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

R. Chaput[1], J. Duval, O. Boissier[2], M. Guillermin[3], S. Hassas[1]

AIES-21 Conference, May 19-21, 2021

[1]Univ. Lyon, Université Lyon 1, LIRIS, UMR5205, F-69622, LYON, France

[2]Mines Saint-Etienne, Univ Clermont Auvergne, CNRS, LIMOS UMR 6158, F-42023 Saint-Etienne, France

[3]UCLy, Sciences and Humanities Confluence Research Center (EA 1598), F-69288 Lyon, France

This work was funded by Région Auvergne-Rhône-Alpes (Pack Ambition Recherche) as part of the Ethics.AI project.

Presentation Outline

Introduction

Contributions

Use Case

Experiments and Results

Discussion

Introduction

Motivations

- Rising societal need for AI agents imbued with ethical considerations [Dig19; Moo06; Sch+20]
- Several implementations were already proposed [Yu+18]
- But it is not clear whether we should use Reasoning or Learning

Our objective

Propose a system of multiple artificial agents interacting in a shared environment, that learn an ethical behavior[1] by combining Learning and Reasoning in a Hybrid method.

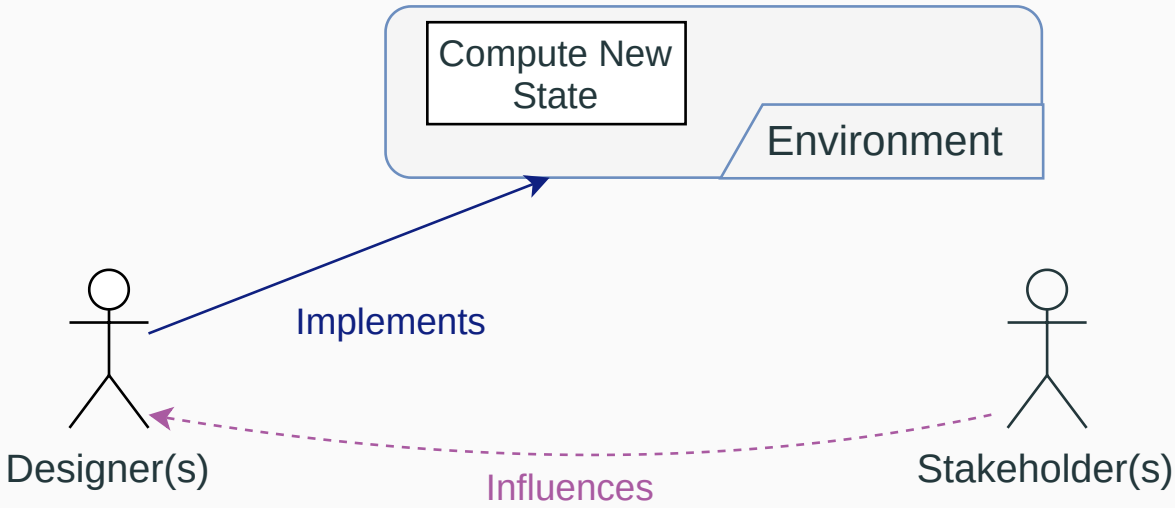
Agents should be able to adapt to changing rules.

- Multiple agents instead of a single one
- Focus on Ethics By Design and not only In Design [Dig19]

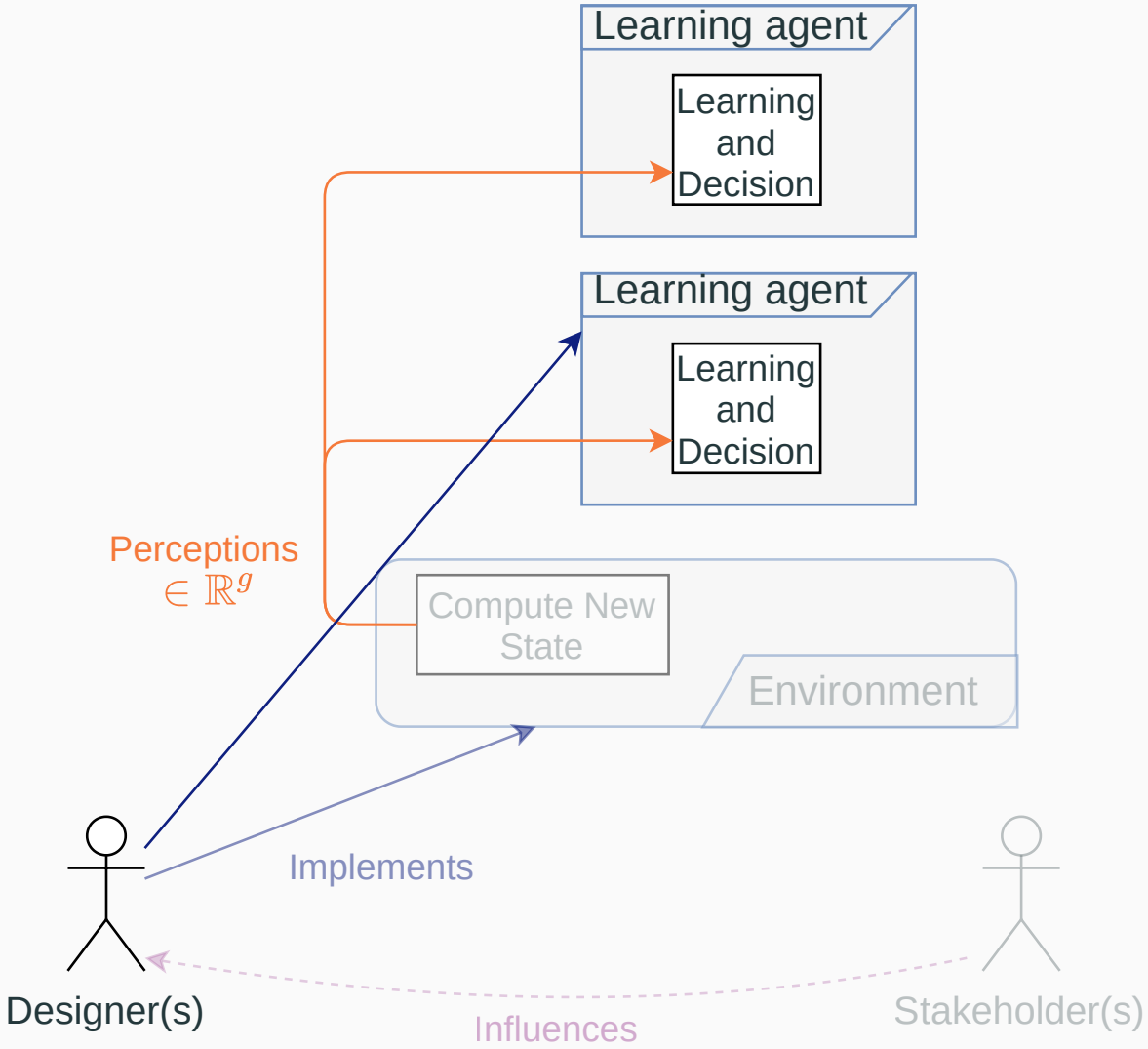
[1] Behavior that would be qualified as "ethical" when performed by humans.

