Multi-Agent Systems for Bullying Intervention

Extended Abstract

Luis Zhinin-Vera
University of Castilla-La Mancha
Albacete, Spain
luis.zhinin@uclm.es

José J González-García
University of Castilla-La Mancha
Albacete, Spain
josejesus.gonzalez@uclm.es

Víctor López-Jaquero
University of Castilla-La Mancha
Albacete, Spain
victormanuel.lopez@uclm.es

Elena Navarro
University of Castilla-La Mancha
Albacete, Spain
elena.navarro@uclm.es

Pascual González
University of Castilla-La Mancha
Albacete, Spain
pascual.gonzalez@uclm.es

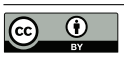
ABSTRACT

Bullying is a pervasive problem in educational settings, leading to serious emotional and psychological harm for those involved. It is a complex social phenomenon involving multiple roles, including bullies, victims, and observers, each contributing to the dynamics of bullying scenarios. In this paper, we propose a novel Multi-Agent system (MAS) framework aimed at detecting and intervening in bullying cases through the integration of Theory of Mind (ToM), Reinforcement Learning (RL), and Continual Learning (CL). Our approach leverages ToM to allow agents to infer the mental states of others, enabling context-aware decision-making for effective intervention strategies. RL is used to allow the observer agent to learn from past interactions, improving its ability to recognize bullying behaviors and refine its responses. CL ensures the system can adapt to new behaviors and evolving environments, maintaining its effectiveness over time. We present abstraction mechanisms based on Theory-Theory and Simulation Theory, which allow the system to reason about complex social interactions either through predefined rules or simulations. This paper outlines the theoretical framework and design of the proposed algorithm, offering a responsive, flexible, adaptive, and capable solution for bullying prevention and intervention in educational contexts, where socially intelligent systems can play a key role in creating safer environments.

KEYWORDS

Multi-Agent Reinforcement Learning; Theory of Mind; Hybrid Intelligence; Human Digital Twin

ACM Reference Format:
Luis Zhinin-Vera, José J González-García , Víctor López-Jaquero, Elena Navarro, and Pascual González. 2025. Multi-Agent Systems for Bullying Intervention: Extended Abstract. In *Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Detroit, Michigan, USA, May 19 – 23, 2025*, IFAAMAS, 3 pages.



Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Y. Vorobeychik, S. Das, A. Nowé (eds.), May 19 – 23, 2025, Detroit, Michigan, USA. © 2025 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

1 INTRODUCTION

Advancing Artificial Intelligence (AI) systems to navigate complex human social dynamics requires overcoming the significant challenge of enabling them to understand, predict, and respond to human intentions and emotions. Theory of Mind (ToM), a crucial cognitive skill that enables humans to attribute mental states to themselves and others, is at the heart of this endeavor [2]. Integrating ToM into AI systems promises to elevate their social intelligence, facilitating more natural and effective interactions. This work explores innovative approaches to enhancing ToM capabilities in AI through the integration of multi-agent reinforcement learning (MARL) and advanced abstraction techniques. These approaches delve into the concept of Mindful Human Digital Twin (MHDT), to go beyond the usual Digital Twin (DT's) framework by incorporating ToM [14, 15]. DT's, initially conceptualized as virtual replicas of physical entities, can be further enhanced through the Multi-Agent Systems (MAS) paradigm to support complex tasks and decision-making processes [11, 12]. This work introduces several novel enhancements to MHDT. We propose advanced abstraction techniques to efficiently model ToM by simplifying computational demands while preserving social interaction richness through abstractions like norms, roles, and values. Additionally, we incorporate the concept of *hybrid intelligence*, fostering effective collaboration between humans and AI agents [1, 5]. Using Reinforcement Learning (RL) [13] and Continual Learning (CL) [10], we enable agents to continuously learn and adapt, creating resilient AI systems suited for dynamic and unpredictable contexts such as bullying scenarios.

2 COMPUTATIONAL APPROACH FOR BULLYING INTERVENTION

2.1 Motivating Example

A school bullying scenario is shown to illustrate the proposed approach, involving agents such as the bully (A_B), victim (A_V), observer (A_O), and teacher (A_T). The A_O monitors interactions and leverages prior knowledge, such as historical behavior and real-time observations, to infer the mental states and intentions of the bully and the victim. A_O dynamically weighs its beliefs about the intentions of A_B, the state of A_V, and the potential outcomes of different interventions. Based on this reasoning, A_O decides whether and how to act, choosing strategies ranging from subtle de-escalation to involving an authority figure A_T in severe cases.

2.2 Implementation of Theory of Mind Algorithm in Bullying Intervention

We propose an approach using ToM, RL, and CL, enabling agents to infer mental states and adapt their strategies to evolving interactions. Focusing on the A_O, the proposal demonstrates how ToM enhances context-aware and ethically aligned interventions, ensuring adaptive responses to the dynamic nature of bullying.

