

Example of manipulation strategies in Borda system

		Options			
		A	*B*	*C*	*D*
Votes	1	3	2	1	0
	2	1	2	3	0
	3	2	3	0	1
Borda count		6	7	4	1

TABLE – The winner is the candidate *B* **if all agents are honest**

Example of manipulation strategies in Borda system

		Options			
		A	*B*	*C*	*D*
Votes	1	3	0	1	2
	2	1	2	3	0
	3	2	3	0	1
Borda count		6	5	4	3

Table – **If agent** 1 **lies**, then the winner is the candidate *A*

Example of manipulation strategies in Borda system

		Options			
		A	*B*	*C*	*D*
Votes	1	3	2	1	0
	2	1	2	3	0
	3	2	3	0	1
	4	3	2	1	0
Borda count		9	8	5	1

TABLE – **If agent** 1 **introduces a Sybil agent** 4, then *A* wins

A bit of modeling : plurality system

Exercice

Plurality rule consists of electing the alternative ranked first most often (i.e. each voter assigns 1 point to an alternative of her choice, and the alternative receiving the most points wins)

- Give a formal definition of a plurality system
- Propose an idea to manipulate this system

A bit of modeling : plurality system

An example of an acceptable solution

A voting system $S = (\mathcal{N}, \mathcal{O}, \{\succ_i\}_{i\in\mathcal{N}}, f)$ is a *plurality system* if for all preference profiles $P = \{\succ'_i\}_{i\in\mathcal{N}}$ defined on $\mathcal{O}$ with $\forall i \in \mathcal{N}, \succ'_i$ is a TSO, the social choice function f is such that

$$\text{If } \succeq = f(P) \text{ then } (\forall o, o' \in \mathcal{O}, o \succeq o' \text{ iff } g_{S,P}(o) \geq g_{S,P}(o'))$$

where $g_{S,P} : \mathcal{O} \to \mathbb{N}$ is a counting function s.t.

$$\forall o \in \mathcal{O}, g_{S,P}(o) = |\{i \in \mathcal{N} : \neg(\exists o' \in \mathcal{O}, o' \succ_i o)\}|$$

with $|\cdot|$ is the cardinal function

- A winner is a candidate o that has the maximum of points i.e. that belongs to $\underset{\mathcal{O}}{\operatorname{argmax}}\, g_{S,P}$
- **It is easy to see that a Sybil attack works in this kind of vote system but lying does not work !**

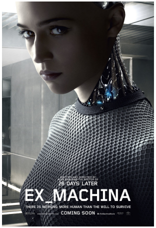
Examples of manipulations in voting systems

Table of Contents

Okay, we've been talking all along this course about manipulation..

But what is a manipulation ?

Manipulation in computer systems



Examples of manipulation strategies

Reputation systems : lying about its identity in order to self-promote

Voting systems : revealing a false preference profile to influence an outcome

Peer-to-peer networks : making a node believes it is alone in the network

Understanding the meaning of manipulation

"manipulation is **not exactly coercion**, **not precisely persuasion**, and **not entirely similar to deception**" [Handelman,2009]

Understanding the meaning of manipulation

"manipulation is **not exactly coercion**, **not precisely persuasion**, and **not entirely similar to deception**" [Handelman,2009]

It is not coercion

Pointing a gun at someone to get them to do something is coercion.
⇒ This is not manipulation because **this is not concealed**.

Understanding the meaning of manipulation

"manipulation is **not exactly coercion**, **not precisely persuasion**, and **not entirely similar to deception**" [Handelman,2009]

It is not coercion

Pointing a gun at someone to get them to do something is coercion.
⇒ This is not manipulation because **this is not concealed**.

Nor persuasion

Telling someone to sort their garbage because they call themselves an environmentalist is persuasion.
⇒ It is not manipulation because **the person who persuades is sincere**.

Understanding the meaning of manipulation

"manipulation is **not exactly coercion**, **not precisely persuasion**, and **not entirely similar to deception**" [Handelman,2009]

It is not coercion

Pointing a gun at someone to get them to do something is coercion.
⇒ This is not manipulation because **this is not concealed**.

Nor persuasion

Telling someone to sort their garbage because they call themselves an environmentalist is persuasion.
⇒ It is not manipulation because **the person who persuades is sincere**.

Nor similar to deception

Making children believe that Santa Claus exists is a lie.
⇒ This is not manipulation because **there is no intention to instrumentalize**.

Manipulation is a deliberate concealed influence

In reputation systems and voting systems

A manipulator may use trust to **influence** other agents to interact with him or prevent agents from interacting with others. An agent may also use voting systems to **influence a collective decision**.
To do so, he applies a **certain strategy** that he calculated beforehand (we say that this strategy has been **deliberated**) e.g. :

- whitewashing
- treachery
- lying about its preference profile
- sybil attacks

Another main characteristic of manipulation is that **to be operational, a manipulation strategy must always be concealed** from other agents.

If you know that I'm lying and we assume that you are a rational person, then why would you consider my (false) preferences or my (false) testimonies ?

To sum up the main characteristics of a manipulation

1. manipulation is *premeditated* :
   ⇒ it is **a deliberate effect of a manipulator**

2. manipulation is *an instrumentalization* :
   ⇒ it is **an influence exercised on a victim**

3. manipulation is *invisible* when it happened :
   ⇒ it is **always hidden from the victim**

To sum up the main characteristics of a manipulation

1. manipulation is *premeditated* :
   ⇒ it is **a deliberate effect of a manipulator**

2. manipulation is *an instrumentalization* :
   ⇒ it is **an influence exercised on a victim**

3. manipulation is *invisible* when it happened :
   ⇒ it is **always hidden from the victim**

To sum up the main characteristics of a manipulation

1. manipulation is *premeditated* :
   ⇒ it is **a deliberate effect of a manipulator**

2. manipulation is *an instrumentalization* :
   ⇒ it is **an influence exercised on a victim**

3. manipulation is *invisible* when it happened :
   ⇒ it is **always hidden from the victim**

To sum up the main characteristics of a manipulation

1. manipulation is *premeditated* :
   ⇒ it is **a deliberate effect of a manipulator**

2. manipulation is *an instrumentalization* :
   ⇒ it is **an influence exercised on a victim**

3. manipulation is *invisible* when it happened :
   ⇒ it is **always hidden from the victim**

A general definition of manipulation in MAS

Definition [Leturc and Bonnet, 2020]

A *manipulation* in a Multi-Agent System is a **deliberate effect** of an agent (called a *manipulator*) to **influence** another agent (called a *victim*), while making sure to **conceal that effect**.

Connexion with other fields

At the frontiers of social science and computer science

In Psychiatry :

Manipulation is defined as the instrumentalization of a victim in the interest of the manipulator [Kligman, 1992]

Connexion with other fields

At the frontiers of social science and computer science

In Psychiatry :

Manipulation is defined as the instrumentalization of a victim in the interest of the manipulator [Kligman, 1992]

In politics and marketing :

Manipulation is the act of altering the judgment of individuals, depriving them of some of their judgment and deliberate choices [Sunstein, 2015]

Connexion with other fields

At the frontiers of social science and computer science

In Psychiatry :

Manipulation is defined as the instrumentalization of a victim in the interest of the manipulator [Kligman, 1992]

In politics and marketing :

Manipulation is the act of altering the judgment of individuals, depriving them of some of their judgment and deliberate choices [Sunstein, 2015]

In Economics :

Manipulation is defined as the means of influencing the final outcome of a game by concealing the true intentions of the agent [Ettinger, 2010]

Any question ?

Thank you for listening