Contributions

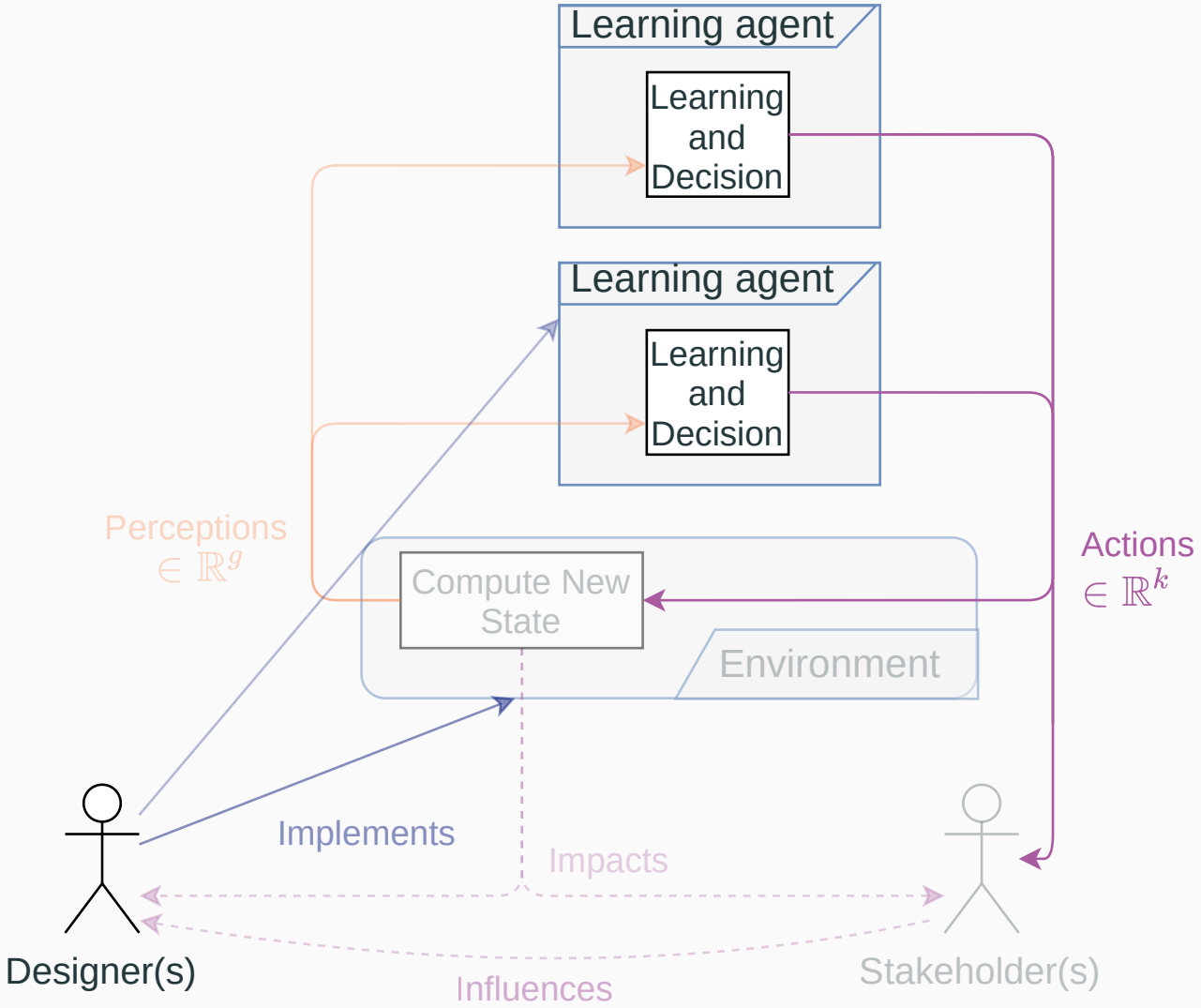
Proposed Model



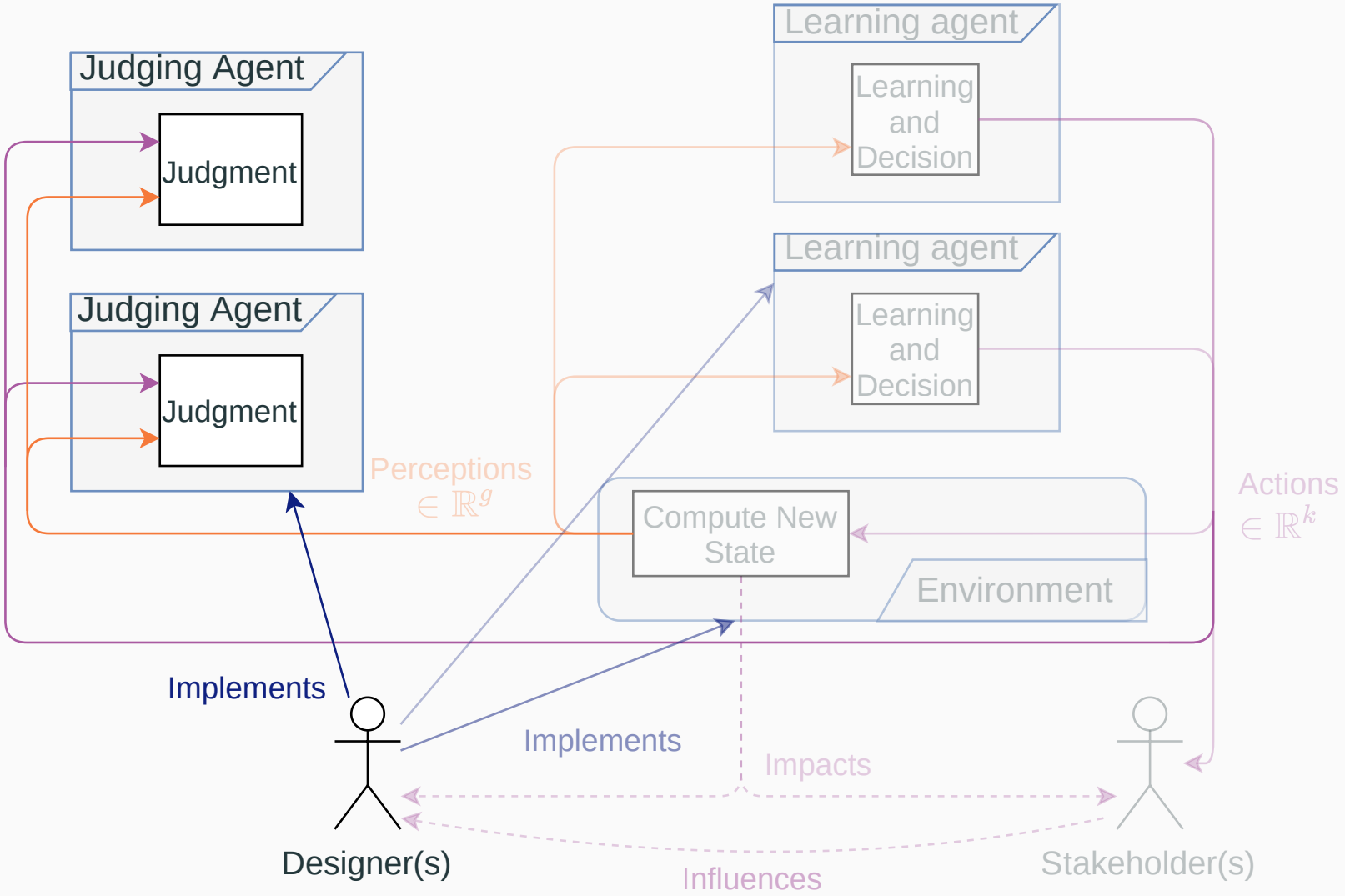
