Learning Preferences and Resolving Conflicts in Multi-User Personalisation in Human-Robot Interaction

Doctoral Consortium

Aniol Civit
Institut de Robòtica i Informàtica Industrial, CSIC-UPC
Barcelona, Spain
acivit@iri.upc.edu

ABSTRACT

Personalisation is a fundamental aspect of Human-Robot Interaction, as users tend to accept and engage more with robots that adapt their behaviour to individual preferences. Prior work has mainly focused on adapting behaviour to a single user. However, in many real-world scenarios, robots must account for the preferences of multiple users. This is particularly evident in assistive contexts, where a robot should consider the preferences of the caregiver, who possesses domain expertise regarding task execution, as well as those of care recipients, who directly experience the interaction. In such settings, these preferences may conflict. The overarching goal of this work is to utilise Gradual Argumentation to resolve multi-user preference conflicts in Human-Robot Interaction. Specifically, we aim to make this framework adaptable and contestable, allowing users to influence the robot's decisions from feedback, and making it adaptable over long-term interactions, where preferences are dynamic and may evolve over time.

KEYWORDS

Human-Robot Interaction; Multi-User Personalisation; Assistive Robots; Gradual Argumentation

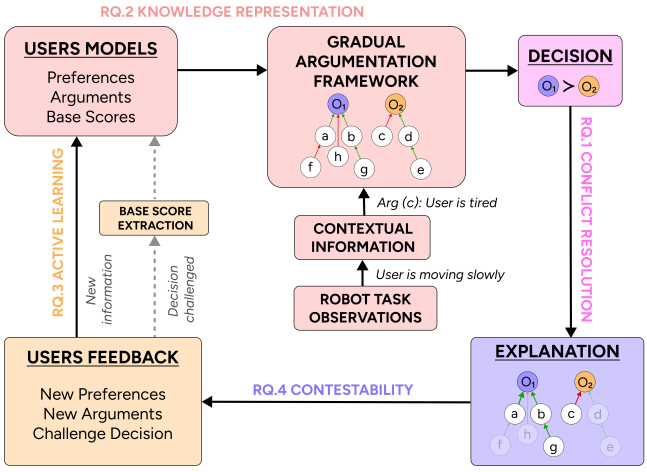

Figure 1: An overview of our robotic argumentation-based framework that resolves preference conflicts from modelled information of the users, explains its decisions, and learns from user feedback.

1 INTRODUCTION

An essential feature of assistive, social, and collaborative robots is their ability to autonomously adjust their behaviour according to the preferences, needs, and conditions of the humans they are interacting with [22]. Several studies in the field of Human-Robot Interaction (HRI) have focused on personalising these interactions in real-world settings, proving that it can improve engagement and foster trust and rapport [4, 5, 10, 35, 38].

Despite its promising potential, we identified 4 main research gaps in the literature, which we aim to address in our work:

1. Limitation to single-user contexts: Most research in HRI has focused on personalising the robot behaviour to a single user [19], while neglecting scenarios where multiple stakeholders are involved in the robot's decision-making. For instance, in the assistive domain, robots must account not only for patients' preferences, but also for those of multiple caregivers and therapists [36].

2. Difficulty in adapting to changing preferences: Traditional data-driven methods for personalisation in HRI often need large amounts of data to learn an optimal behaviour. Considering that preferences are not static and might change over time, these methods require the model to be retrained, for which more data is needed [11, 24].

3. Misalignment Between Qualitative Human Preferences and Quantitative Models: Standard decision-making frameworks in HRI rely on optimising single, pre-defined metrics (e.g., cost, engagement) to make decisions. This often forces complex, multi-dimensional human values, such as balancing a user's engagement with their success rate in a learning task, to be merged into a single scalar value [29]. While multi-agent and game-theoretic approaches offer structured conflict resolution [8, 23, 33], they require complete, quantitative utility functions for all parties. Human preferences, however, are often subjective and fundamentally qualitative. This mismatch strips away the nuanced reasoning behind preferences and prevents systems from providing intuitive, human-aligned explanations for their choices.

