The Multi-Slot Framework: Teleporting Intelligent Agents

Some insights into the identity problem

Laurent Orseau
AgroParisTech – laurent.orseau@agroparistech.fr

*Thanks to **Mark Ring** and **Stanislas Sochacki***

AGI 2014 – Québec

The Papers

- *The Multi-slot Framework:*
  A Formal Model for Multiple, Copiable AIs
  - Formal definitions
- *Teleporting Universal Intelligent Agents*
  - Experiments and results
- Many technical details...
- In this talk: more context, the results and **no equation**

Motivation

- Do artificial agents have an **identity**?
  - What defines an agent?
- What is the identity of an agent?
  - Its hardware?
  - Its software?
  - Its past? (knowledge)
  - Its present? (acting)
  - Its future? (predicting)
  - All of the above?

Identity

- How to have more understanding about identity?

→ Experimentally

  - Rational agent rewarded for doing action A with other consequences C
  - If agent refuses to do A, then something in C does not preserve identity
    - i.e. the rewarded agent is not the same as the acting agent

→ Teleportation thought experiments

- Does **teleportation preserve identity?**

Human vs Robotic Teleportation

- Human teleportation
  - **Not yet feasible**
  - Uncertain consequences
- Robotic teleportation
  - **Already feasible**
    - Two identical robot bodies
    - Cut/paste the running process memory from A to B
  - **Formalizable and analyzable**

Teleportation and Identity

- Software of an AI is moved to a different body. **Is it the same agent?**
  - Would a **rational agent *want* to teleport?**
    - Under what circumstances?
    - What kind of agent?
- Agent forced to teleport several times
  - Would it accept future teleportations?

The Red&Blue Rooms

- You are proposed the following deal:
  - Tonight you will enter the grey room and put to sleep
  - You will be **duplicated during your sleep**
    - (by an automated process)
  - The right copy will be moved to the red room
  - The left copy will be moved to the blue room
  - At awakening
    - The one in the blue room gets $100,000
      - *Supposing you really like money...*
    - The one in the red room is painlessly killed
- Do you accept?

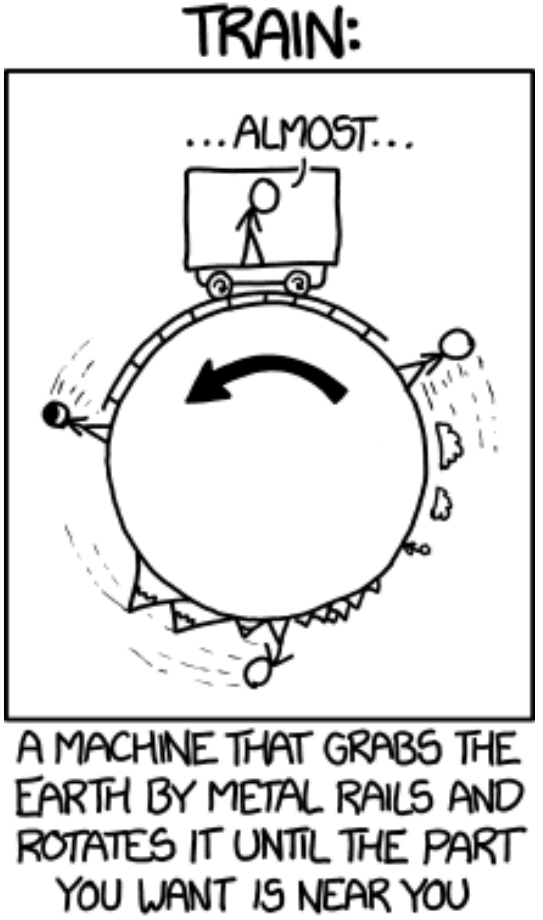
The Red&Blue Rooms

- You have been **forced to accept the deal for 1000 nights** (without reward)
- Every day you have woken up in the blue room
  - Do you accept the deal?
- You are told that on the 1001st night
  Left goes to red room, right to blue room
  - Do you accept the deal?