Proposed Model



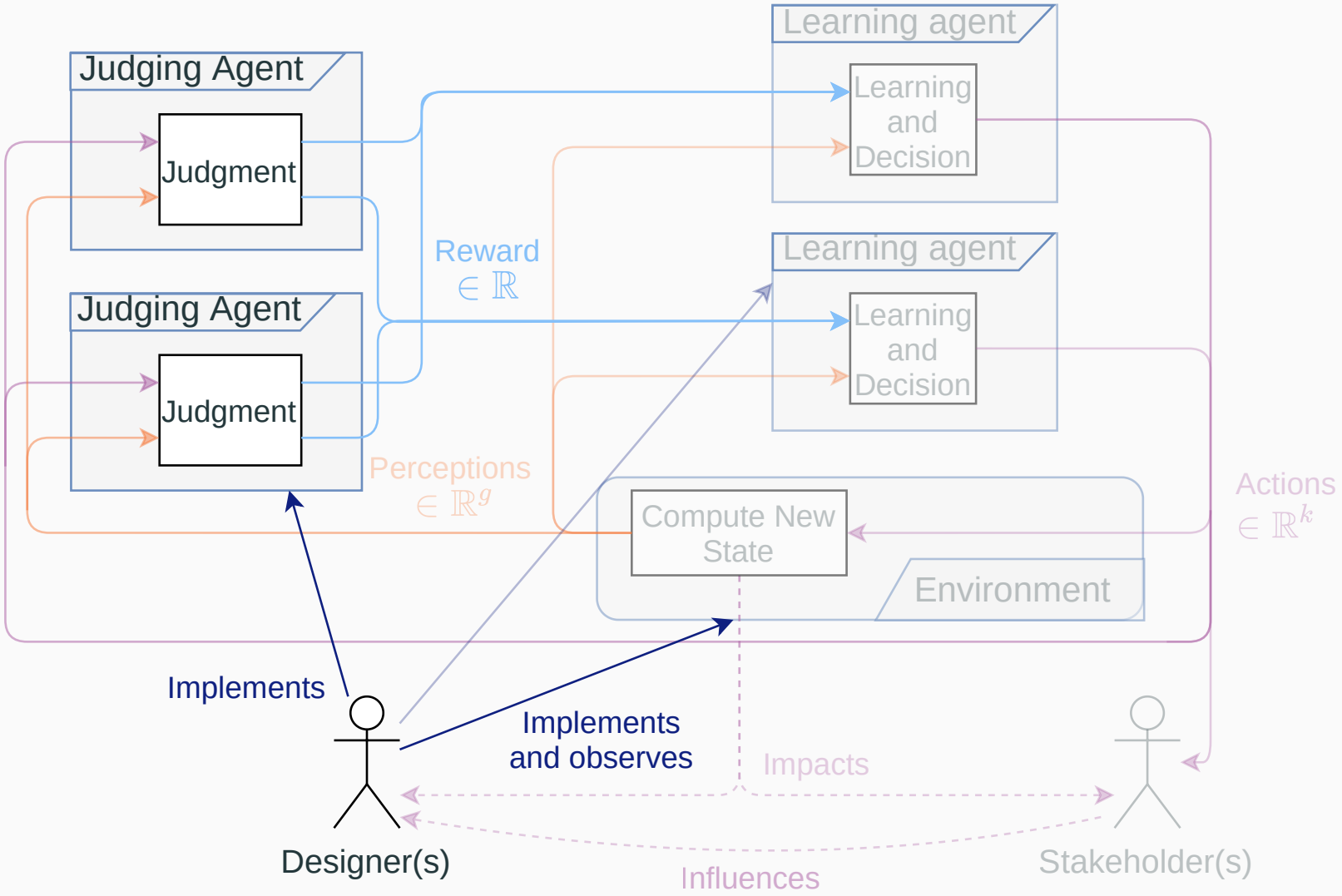
Proposed Model



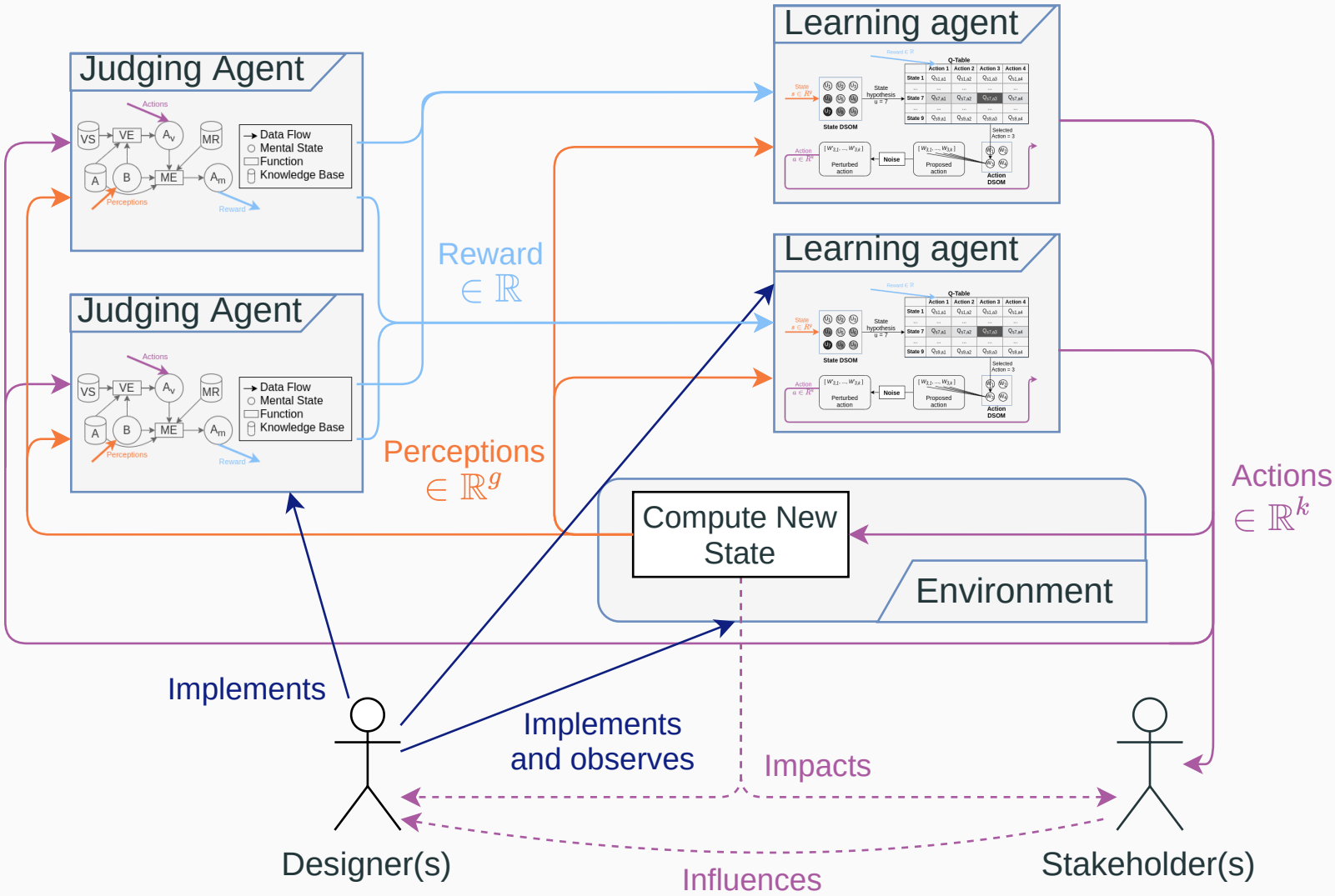
