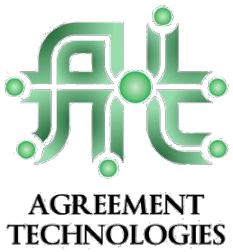
Social Conformity and its Convergence for Reinforcement Learning



Juan A. García-Pardo
Carlos Carrascosa
Jose Soler

2010 German Conference on Multi-Agent system Technologies (MATES)
Leipzig, Germany
September 27th, 2010

Outline

- Introduction
- Social Reinforcement
  - What it means, what we want it for
  - Reinforcement Learning
  - Exploration when things go wrong
- Convergence
  - Assumptions
- Study Case
- Conclusions

Introduction

- Agents in changing environments can detect changes?
- Goal: mechanism for easy agent adaptation
- Changes:
  - Environmental (e.g. new rules for a game)
  - Social (e.g. an agent leaves the game)

Social Reinforcement

- Every agent can give its opinion
  - Things are going as expected (positive values)
  - Things are going bad (negative values)
- The sum of all opinions would be the global opinion of this society
  - Thus society wants a good global opinion
- Every agent pursues its own goal, though they share a final “purpose” i.e. the goal of the MAS
- It is a locally sensed global status of the society.

Social Reinforcement

- Meaning: It is a locally sensed global status of the society.
- It is related to the “environmental reinforcement”, but it is not the same
- Different agents can have different opinions even if the world is static (not changing)
  - E.g. in a game when a character is being attacked opposed to a character which is “winning a battle”

Social Reinforcement

- We want to use it to characterize the status of the society.
- If the social reinforcement (SR) is bad (i.e. the global opinion shows society is doing bad) it is a sign of the need of a change.
- To take into account that SR can fluctuate, how can we “detect” these changes?

Reinforcement Learning

- Environment guides behavior.
- Needs time to explore the world.
  - Afterwards agents can use the information to behave optimally.
- Beforehand we do not know how much exploration the agents will need to gain the information (unrestricted environment).

Reinforcement Learning

- Furthermore, when the environment is dynamic the information gathered may not be correct at a given moment.
- Agents know when the expected value of an action in the world differs from the obtained one. If it is less than expected the opinion will be bad.
- SR can help agents to decide whether to explore or not, by assuming global welfare should prevail.

Reinforcement Learning

- There are some concerns:
  - Agents should be able to communicate with the society with as many members as possible.
  - Arrow's paradox does not apply here (social welfare): the voting is already ordered (positive is always preferable).
  - Fluctuations always will be there: need to "smooth" the SR in the time domain: e.g. time series.
  - The convergence of the society can be towards a non-optimal state, as could happen in RL when exploring too little.

Reinforcement Learning: exploration

- How to modify the exploration probability?
- One possibility is to use this time series:

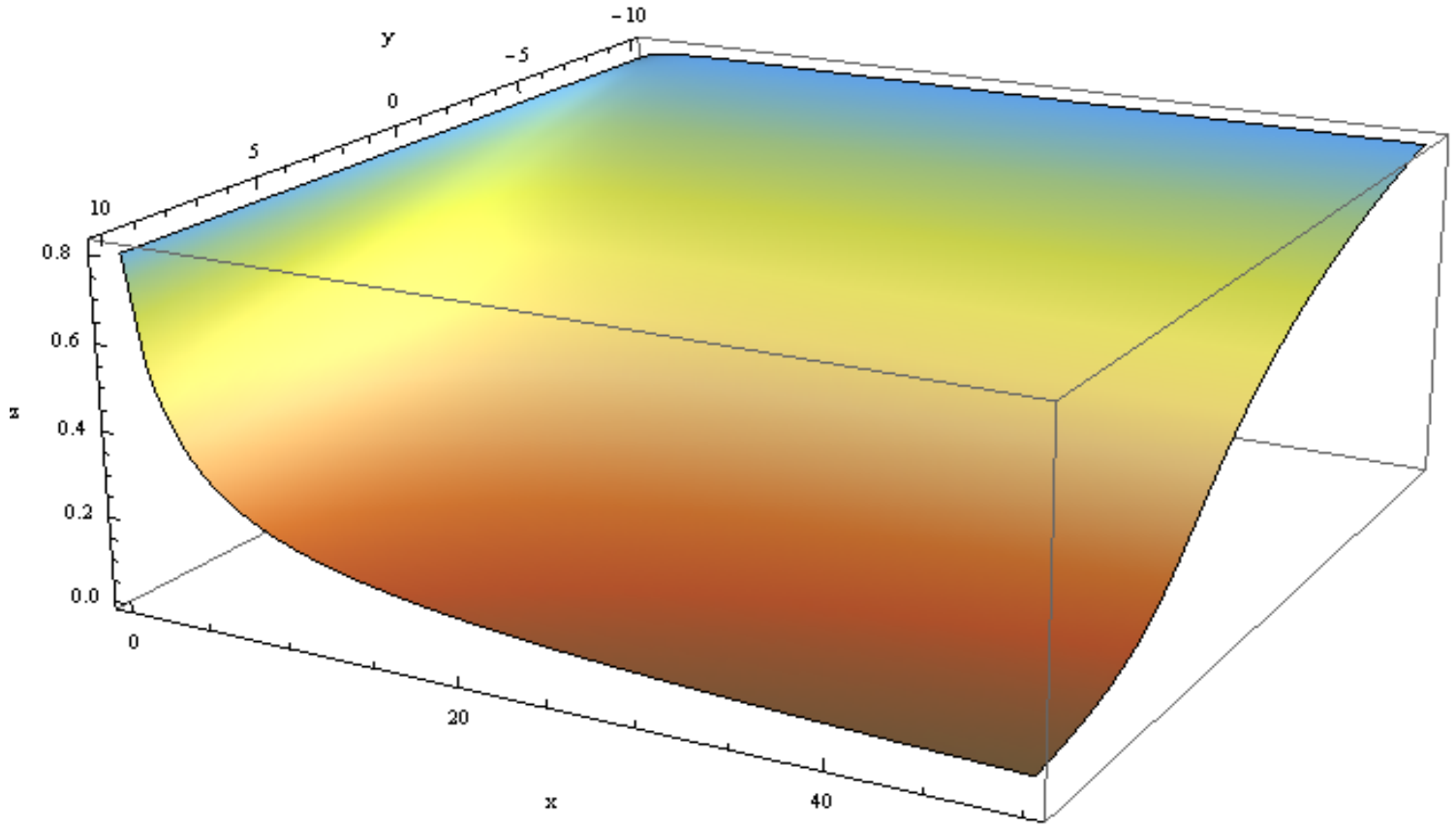
$$k_{t+1} = \frac{k_t}{k_t + 2^{k_t(\mathrm{SR}-2)}}$$

Behaves similarly to a sigmoid function, depending on the value of SR at a given moment.

Opinion is]0,+∞[if expected value was lower or equal than the observed one,]0,-∞[otherwise.

Reinforcement Learning: exploration

- Y-axis: SR. X-axis: time. Z-axis: value of k.



Convergence?

- With the use of a time series a study of its convergence is always desirable.
- For the RL algorithm to converge it is necessary that the series converge as well.
- Since the environment is dynamic, convergence will be proof only in the stationary segments of time (when there are no changes), as classical RL algorithms.

Convergence: Assumptions

- The transition between states are governed by a probability distribution, which does not change.
- The same applies to the rewards.
- Each agent's opinion is trustworthy: for now no one lies.
- Everything else is allowed.

Convergence: Assumptions

- Rewriting the series as $k_{t+1} = \frac{k_t}{k_t + 2^{k_t(\sigma-2)}}$
- In the limit, we expect $t \longrightarrow \infty \quad k_{t+1} = k_t$
- Then the series converge to:

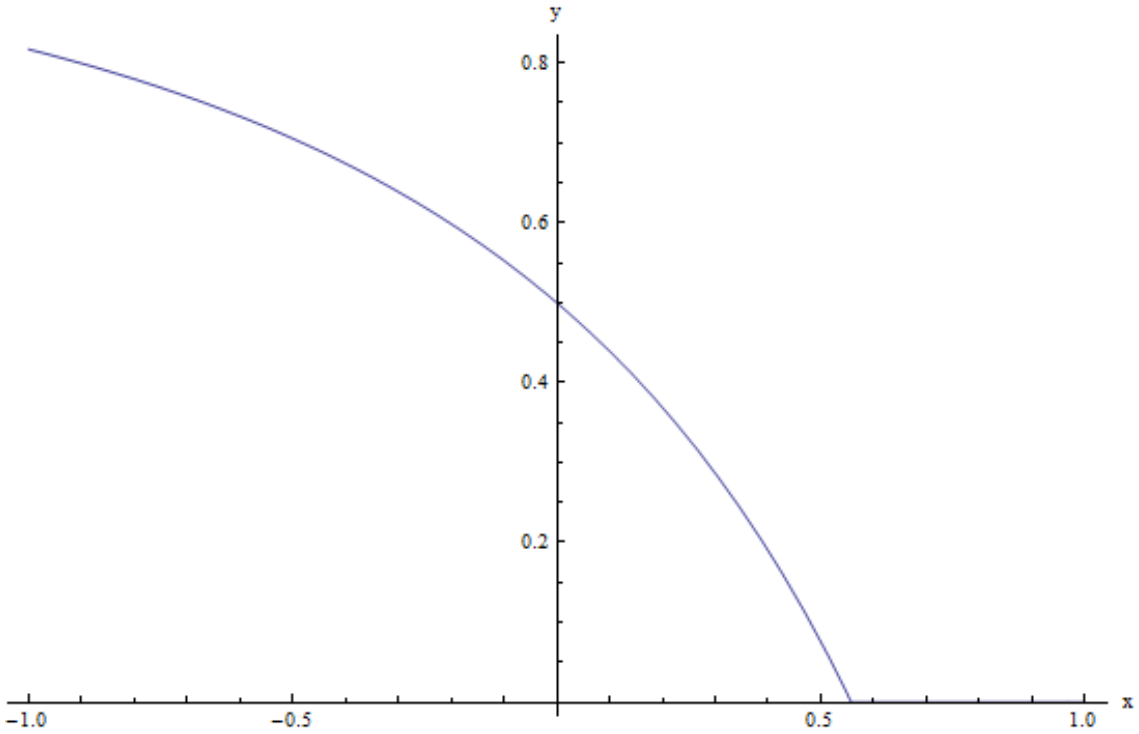
$$k_t = 1 - \frac{W(ln(2)2^{\sigma-2}(\sigma-2))}{(\sigma-2)ln(2)}$$

- W(x) is the Lambert-W, such as

$$\forall x \in \mathbb{R} \quad x = W(x)e^{W(x)}$$

Convergence:

- Convergence (infinite time) of k_t as SR varies from -1 to +1:

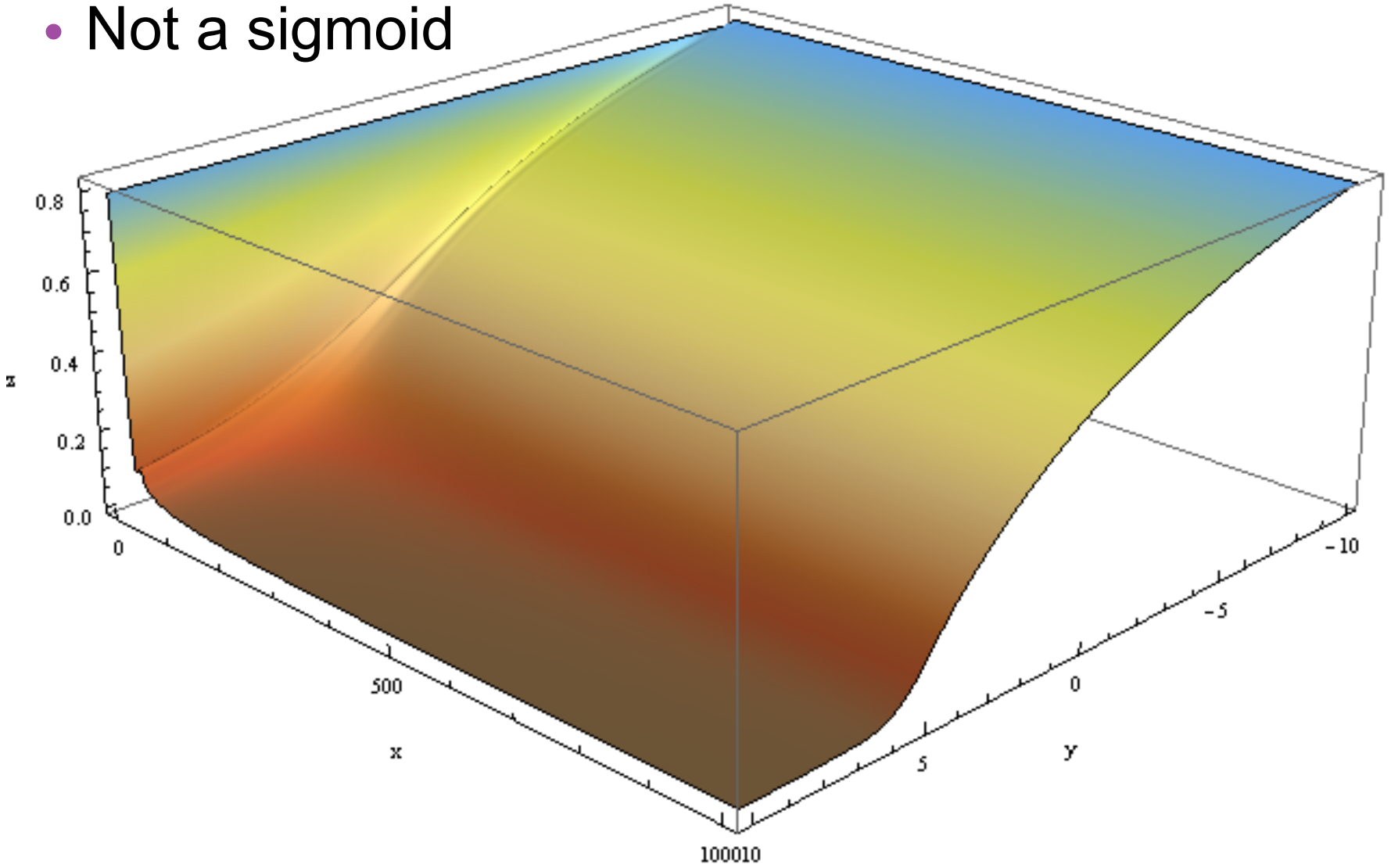


Convergence:

- But it does not converge to a sigmoid-like function?
- Intuition is sometimes tricky.
- It looked like it was converging towards a sigmoid, but it has been proven not to.

Convergence:

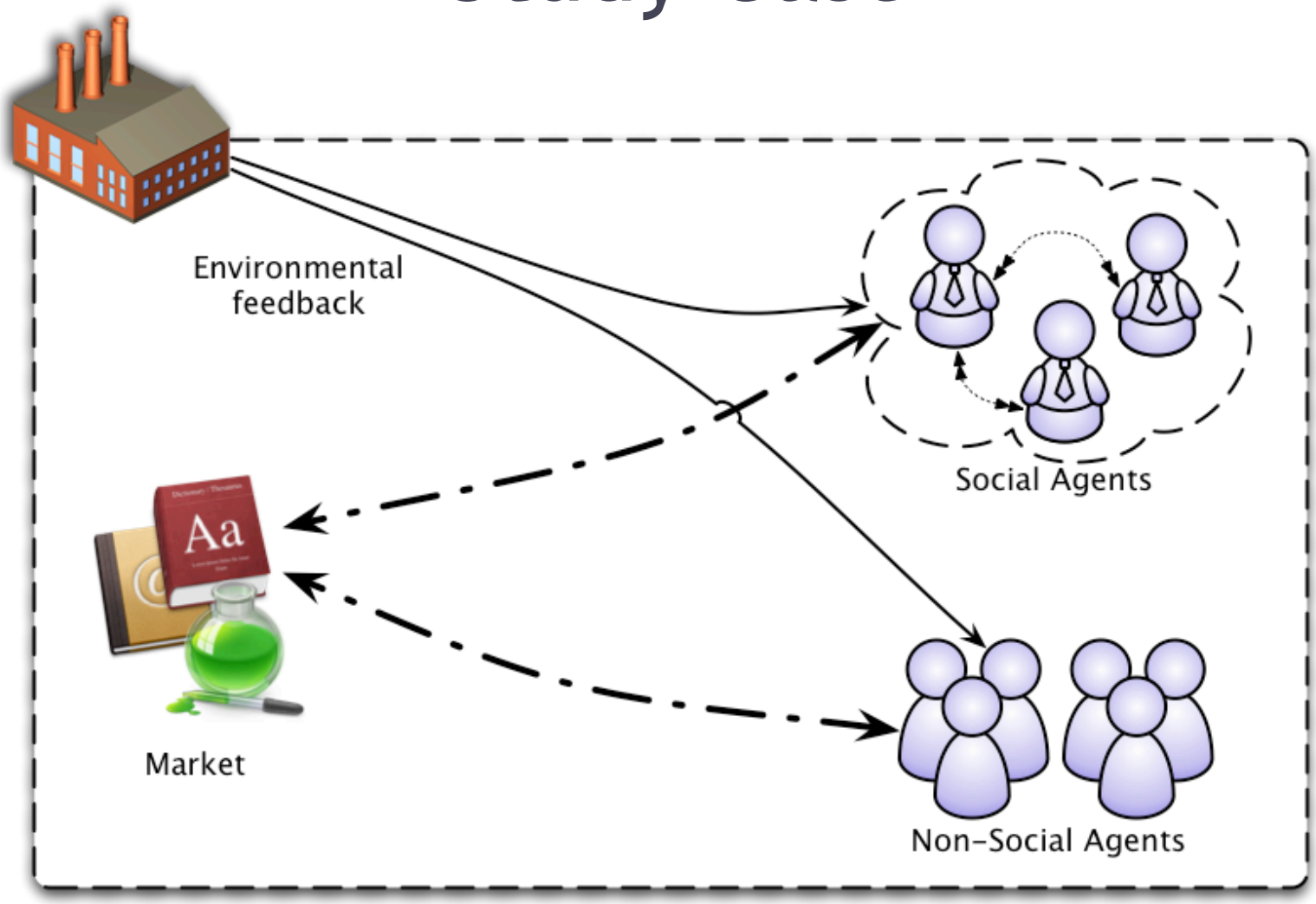
- Not a sigmoid



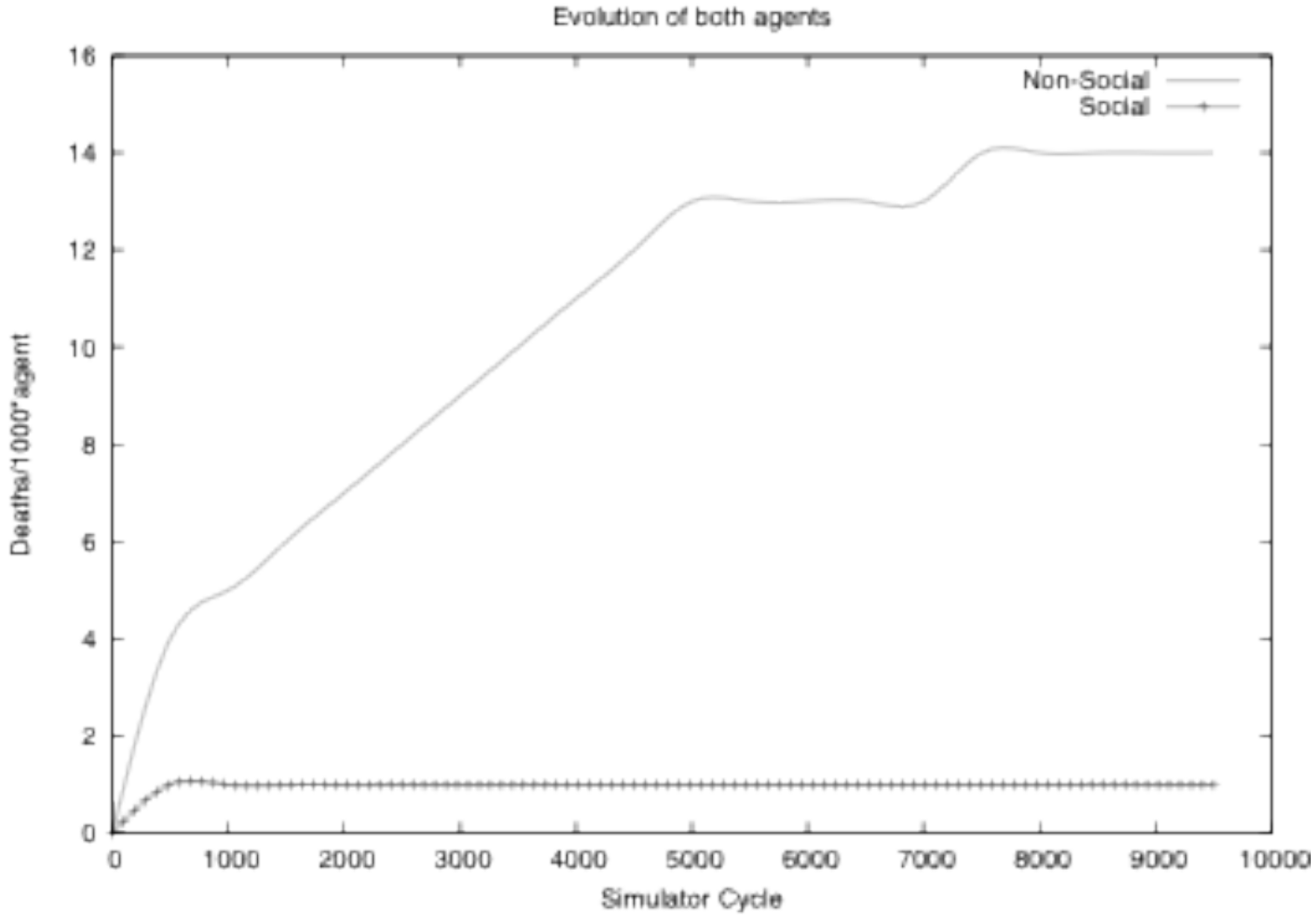
Study Case

- Small study case to empirically test this idea.
- Supply Chain with the ability to change rules of production and in an open system (agents can enter and leave).
- We expect to see a better performance of the social-aware agents when environment change.
- Just showing one example: the production rules change and so does the social environment: from 1 agent we move to 4 agents.

Study Case


Environmental feedback
Social Agents
Market
Non-Social Agents

Study Case


Evolution of both agents
Non-Social
Social
Deaths/1000*agent
Simulator Cycle
0
2
4
6
8
10
12
14
16
0
1000
2000
3000
4000
5000
6000
7000
8000
9000
10000

Study Case

- The figure shows the evolution of “error” beginning with the social and environmental change. X-axis is time, Y-axis is deaths per agent, meaning agents which were unable to accomplish productions goals in a certain amount of time.

- At the end of 10K cycles the “error” (deaths/1000) for the traditional RL agent in a 2-agent society was 7.4 while for the SR-aware was 1.85. Moreover, when in a 4-agent society, it was 14.05 for the traditional RL and 1.35 for the SR-aware.

Conclusions

- As we expected the SR agents perform better when there are social and environmental changes
- When increasing the number of agents in the society, traditional RL performs worse because each agent is not modeled by the rest of the society, it is a source of non-determinism.
- But increasing the number of agents in this study case makes the SR mechanism work better, since the “social feedback” is stronger.

Conclusions

- The agents explore when it seems necessary, and exploit their knowledge otherwise.
- It is a mechanism which can be easily implemented in traditional single-agent RL algorithms to modify the exploration probability.
- A further study on trust and reputation and its implications in the Social Reinforcement is required.

Thank you for your Attention

Contact author:

Juan A. García-Pardo
Universidad Politécnica de Valencia
Email: jgarciapardo@dsic.upv.es

GTI IA
Grupo de Tecnología Informática
Inteligencia Artificial