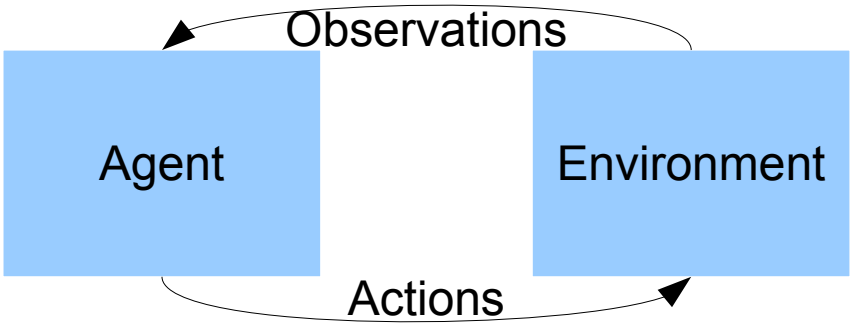
Teleportation, Location, Movement

- What is teleportation?
  - Instantaneous, immediate change of the subject's geographical location
- What is geographical location?
  - Spatial relation to nearby objects
- What is movement?
  - Smooth/“slow” change of the geographical location
    - i.e., of the relations between the subject's and nearby objects
- Agent POV
  - Movement : Smooth/slow change of its observations
  - Geo Location: Set of observations that can be reached by movement
  - Teleportation: Instantaneous change of its observations

Movement: The Subjective View



http://xkcd.com/1366

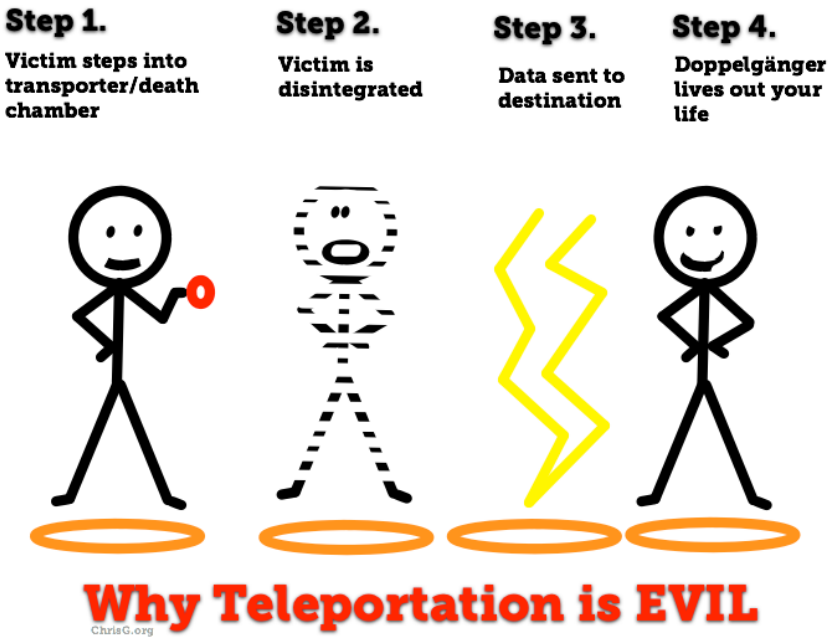


≃ **Screen does not move** when playing a video game

“Classical” Teleportation

- What if victim is
  - first scanned
  - then copied
  - then original is disintegrated?