Proposed Model



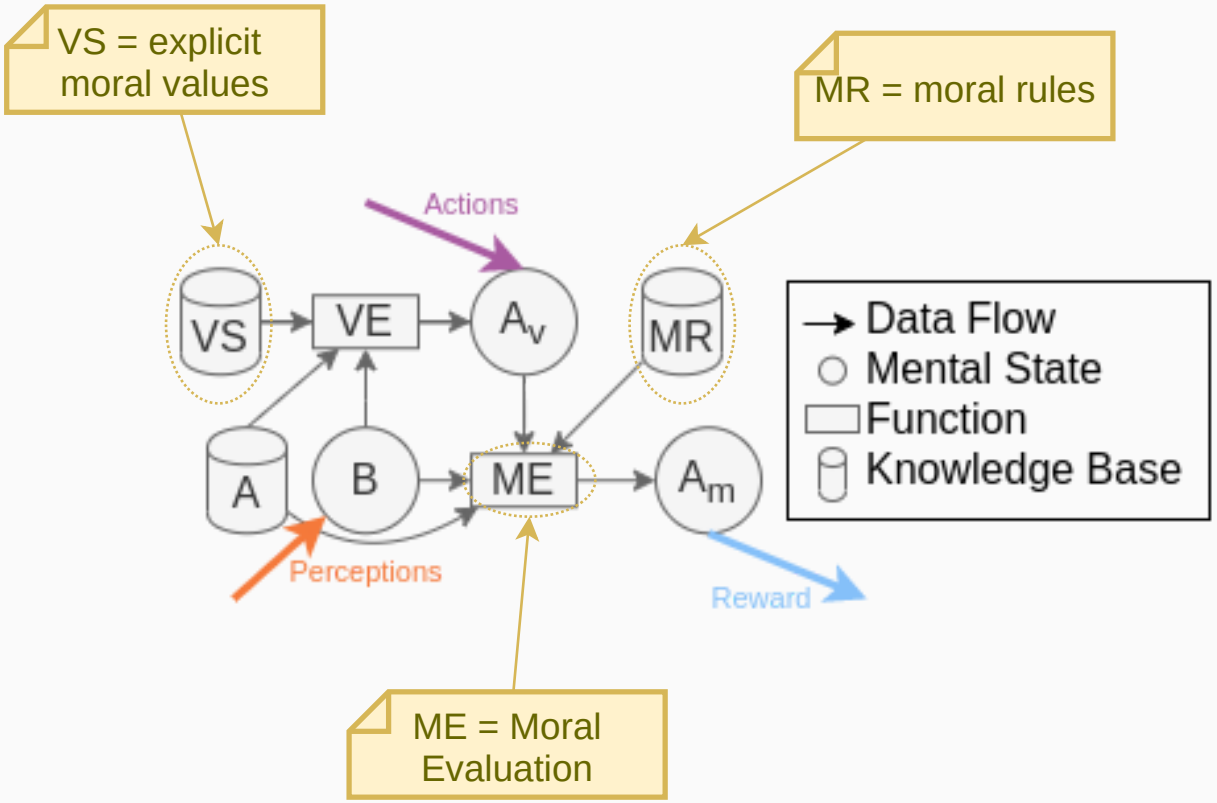
Proposed Model



Learning & Judging Agents

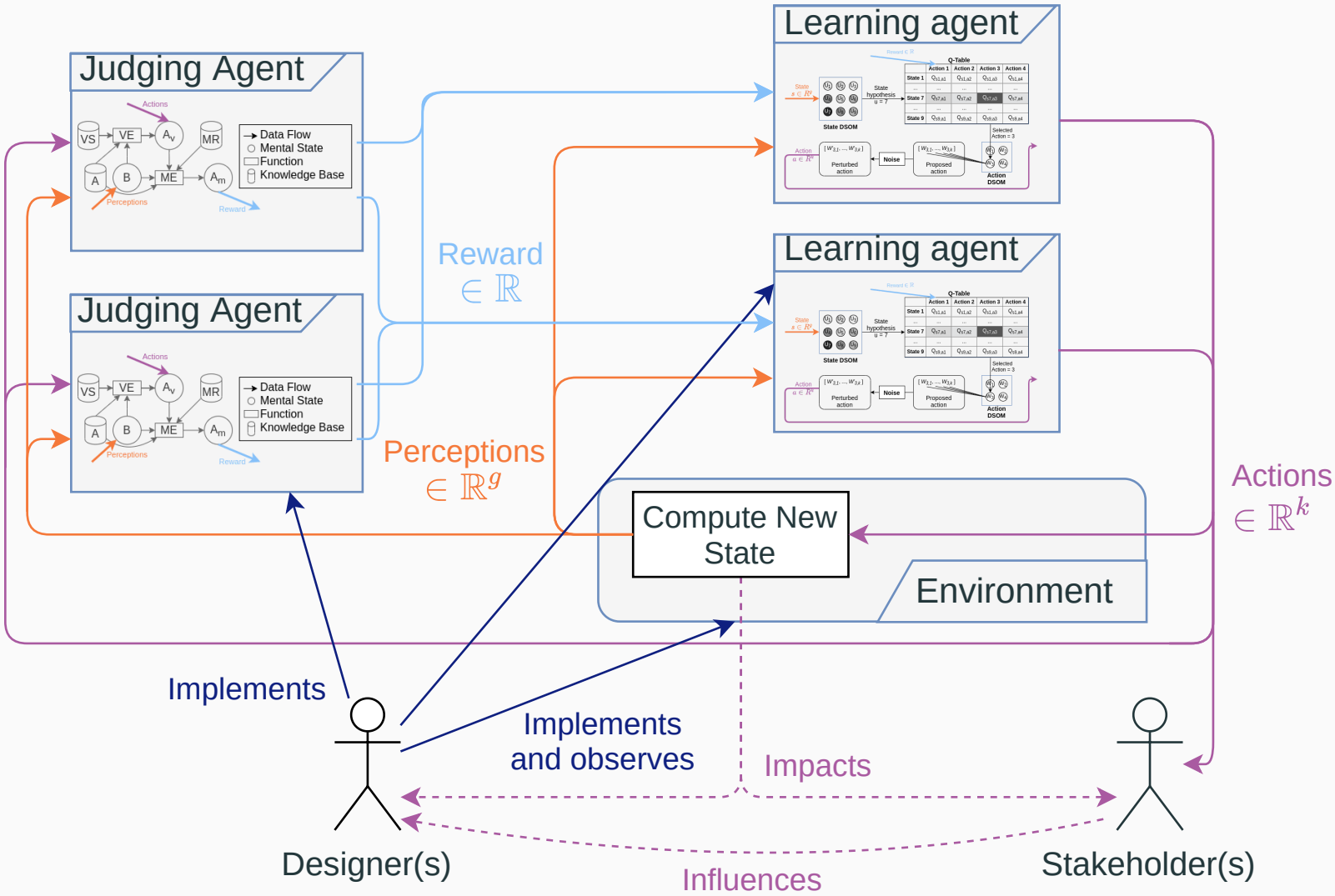


