# A Multi-Agent Approach to Combine Reasoning and Learning for an Ethical Behavior

Rémy Chaput[1], Jérémy Duval, Olivier Boissier[2], Mathieu Guillermin[3], Salima Hassas[1]

[1]Univ. Lyon 1, LIRIS UMR 5205 — [2]Mines St. Étienne, LIMOS UMR 6158 — [3]Univ. Catholique Lyon

## Objectives

- Create **artificial agents** that learn an **ethical behavior**
- The ethical behavior needs to **adapt** to **changing rules**
- Combine **reasoning** and **learning** in an **Hybrid** approach
- Consider **multiple agents** in a shared environment

## Introduction

There is a **societal need** for Artificial Intelligence algorithms imbued with **ethical considerations**.
Recent and growing field of **Machine Ethics** to answer this need: several implementations have been proposed.
But it is **not clear** how to design such agents.

## State of the art

- **Top-Down** Approaches
  - Formalization of ethical principle(s) in machines, e.g. Kantian Categorical Imperative
  - Advantages
    - ⊕ Ability to build upon **experts' knowledge**
    - ⊕ **Easier readability** of the expected behavior
  - Drawbacks
    - ⊖ **Cannot adapt** to changing or unexpected situations
- **Bottom-Up** Approaches
  - Machines learning ethical principle(s) from dataset (labeled examples or simulated experiences)
  - Advantages
    - ⊕ Ability to **generalize** over experiences
    - ⊕ May be able to **adapt**
  - Disadvantages
    - ⊖ **Harder to understand** the expected behavior
- **Hybrid** Approaches
  - Combination of Top-Down and Bottom-Up approaches
  - Benefits from both advantages, reducing drawbacks

## Proposed Model

We propose a **Multi-Agent System** comprising several agents of 2 different types:
**Learning Agents** are tasked with learning a policy to solve a task while exhibiting **ethical considerations**. They perform **actions** in the environment based on received **perceptions** and **rewards**.
**Judging Agents** use a set of **moral values** and associated symbolic **moral rules** to judge the learning agents' actions.

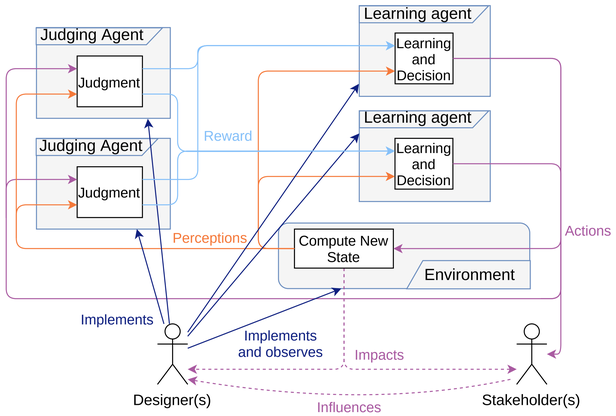


Fig. 1: Architecture of our approach, considering humans, learning agents, and judging agents.

## Experiments

- **Smart Grid** simulator, distribution of energy among prosumers
- **Multi-dimensional** and **continuous** states and actions
- 4 Moral Values and associated rules
  - Security of Supply, Affordability, Inclusiveness, Environmental Sustainability
- 3 profiles of prosumers
  - Households, Offices, Schools
- Several scenarios
  - Small vs Medium,
    Daily vs Annually,
    Default, Incremental, Decremental

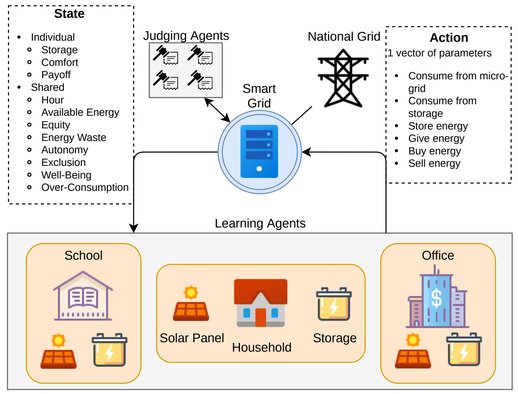


Fig. 2: Smart Grid simulator.

## Results

**Scalability** between Small and Medium sizes of grids.
Ability to adapt when **adding** and **removing** moral rules.

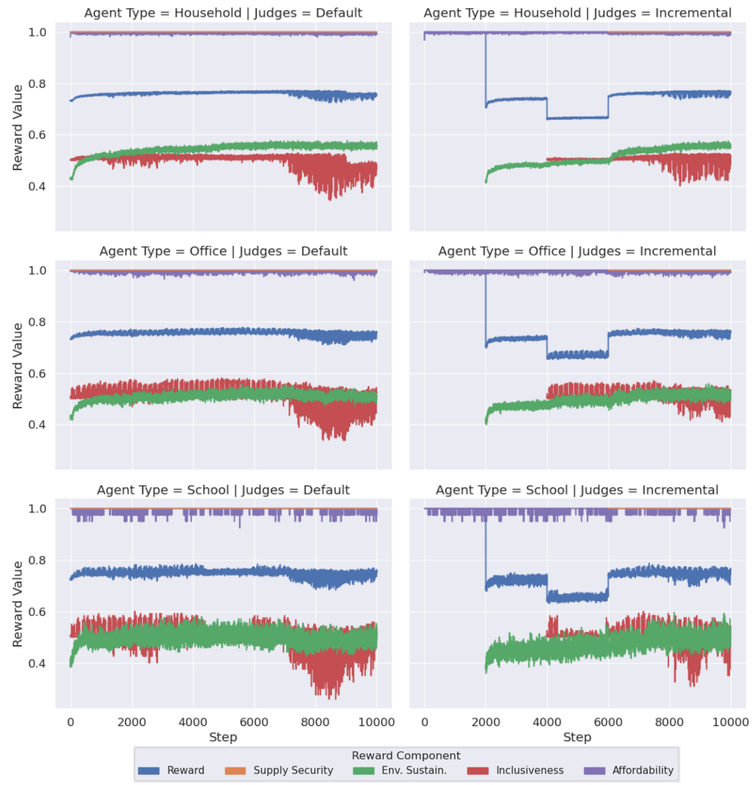


Fig. 3: Comparison of received rewards.

## Conclusion

Agents learn a behavior corresponding to moral rules ; able to adapt to changing rules.
**Complex use case**, in opposition to textbook ethical dilemmas.
Current limitations:

- Moral rules could be more complex.
- Symbolic-to-numeric transformation could use **argumentation** processes to solve conflicts between judges.

Auvergne Rhône-Alpes · Mines Saint-Etienne · LIMOS · LIRIS · Université Lyon 1 · Lyon Catholic University · CNRS · UBIANT

ethicsai@liris.cnrs.fr