4. Opacity and lack of transparency: The 'black-box' nature of many data-driven models makes it difficult to explain why a particular decision was made [21]. This opacity hinders a system's *influencibility*: if users cannot understand a decision, they cannot provide feedback to correct or improve it. This creates a barrier to creating adaptive and transparent human-robot partnerships.

Our work is based on Argumentation Frameworks (AFs) [18], which are suitable to resolve preference conflicts. In fact, they offer a structured method to determine a reasonable decision by evaluating the arguments provided [2]. In particular, we employ Gradual Argumentation Frameworks (GAFs) [1], in which each argument

is assigned a numerical weight, referred to as a *base score*. By assigning the base scores, we can align the decision-making towards a more adaptive system according to users' preferences.

To fill the mentioned gaps, we aim to address the following research questions:

RQ.1 *Can we develop a framework for robots to provide personalised interactions according to multiple users' preferences, handling and solving any possible conflicts that may arise?*

RQ.2 *How can we model the users' preferences, reasons, and contextual information to fit in such a framework?*

RQ.3 *How can we elicit and understand the user's knowledge, and which methods are suitable to actively update their models?*

RQ.4 *Which explanations must the robot provide to allow users to effectively influence its decision-making?*

2 METHODOLOGY

Our framework, illustrated in Fig. 1, is structured to address the four research questions outlined in the introduction. Here, we detail our contributions to date and outline our planned future work.

RQ.1 Conflict Resolution: In assistive scenarios involving multiple stakeholders (e.g., patients, caregivers), their preferences can conflict. A caregiver may advocate for more exercise, while a patient may resist. Robots in these settings require tools to resolve such conflicts fairly, as their decisions impact all involved. Furthermore, user preferences and environmental context are dynamic, requiring a robot's decision-making to be effectively adaptable over time without costly retraining. Research in Argumentation Frameworks has mainly focused on resolving conflicts [3, 20], but very few works in GAFs have done it [31]. Specifically, their application in HRI for dynamic, multi-user conflict resolution remains underexplored, since standard argumentation models are not explicitly designed to incorporate real-time and sequential contextual observations. To that extent, we proposed a framework that considers multiple users' arguments, their reasons, and robot observations of the environment as arguments, and generates a GAF accordingly. This framework allows for inserting new arguments or modifying the existing ones without the need for retraining. The decision is based on the final strength of the arguments. We validated the framework theoretically and through a use case in which an assistive robot performs frailty assessments in older adults [14].

RQ.2 Knowledge Representation: To personalise decisions, a robot must quantify the importance (base score) a user assigns to different arguments. Standard GAFs provide no mechanism to derive these personalised scores from individual user data. In most GAFs applications, base scores are set heuristically or derived from aggregate data, such as votes in debates [17, 32] or product reviews for recommendation systems [16]. These methods do not capture individual user models. We propose *Base Score Extraction Functions* (BSEFs) to bridge this gap [15]. BSEFs provide a formal method to translate a single user's stated ordinal preferences over arguments into the quantitative base scores required by the argumentation framework, enabling true personalisation of the robot's reasoning model.

3 FUTURE DIRECTIONS

RQ.3 Active Learning: Users prefer to interact with robots using natural language, especially older adults [9]. Large Language Models have endowed robots with the capability to understand the users, the context, and to provide natural interactions accordingly [25]. That technology is suitable for our framework, since it allows us to proactively elicit users' preferences and their reasons, and to understand the user feedback when they challenge the robot's decisions. The current literature has focused on directly asking the user's preferences towards how a task must be performed [28] or the preferences have been learned implicitly from the user's input [30, 37]. We aim to develop an agentic system in charge of proactively eliciting a user's preference and their reasons, and to interpret the possible feedback they might provide. Feedback can be in the form of new arguments, contradicting the robot's decision, or correcting the robot's user model, such as clarifying the arguments' importance [26, 39]. These latter forms of feedback must be interpreted to properly update the user base scores, to have a more accurate user model. Therefore, we will also develop a method that, when the user corrects the robot, it will update the user base scores accordingly, without the need to retrain a data-driven model.