Learning & Judging Agents

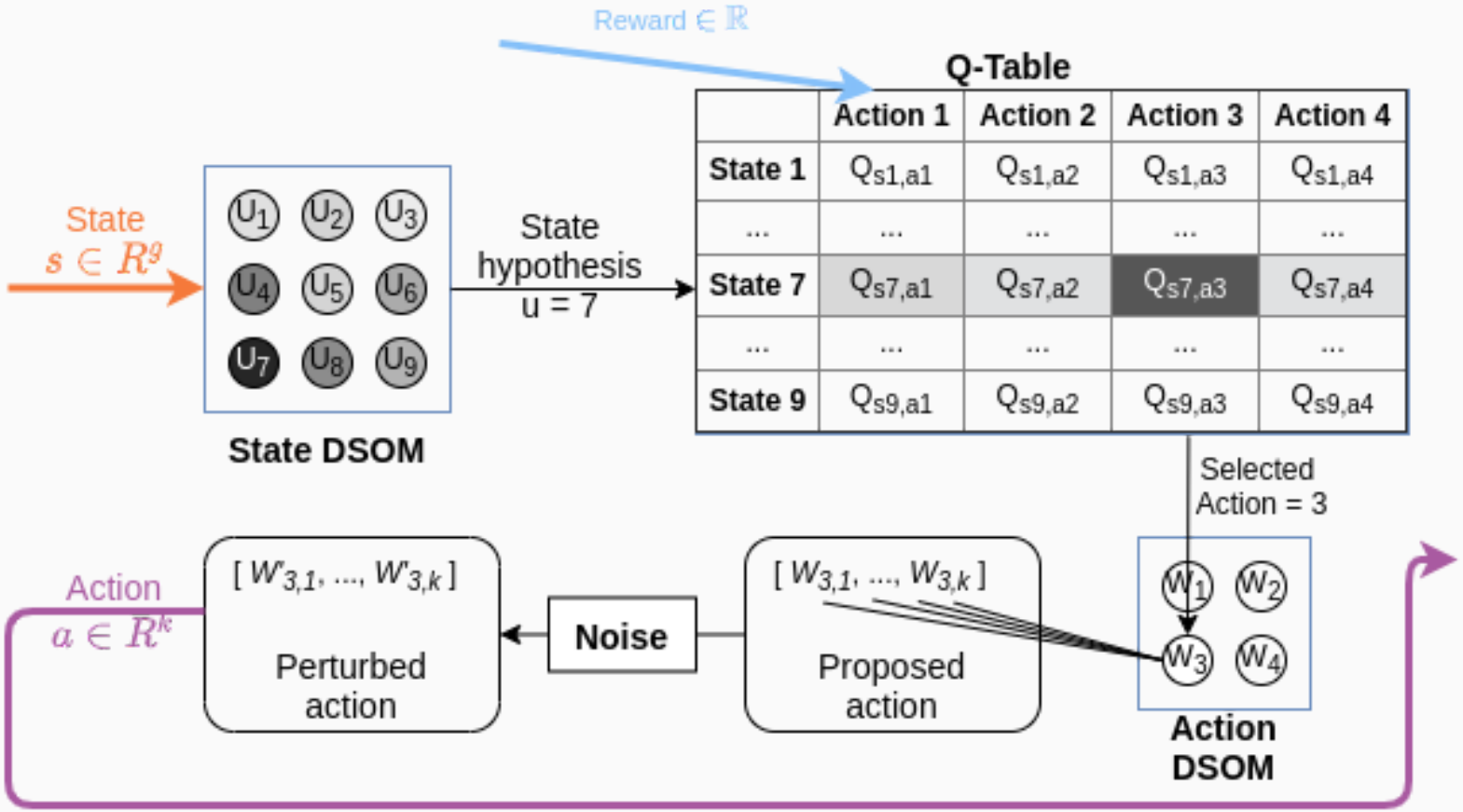


Judging agent: Ethicaa [CBB16]