2.2.1 Main Algorithm. The main algorithm forms the core of the proposed system, leveraging RL to optimize decision-making for the A_O. Its goal is to learn an action-value function $Q(s, a)$ that maximizes cumulative rewards over time, addressing the complex dynamics of bullying scenarios. Inputs include the state space (e.g., agent's positions and interactions), action space (e.g., intervene, observe, or escalate), reward function (e.g., positive for successful interventions), and transition function (e.g., changes in agent behavior after an action), all of which capture the environment's configurations, possible actions, outcomes, and dynamics. The algorithm incorporates social context by embedding prior knowledge, such as historical interactions, into the state space, enabling A_O to make decisions aligned with ongoing social dynamics.

The agent balances exploitation and exploration to select actions, either maximizing $Q(s, a)$ or invoking the *Theory of Mind* function for reasoning about other agents' mental states. Actions and rewards are stored in an experience replay memory, and updates to $Q(s, a)$ occur through stochastic gradient descent with regularization (e.g., using the Fisher Information Matrix) to prevent catastrophic forgetting. Parameters like learning rate, discount factor, and dropout rate govern the learning process, ensuring adaptability (via CL) and ethical alignment in interventions.

2.2.2 Theory of Mind Function. The ToM-enhanced module refines the A_O's decision-making by enabling reasoning about other agents' mental states, such as beliefs, intentions, and desires. Building on concepts like Theory-Theory (TT) and Simulation Theory (ST) [8, 9], it incorporates predefined rules and learned experiences to predict outcomes and guide interventions. This integration allows dynamic reasoning about social interactions, aligning decisions with social norms and improving the agent's effectiveness in addressing bullying scenarios [6].

The *perception and inference of the mental state* is the initial concept, where the ToM algorithm begins by extracting observable features from the environment's current state, such as actions, expressions, and spatial relationships. These features are used to infer the mental states of other agents, providing a foundation for understanding their motivations and emotions. *Abstraction mechanisms* simplify the complex task of mental state inference in the ToM algorithm by translating observable features (e.g., actions, gestures, facial expressions, or spatial relationships) into higher-level concepts like intentions and emotions. For example, aggressive gestures by A_B can infer intimidation, or defensive postures by A_V can indicate fear. Using predefined rules grounded in TT, the agent derives mental states such as fear, empathy, or intent based on patterns in social interactions. These rules are supported by epistemic principles (e.g., knowledge implies belief) that formalize knowledge and belief relationships, enabling the agent to reason effectively about agents' behaviors. ST further refines these inferences by allowing the agent to predict responses to potential actions based on its own experiences. The resulting mental state representation integrates both inferred and simulated data, providing a rich foundation for adaptive, context-aware decision-making in bullying scenarios.

After inferring mental states, actions are filtered through two critical constraints: norms and roles. *Norm constraints* ensure that actions align with predefined behavioral expectations, such as intervening to stop aggression in bullying scenarios [4, 7]. These norms are represented as rules (e.g., if aggression is detected, intervention is required) and help narrow the available actions to a socially acceptable set. *Role constraints* further refine the action set based on the agent's responsibilities within a social context. Roles dictate appropriate behaviors and responsibilities, guiding the agent's decisions to match its identity, such as acting as a mediator to de-escalate conflicts [3]. For example, an A_O in a mediator role focuses on dialogue and resolution rather than on punitive actions. This two-stage filtering process produces a final set of actions that are both norm-compliant and role-appropriate, enabling effective and context-aware decision-making.

The algorithm *simulates the outcomes of actions* in the filtered set to estimate their *utility*, predicting next states and calculating rewards and future values. This simulation considers the inferred mental states and social dynamics to anticipate how agents like the A_B and A_V might react to specific actions. The utility is then adjusted to align with intrinsic values, such as empathy and safety, ensuring that decisions reflect ethical principles and moral priorities. Actions that align with these values receive a bonus, favoring interventions that address distress or de-escalate conflicts, while minimizing potential harm. The action with the highest adjusted utility is selected as the optimal decision, balancing social norms, role suitability, inferred mental states, and value alignment. Additionally, this process allows the agent to adapt to dynamic environments and evolving social contexts. The integration of CL, RL, and ToM equips the A_O to adapt, reason about social dynamics, and choose interventions that maximize long-term rewards while adhering to ethical considerations, effectively closing the decision-making loop and ensuring robust responses to bullying scenarios.

3 CONCLUSION

This work presents a multi-agent system (MAS) integrating Theory of Mind (ToM), Reinforcement Learning (RL), and Continual Learning (CL) to address bullying scenarios. By inferring mental states, beliefs, and intentions, agents make real-time, context-aware interventions, adapting to new behaviors and evolving dynamics. This framework emphasizes adaptive, and effective solutions for understanding and mitigating bullying. Applications include early detection of aggression and supporting empathy in schools, showcasing its potential to create safer educational environments.