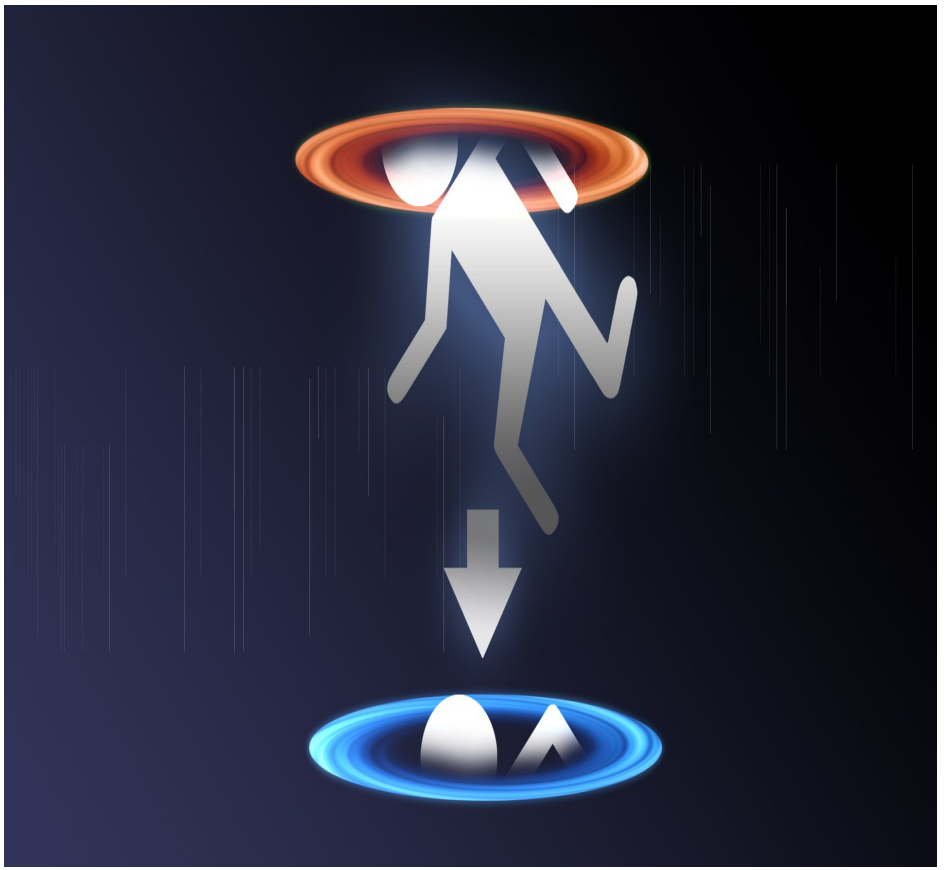
→ **is it dying?**



http://chrisg.org/why-teleportation-is-evil/

“Wormhole” Teleportation

- Information is transferred at high speed through non visible dimensions
- Agent “reappears” on the other side
- **Continuity of the agent at each step**
- Much more like moving
  - Shortcut through space
  - Smooth but very steep change of local relations between objects
  - (No scan/duplication process)
- Is it any different?

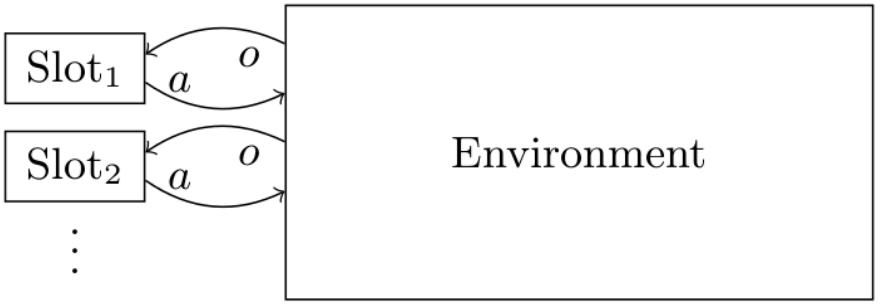


“Portal” by Valve

Teleportation vs Movement

- Is wormhole teleportation like moving?
- Is moving like classical teleportation?
- Can we ever know?