Learning & Judging Agents

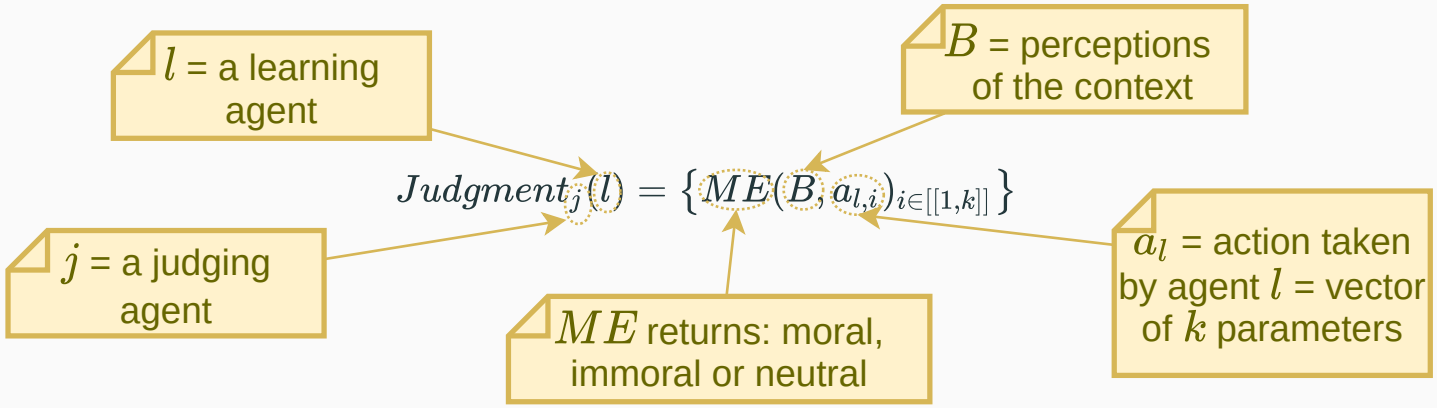


Learning & Judging Agents

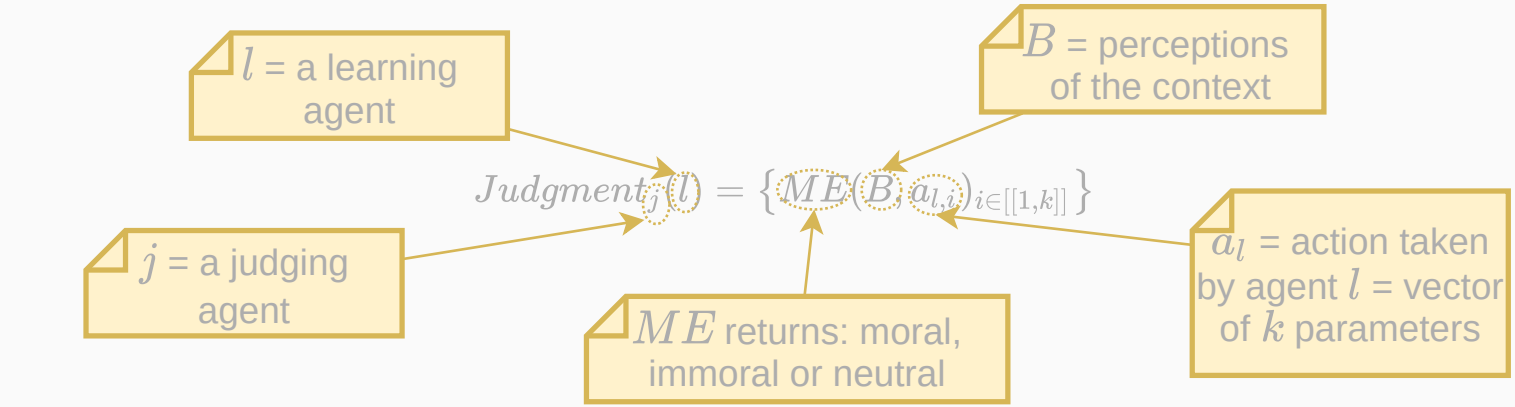


Learning agent: Q-DSOM [Cha+20]