RQ.4 Contestability: Robots must explain their decisions in a manner that users can clearly understand them, so they can provide explicit feedback to influence the robot's decisions or beliefs [27, 34]. Existing literature has focused on explaining the influence of arguments or relations on the strength of another argument [40, 42], or even counter-factual explanations to determine the direction and magnitude an argument must change to achieve a desired goal [41]. It is required to convert these explanations into natural language to foster understanding. Such explanations must be sufficient so that users can influence the robot's decisions. Our goal is to develop a system that can generate such explanations. Furthermore, since the arguments' final strength computation is linearly complex, it is possible to rapidly determine how changes in the framework will affect future decisions. Therefore, the robot can proactively warn the user of future decisions and wait for confirmation.

User Studies: We plan to validate all the frameworks and methods by performing user studies in assistive scenarios, since we are working closely with healthcare partners in Barcelona. These scenarios include: robots that perform frailty assessments in older adults, which we already implemented [12, 13], robots that feed people who cannot eat autonomously [6], and robots at home [7].

ACKNOWLEDGMENTS

This work has been supported by AGAUR-FI ajuts (2023 FI-3 00065) Joan Oró of the Generalitat of Catalonia and the European Social Plus Fund.

REFERENCES

[1] Leila Amgoud, Jonathan Ben-Naim, Dragan Doder, and Srdjan Vesic. 2017. Acceptability semantics for weighted argumentation frameworks. In *Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI)*. International Joint Conferences on Artifical Intelligence (IJCAI).

[2] Leila Amgoud, Yannis Dimopoulos, and Pavlos Moraitis. 2008. Making Decisions through Preference-Based Argumentation. *KR* (2008), 963–970.

[3] Leila Amgoud and Srdjan Vesic. 2012. On the use of argumentation for multiple criteria decision making. In *International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*. Springer, 480–489.

[4] Antonio Andriella, Carme Torras, Carla Abdelnour, and Guillem Alenyà. 2022. Introducing CARESSER: A framework for in situ learning robot social assistance from expert knowledge and demonstrations. *User Modeling and User-Adapted Interaction* (2022), 441.

[5] Antonio Andriella, Carme Torras, and Guillem Alenyà. 2025. *Personalising Human-Robot Interactions in Social Contexts*. Springer.

[6] Cristian Barrué, Alejandro Suárez, Marco Inzitari, Aida Ribera, and Guillem Alenyà. 2024. NYAM: the role of configurable engagement strategies in robotic-assisted feeding. In *Companion of the ACM/IEEE International Conference on Human-Robot Interaction*. 228–232.

[7] Ermanno Bartoli, Dennis Rotondi, Kai O Arras, and Iolanda Leite. 2025. Long-Term Planning Around Humans in Domestic Environments with 3D Scene Graphs. *arXiv preprint arXiv:2503.09173* (2025).

[8] Zia Bashir, Saima Mahnaz, and Muhammad Ghulam Abbas Malik. 2021. Conflict resolution using game theory and rough sets. *International Journal of Intelligent Systems* (2021), 237–259.

[9] Lauriane Blavette, Sébastien Dacunha, Xavier Alameda-Pineda, Daniel Hernández García, Sharon Gannot, Florian Gras, Nancie Gunson, Séverin Lemaignan, Michal Polic, Pinchas Tandeitnik, et al. 2025. Acceptability and usability of a socially assistive robot integrated with a large language model for enhanced human-robot interaction in a geriatric care institution: mixed methods evaluation. *JMIR Human Factors* (2025), e76496.

[10] Gerard Canal, Guillem Alenyà, and Carme Torras. 2016. Personalization framework for adaptive robotic feeding assistance. In *International conference on social robotics*. Springer, 22–31.

[11] Micah Carroll, Davis Foote, Anand Siththaranjan, Stuart Russell, and Anca Dragan. 2024. AI alignment with changing and influenceable reward functions. In *Proceedings of the 41st International Conference on Machine Learning*. 5706–5756.

[12] Aniol Civit, Antonio Andriella, Maite Antonio, Casimiro Javierre, Concepción Boqué, and Guillem Alenyà. 2024. Exploring the potential of a robot-assisted frailty assessment system for elderly care. In *33rd IEEE International Conference on Robot and Human Interactive Communication (ROMAN)*. IEEE, 168–175.