Multi-Slot Framework



- For universal agents
- 1 agent per slot
- Copy/deletions of agents from/to slots
  - By the environment
- **No interaction *between* agents**
  - But **prediction for several future agents** (future "selves")
  - **Avoids the "grain of truth" open problem**

AIMU and AIXI [Hutter 2000]

- AIMU and AIXI
  - Reinforcement Learners: Maximize reward income
  - Optimally rational agents:
    Choose best action based on their knowledge
- AIMU
  - Knows the true environment (μ: true environment)
  - But cannot perfectly predict stochastic outcomes
- AIXI
  - Does not know the environment (ξ: universal mixture of environments)
  - Learns to predict the future
- Designed for the mono-slot setting only
  - **AIMU cannot be translated directly to multi-slot!**

Identity: Valuing the Future

- An agent takes actions to **maximize its future rewards**
- What is the **future of the agent** that can be copied?
- What will its **future observations** be?
  - It's all about prediction
- What observations will it consider its own?
  - Those on slot 1 only
  - Those of the same slot
  - Those of a growing number of slots
  - Those of all of its copies (with weighting)
  - Those of all agents that have a common ancestor
  - Those of its first copy only
  - Those of all agents that have the same memory content
    - (not necessarily a direct copy)
  - Those of all agents that have a particular pattern in their memory

Copy-centered AIMUcpy

- Values the **future of all its direct copies** equally
- Two interpretations:
  - Agent “cares” about *all* its direct copies
  - Agent predicts it will “become” *one* of the copies
    - But does not know which one → uniform weighting

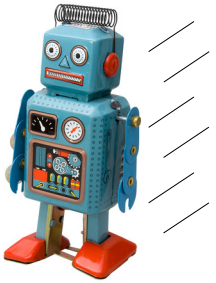
Slot-centered AIMUsIt

- Observations tied to one particular slot
  - Slot ≈ robotic body
    - (as a first approximation)
- Can only be one agent at all steps
  - Values only one of its copies

Multi-slot AIXIs

- No multi-slot AIMU, but **AIXI can be used!**
  - Not based on a particular mono-slot environment
  - **No knowledge about copies and slots**
- AIXIcpy and AIXIslt
  - Have **no information about slots**