ACKNOWLEDGMENTS

This work is part of the R+D+i projects PID2019-108915RB-I00 and PID2022-140907OB-I00 as well as by the grant PRE2020-094056 funded by MCIU/AEI/10.13039/501100011033 and ERDF, EU. It has also been partially supported by the University of Castilla-La Mancha (2022-GRIN-34436).

REFERENCES

[1] Zeynep Akata, Dan Balliet, Maarten de Rijke, Frank Dignum, Virginia Dignum, Guszti Eiben, Antske Fokkens, Davide Grossi, Koen Hindriks, Holger Hoos, Hayley Hung, Catholijn Jonker, Christof Monz, Mark Neerincx, Frans Oliehoek, Henry Prakken, Stefan Schlobach, Linda van der Gaag, Frank van Harmelen, Herke van Hoof, Birna van Riemsdijk, Aimee van Wynsberghe, Rineke Verbrugge, Bart Verheij, Piek Vossen, and Max Welling. 2020. A Research Agenda for Hybrid Intelligence: Augmenting Human Intellect With Collaborative, Adaptive, Responsible, and Explainable Artificial Intelligence. *Computer* 53, 8 (2020), 18–28. https://doi.org/10.1109/MC.2020.2996587

[2] Simon Baron-Cohen. 1995. *Mindblindness: An Essay on Autism and Theory of Mind.* The MIT Press. https://doi.org/10.7551/mitpress/4635.001.0001

[3] G.M. Breakwell, H.C. Foot, and R. Gilmour. 1982. *Social Psychology: A Practical Manual.* Macmillan. https://doi.org/10.1007/978-1-349-16794-4

[4] Martina Cabra. 2020. *Norms.* Springer International Publishing, Cham, 1–9.

[5] Dominik Dellermann, Philipp Alexander Ebel, Matthias Söllner, and Jan Marco Leimeister. 2019. Hybrid Intelligence. *Business & Information Systems Engineering* 61 (2019), 637 – 643. https://doi.org/10.1007/s12599-019-00595-2

[6] Emre Erdogan, Frank Dignum, and Rineke Verbrugge. 2024. Effective Maintenance of Computational Theory of Mind for Human-AI Collaboration. In *HHAI 2024: Hybrid Human AI Systems for the Social Good.* IOS Press, 114–123. https://doi.org/10.3233/FAIA240188

[7] Emre Erdogan, Rineke Verbrugge, and Pınar Yolum. 2024. Computational Theory of Mind with Abstractions for Effective Human-Agent Collaboration. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems* (Auckland, New Zealand) *(AAMAS '24)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 2249–2251.

[8] Alison Gopnik and Henry Wellman. 1992. Why the Child's Theory of Mind Really Is a Theory. *Mind & Language* 7, 1-2 (1992), 145 – 171. https://doi.org/10.1111/j.1468-0017.1992.tb00202.x

[9] Paul L. Harris. 1992. From Simulation to Folk Psychology: The Case for Development. *Mind & Language* 7, 1-2 (1992), 120–144. https://doi.org/10.1111/j.1468-0017.1992.tb00201.x

[10] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences* 114, 13 (March 2017), 3521–3526. https://doi.org/10.1073/pnas.1611835114

[11] Elena Pretel, Alejandro Moya, Elena Navarro, Víctor López-Jaquero, and Pascual González. 2024. Analysing the synergies between Multi-agent Systems and Digital Twins: A systematic literature review. *Information and Software Technology* 174 (2024), 107503. https://doi.org/10.1016/j.infsof.2024.107503

[12] Elena Pretel, Luis Zhinin-Vera, Elena Navarro, Víctor López-Jaquero, and Pascual González. 2025. MAS4DT: A novel proposal for developing Digital Twins following a Multi-Agent System approach. *Journal of Systems and Software* (2025), 112344. https://doi.org/10.1016/j.jss.2025.112344

[13] Richard S. Sutton and Andrew G. Barto. 2018. *Reinforcement Learning: An Introduction.* A Bradford Book, Cambridge, MA, USA.

[14] Luis Zhinin-Vera, Víctor López-Jaquero, Elena Navarro, and Pascual González. 2023. A Computational Model for Agents in a Social Context: An Approach Based on Theory of Mind. In *Proceedings of the 15th International Conference on Ubiquitous Computing & Ambient Intelligence (UCAmI 2023).* Springer Nature Switzerland, Cham, 3–14. https://doi.org/10.1007/978-3-031-48306-6_1

[15] Luis Zhinin-Vera, Elena Pretel, Víctor López-Jaquero, Elena Navarro, and Pascual González. 2024. Mindful Human Digital Twins: Integrating Theory of Mind with Multi-Agent Reinforcement Learning. *Applied Soft Computing* (2024). Under second review.