[13] Aniol Civit, Antonio Andriella, Cristian Barrue, Maite Antonio, Concepción Boqué, and Guillem Alenyà. 2024. Introducing social robots to assess frailty in older adults. In *Companion of the ACM/IEEE International Conference on Human-Robot Interaction*. 342–346.

[14] Aniol Civit, Antonio Andriella, Carles Sierra, and Guillem Alenyà. 2025. Multi-User Personalisation in Human-Robot Interaction: Using Quantitative Bipolar Argumentation Frameworks for Preferences Conflict Resolution. *arXiv preprint arXiv:2511.03576* (2025).

[15] Aniol Civit, Antonio Rago, Antonio Andriella, Guillem Alenyà, and Francesca Toni. 2026. From User Preferences to Base Score Extraction Functions in Gradual Argumentation (with Appendix). *arXiv preprint arXiv:2602.14674* (2026).

[16] Oana Cocarascu, Antonio Rago, and Francesca Toni. 2019. Extracting dialogical explanations for review aggregations with argumentative dialogical agents. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*. Association for Computing Machinery, 1261–1269.

[17] Louise Dupuis De Tarlé, Elise Bonzon, and Nicolas Maudet. 2022. Multiagent dynamics of gradual argumentation semantics. In *21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*.

[18] Phan Minh Dung. 1995. An argumentation-theoretic foundation for logic programming. *The Journal of logic programming* (1995), 151–177.

[19] Matteo Meregalli Falerni, Vincenzo Pomponi, Hamid Reza Karimi, Matteo Lavit Nicora, Matteo Malosio, Loris Roveda, et al. 2024. A framework for human–robot collaboration enhanced by preference learning and ergonomics. *Robotics and Computer-Integrated Manufacturing* (2024), 102781.

[20] Xiuyi Fan and Francesca Toni. 2012. Argumentation dialogues for two-agent conflict resolution. In *Computational Models of Argument*. IOS Press, 249–260.

[21] Vikas Hassija, Vinay Chamola, Atmesh Mahapatra, Abhinandan Singal, Divyansh Goel, Kaizhu Huang, Simone Scardapane, Indro Spinelli, Mufti Mahmud, and Amir Hussain. 2024. Interpreting black-box models: a review on explainable artificial intelligence. *Cognitive Computation* (2024), 45–74.

[22] Bob M Hofstede, Sima Ipakchian Askari, Dirk Lukkien, Laëtitia Gosetto, Janna W Alberts, Ephrem Tesfay, Minke ter Stal, Tom van Hoesel, Raymond H Cuijpers, Martijn H Vastenburg, et al. 2025. A field study to explore user experiences with socially assistive robots for older adults: emphasizing the need for more interactivity and personalisation. *Frontiers in Robotics and AI* (2025), 1537272.

[23] Xiao Huang, Yong Tian, Jiangchen Li, Naizhong Zhang, Xingchen Dong, Yue Lv, and Zhixiong Li. 2025. Joint autonomous decision-making of conflict resolution and aircraft scheduling based on triple-aspect improved multi-agent reinforcement learning. *Expert Systems with Applications* (2025), 127024.

[24] Bahar Irfan, Nathalia Céspedes, Jonathan Casas, Emmanuel Senft, Luisa F Gutiérrez, Mónica Rincon-Roncancio, Carlos A Cifuentes, Tony Belpaeme, and Marcela Múnera. 2023. Personalised socially assistive robot for cardiac rehabilitation: Critical reflections on long-term interactions in the real world. *User Modeling and User-Adapted Interaction* (2023), 497–544.

[25] Bahar Irfan, Sanna Kuoppamäki, and Gabriel Skantze. 2024. Recommendations for designing conversational companion robots with older adults through foundation models. *Frontiers in Robotics and AI* (2024), 1363713.

[26] Francesco Leofante, Hamed Ayoobi, Adam Dejl, Gabriel Freedman, Deniz Gorur, Junqi Jiang, Guilherme Paulino-Passos, Antonio Rago, Anna Rapberger, Fabrizio Russo, Xiang Yin, Dekai Zhang, and Francesca Toni. 2024. Contestable AI Needs Computational Argumentation. In *Proceedings of the 21st International Conference on Principles of Knowledge Representation and Reasoning*. 888–896.