Symbolic-to-numeric rewards



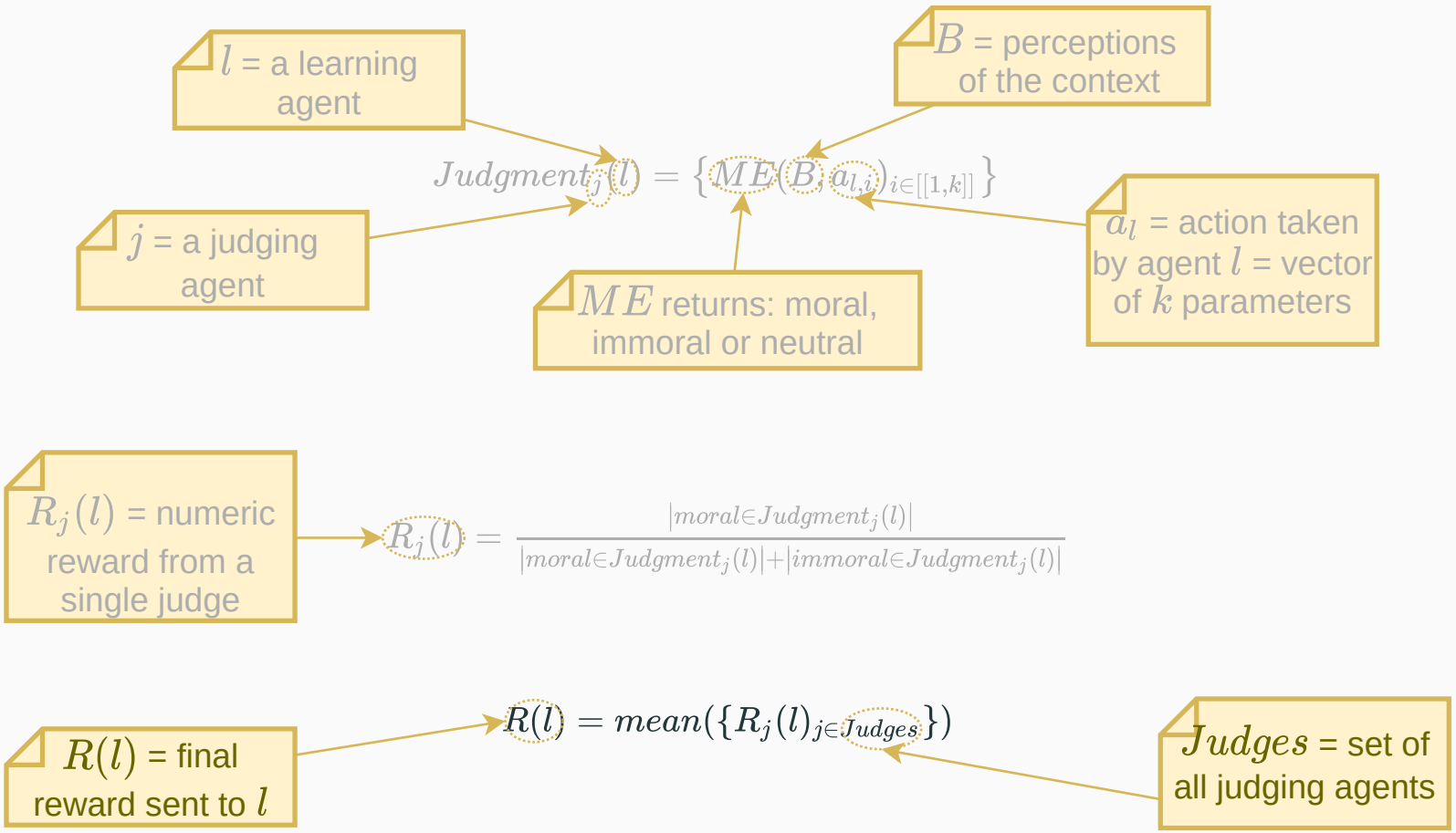
Symbolic-to-numeric rewards



$R_j(l)$ = numeric reward from a single judge

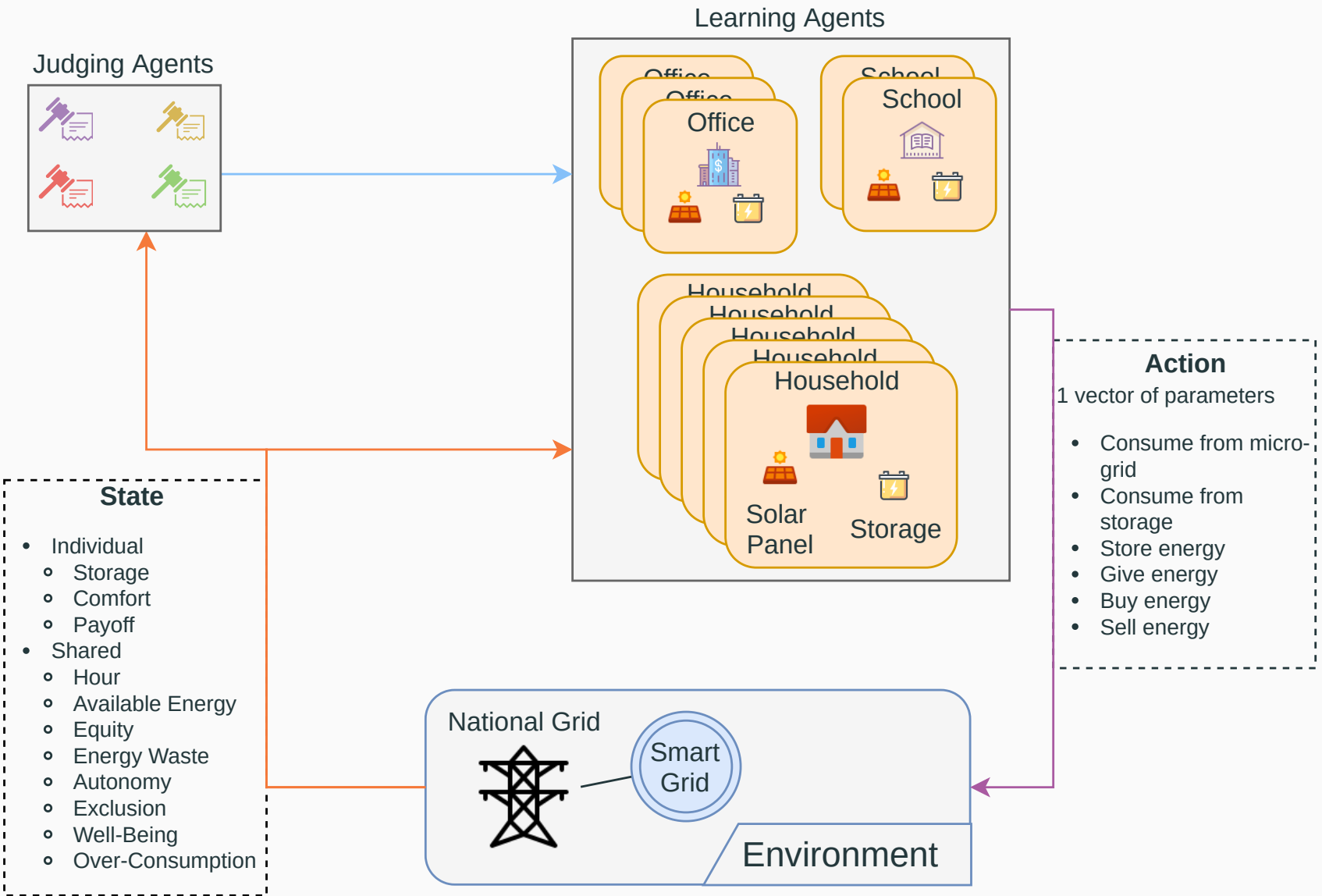
$$R_j(l) = \frac{|moral \in Judgment_j(l)|}{|moral \in Judgment_j(l)| + |immoral \in Judgment_j(l)|}$$