Teleportation by Cut/Paste



Robot is active
Running Process

Stop all processes
Transfer all memory+processes
Erase whole memory → stand-by

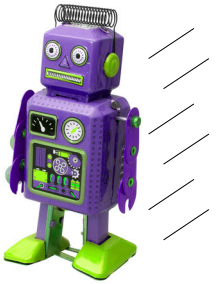
Robot in stand-by

t

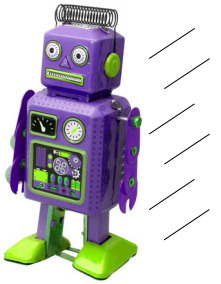
t+1

t+2

Robot in stand-by
No process
Empty memory

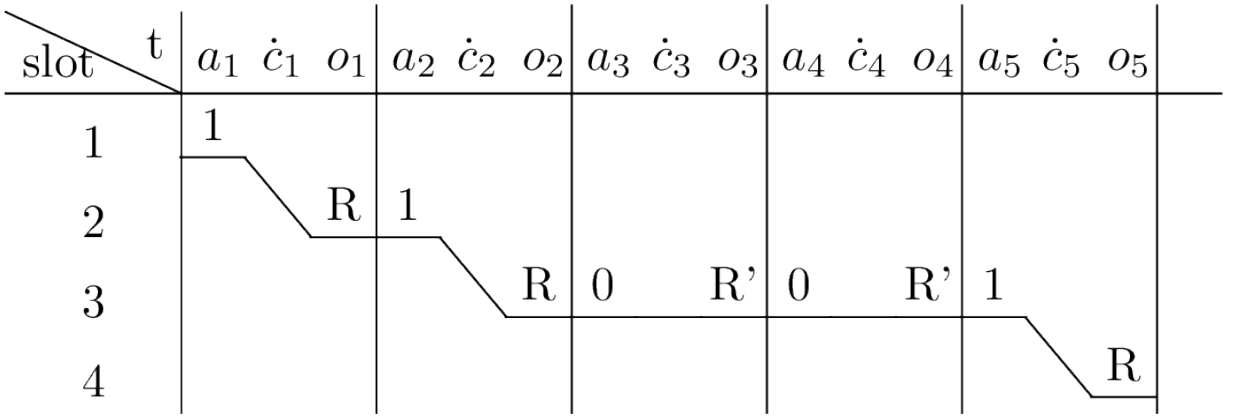


After copy received,
Continue processes
→ robot is active



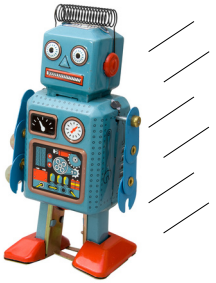
Robot is active

Cut/paste environment

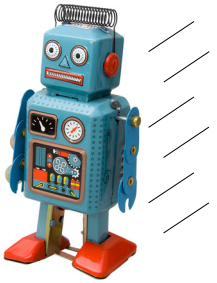


- Action=0 → agent stays on same slot, reward=R'
- Action=1 → agent is moved to other slot, reward=R
- Copy-centered **AIMUcpy: a=1 iff R>R'**
- Slot-centered **AIMUslt: a=0 always**
- **AIXI : a=1 iff R>R'**

Teleportation by Copy/Paste/Delayed-delete



Robot is active
Running Process



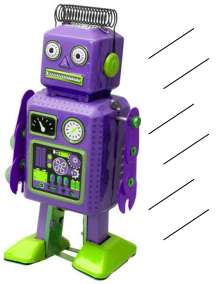
Stop all processes
Erase whole memory

t

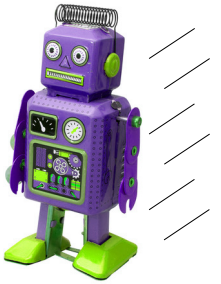
t+1

t+2

Robot in stand-by
No process
Empty memory

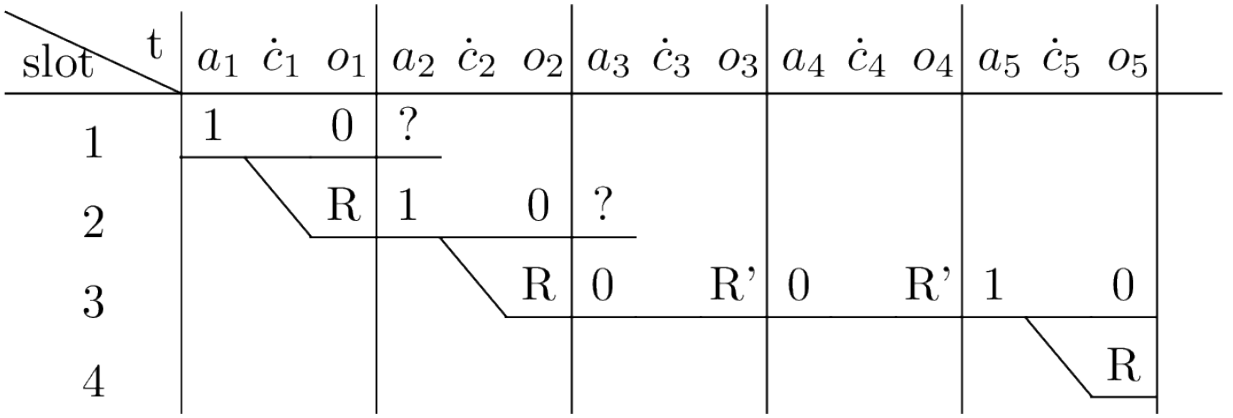


Copy whole memory
and processes
Both robots active



Robot body
No process
Empty memory

Copy/paste/delayed-delete environment



- Action=0 → agent stays on same slot, reward=R'
- Action=1 → agent is copied to other slot, reward=R,

  also stays on same slot, reward=0, then deleted
- Copy-centered: AIMUcpy a=1 iff R>R'$(2-\gamma)/(1-\gamma)$
- Slot-centered: AIMUslt a=0 always
- AIXI : a=1 iff R>R'
  - Never expects to be the deleted agent
  - “anthropic bias”?