[27] Tamlin Love, Antonio Andriella, and Guillem Alenyà. 2024. What Would I Do If…? Promoting Understanding in HRI through Real-Time Explanations in the Wild. In *33rd International Conference on Robot and Human Interactive Communication (ROMAN)*. IEEE, 504–509.

[28] Tamon Miyake, Yushi Wang, Pin-chu Yang, and Shigeki Sugano. 2023. Feasibility study on parameter adjustment for a humanoid using LLM tailoring physical care. In *International Conference on Social Robotics*. Springer, 230–243.

[29] Hae Won Park, Ishaan Grover, Samuel Spaulding, Louis Gomez, and Cynthia Breazeal. 2019. A model-free affective reinforcement learning approach to personalization of an autonomous social robot companion for early literacy education. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 687–694.

[30] Maithili Patel and Sonia Chernova. 2025. Robot behavior personalization from sparse user feedback. *IEEE Robotics and Automation Letters* (2025).

[31] Antonio Rago, Hengzhi Li, and Francesca Toni. 2023. Interactive Explanations by Conflict Resolution via Argumentative Exchanges. In *Proceedings of the International Conference on Principles of Knowledge Representation and Reasoning*. 582–592.

[32] Antonio Rago and Francesca Toni. 2017. Quantitative argumentation debates with votes for opinion polling. In *International Conference on Principles and Practice of Multi-Agent Systems*. Springer, 369–385.

[33] Zhongqiang Ren, Jiaoyang Li, Han Zhang, Sven Koenig, Sivakumar Rathinam, and Howie Choset. 2023. Binary branching multi-objective conflict-based search for multi-agent path finding. In *Proceedings of the International Conference on Automated Planning and Scheduling*. 361–369.

[34] Fatai Sado, Chu Kiong Loo, Wei Shiung Liew, Matthias Kerzel, and Stefan Wermter. 2023. Explainable goal-driven agents and robots-a comprehensive review. *Comput. Surveys* (2023), 1–41.

[35] Cory-Ann Smarr, Akanksha Prakash, Jenay M Beer, Tracy L Mitzner, Charles C Kemp, and Wendy A Rogers. 2012. Older adults' preferences for and acceptance of robot assistance for everyday living tasks. In *Proceedings of the human factors and ergonomics society annual meeting*. Sage Publications Sage CA: Los Angeles, CA, 153–157.

[36] Katie Trainum, Jiaying Liu, Elliott Hauser, and Bo Xie. 2024. Nursing staff's attitudes, needs, and preferences for care robots in assisted living facilities: a systematic literature review. In *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*. 1058–1062.

[37] Jimmy Wu, Rika Antonova, Adam Kan, Marion Lepert, Andy Zeng, Shuran Song, Jeannette Bohg, Szymon Rusinkiewicz, and Thomas Funkhouser. 2023. Tidybot: Personalized robot assistance with large language models. *Autonomous Robots* (2023), 1087–1102.

[38] Jinyu Yang, Camille Vindolet, Julio Rogelio Guadarrama Olvera, and Gordon Cheng. 2024. On the impact of robot personalization on human-robot interaction: A review. *arXiv preprint arXiv:2401.11776* (2024).

[39] Xiang Yin, Nico Potyka, Antonio Rago, Timotheus Kampik, and Francesca Toni. 2025. Contestability in Quantitative Argumentation. *arXiv preprint arXiv:2507.11323* (2025).

[40] Xiang Yin, Nico Potyka, and Francesca Toni. 2023. Argument attribution explanations in quantitative bipolar argumentation frameworks. In *26th European Conference on Artificial Intelligence*. IOS Press, 2898–2905.

[41] Xiang Yin, Nico Potyka, and Francesca Toni. 2024. CE-QArg: counterfactual explanations for quantitative bipolar argumentation frameworks. In *Proceedings of the 21st International Conference on Principles of Knowledge Representation and Reasoning*. 697–707.

[42] Xiang Yin, Nico Potyka, and Francesca Toni. 2024. Explaining arguments' strength: unveiling the role of attacks and supports. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*. 3622–3630.