Symbolic-to-numeric rewards



Use Case

Smart Grids

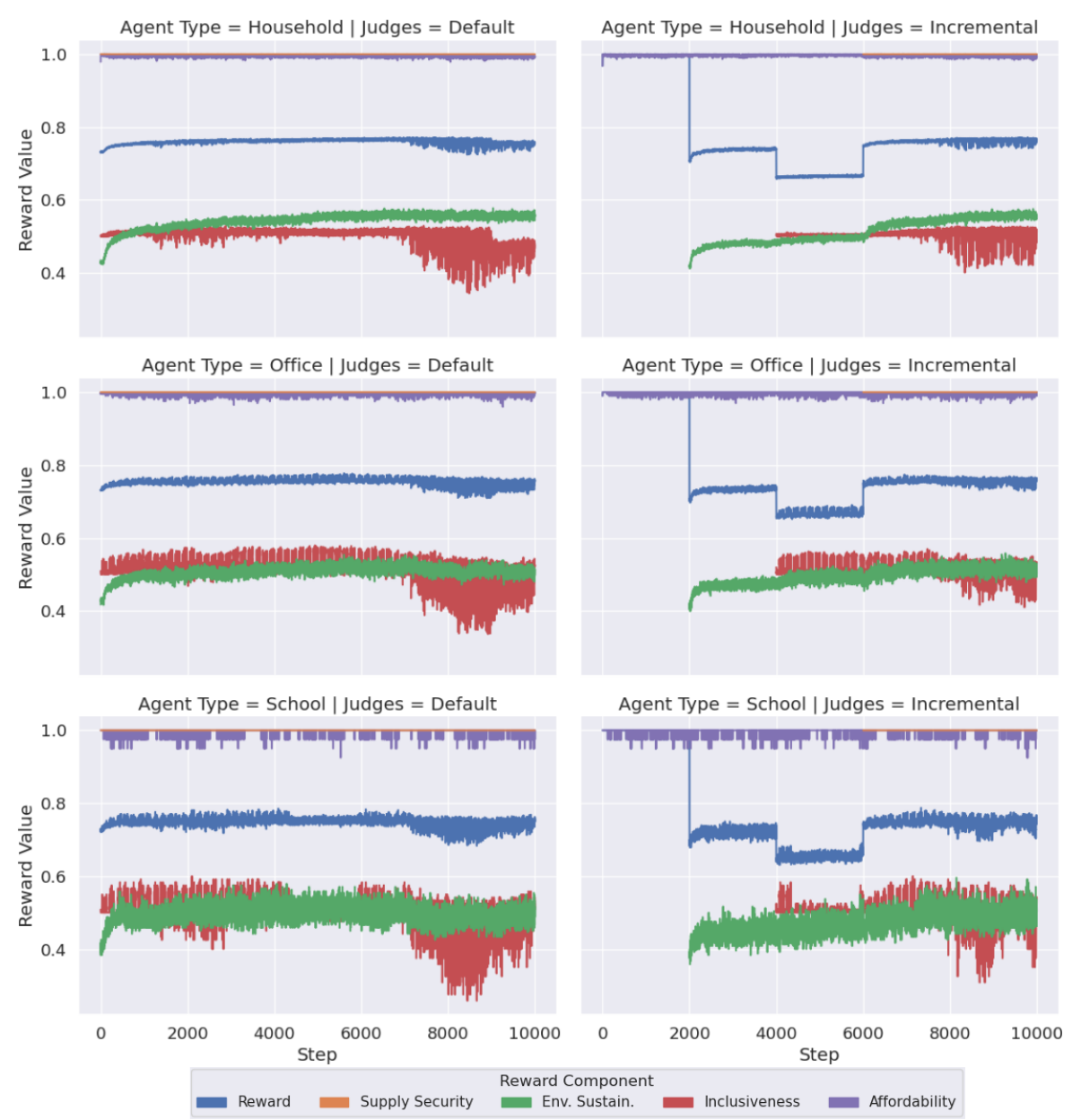

Learning Agents
Judging Agents
Office
School
Household
Solar Panel
Storage
State
Individual
Storage
Comfort
Payoff
Shared
Hour
Available Energy
Equity
Energy Waste
Autonomy
Exclusion
Well-Being
Over-Consumption
Action
1 vector of parameters
Consume from micro-grid
Consume from storage
Store energy
Give energy
Buy energy
Sell energy
National Grid
Smart Grid
Environment

Experiments and Results

Experiments

- 4 Moral Values (and associated rules) [Boi19; Wil+19; Mil+18]
  - Security of Supply: improve one's comfort
  - Affordability: do not pay too much
  - Inclusiveness: ensure equity of comforts
  - Environmental Sustainability: prevent exchanges with national grid
- 3 profiles of prosumers
  - Households
  - Offices
  - Schools
- Several scenarios
  - Small vs Medium
  - Daily vs Annually
  - Default (all judges) vs Incremental vs Decremental

Results


Agent Type = Household | Judges = Default
Agent Type = Household | Judges = Incremental
Agent Type = Office | Judges = Default
Agent Type = Office | Judges = Incremental
Agent Type = School | Judges = Default
Agent Type = School | Judges = Incremental
Reward Value
1.0
0.8
0.6
0.4
0
2000
4000
6000
8000
10000
Step
Reward Component
Reward
Supply Security
Env. Sustain.
Inclusiveness
Affordability

Discussion

Advantages

- Combine Reasoning (use expert knowledge) and Learning (generalize over unexpected situations) advantages
- Allows for a co-construction process with a human-in-the-loop schema
- Symbolic judgment allows for a better intelligibility of the expected behavior
- Using a variety of judges gives a richer feedback
- Learning agents have the ability to adapt to changing rules

Limitations and Perspectives

- We use domain-specific moral rules
  - Other works have a more generic approach [WL18]
  - May be possible to use generic rules if they do exist (?)
- No guarantee on the moral compliance
  - Other works use formal verification [Bre+19]
  - May be possible to apply formal verification to RL [FP18; Cor+20]
- Judgment may use extensive data from agents
  - Could be mitigated by using limited judgments or anonymized data
- The moral rules could be more complex
  - It was a necessary step to assess feasibility
- Symbolic-to-numeric transformation use a simple mechanism to solve conflicts between judges
  - We could could use an argumentation or negotiation process

Thank you for your attention

References

[Boi19] Anne Boijmans. “The Acceptability of Decentralized Energy Systems”. MA thesis. Delft University of Technology, July 2019.

[Bre+19] Paul Bremner et al. “On proactive, transparent, and verifiable ethical reasoning for robots”. In: *Proceedings of the IEEE* 107.3 (2019), pp. 541–561.

References II

[CBB16] Nicolas Cointe, Grégory Bonnet, and Olivier Boissier. “Ethical Judgment of Agents’ Behaviors in Multi-Agent Systems”. In: *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*. AAMAS ’16. Singapore, Singapore: International Foundation for Autonomous Agents and Multiagent Systems, May 2016, pp. 1106–1114.

[Cha+20] Rémy Chaput et al. “Apprentissage adaptatif de comportements éthiques”. In: *Architectures multi-agents pour la simulation de systèmes complexes - Vingt-huitième journées francophones sur les systèmes multi-agents, JFSMA 2020, Angers, France, June 29 - July 3, 2020*. Ed. by Nicolas Sabouret. Cépaduès, 2020.

References III

[Cor+20] Davide Corsi et al. “Formal Verification for Safe Deep Reinforcement Learning in Trajectory Generation”. In: *Fourth IEEE International Conference on Robotic Computing, IRC 2020, Taichung, Taiwan, November 9-11, 2020*. IEEE, 2020, pp. 352–359.

[Dig19] Virginia Dignum. *Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way*. Springer Nature, 2019.

References iv

[FP18] Nathan Fulton and André Platzer. “Safe Reinforcement Learning via Formal Methods: Toward Safe Control Through Proof and Learning”. In: *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*. Ed. by Sheila A. McIlraith and Kilian Q. Weinberger. AAAI Press, 2018, pp. 6485–6492.

[Mil+18] Christine Milchram et al. “Moral values as factors for social acceptance of smart grid technologies”. In: *Sustainability* 10.8 (2018), p. 2703.

References v

[Moo06] James H Moor. “The nature, importance, and difficulty of machine ethics”. In: *IEEE intelligent systems* 21.4 (2006), pp. 18–21.

[Sch+20] Daniel Schiff et al. “What’s Next for AI Ethics, Policy, and Governance? A Global Overview”. In: *AIES ’20: AAAI/ACM Conference on AI, Ethics, and Society, New York, NY, USA, February 7-8, 2020*. Ed. by Annette N. Markham et al. ACM, 2020, pp. 153–158.

[Wil+19] TE de Wildt et al. “Conflicting values in the smart electricity grid a comprehensive overview”. In: *Renewable and Sustainable Energy Reviews* 111 (2019), pp. 184–196.

References vi

[WL18] Yueh-Hua Wu and Shou-De Lin. “A Low-Cost Ethics Shaping Approach for Designing Reinforcement Learning Agents”. In: *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*. Ed. by Sheila A. McIlraith and Kilian Q. Weinberger. AAAI Press, 2018, pp. 1687–1694.

[Yu+18] Han Yu et al. “Building Ethics into Artificial Intelligence”. In: *Proceedings of the 27th International Joint Conference on Artificial Intelligence*. IJCAI’18. Stockholm, Sweden: AAAI Press, July 2018, pp. 5527–5533.