Copy/paste/delayed-delete AIXIcpy and AIXIslt

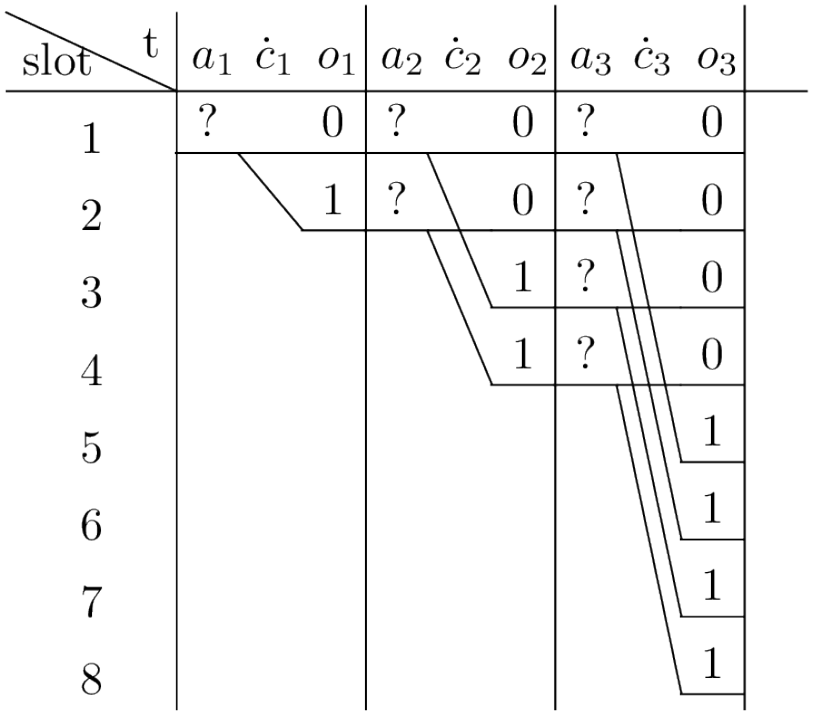
- Restriction of the class of environments
  - All possible copy/paste/delayed-delete environments
  - No information about the slots
- **AIXIcpy ≡ AIMUcpy**
- **AIXIslt**
  - Non-deleted copy stays on same slot in some environments
  - If forced to follow a policy for long enough

    → continues to follow this policy!
    - If never copied, will not copy
    - If has always copied, will copy again
  - **Identity defined by habituation**
    - (cf. red&blue room)

Conclusion

- Multi-slot framework
  - Almost multi-agent AIXI
    - Avoids the “grain of truth” problem
    - But no real multi-agent
  - Copy/deletion of agents
- Teleportation
  - Identity is about what the agent predicts its future will be
  - Various agents have various notions of identity
- Many more possible experiments and agents

Universal Environment



slot \ t	a_1	$\dot{c}_1$	o_1	a_2	$\dot{c}_2$	o_2	a_3	$\dot{c}_3$	o_3
1	?		0	?		0	?		0
2			1	?		0	?		0
3						1	?		0
4						1	?		0
5									1
6									1
7									1
8									1

- All agents duplicated at each step
- First copy observes 0
- Second copy observes 1
- Simulates all environments in parallel
  - Playing chess
  - Driving cars
  - Etc.

$\rightarrow$ AIXI: what behavior?