Engineering Inter-Agent Explainability in BDI Agents

Katharine Beaumont* Elena Yan** Samuele Burattini*** Rem Collier*

*UCD School of Computer Science, University College Dublin, Belfield, Dublin 4, Ireland
**Mines Saint-Etienne, Univ Clermont Auvergne, INP Clermont Auvergne, CNRS, UMR 6158 LIMOS, F-42023 Saint-Etienne France
***Department of Computer Science and Engineering, Alma Mater Studiorum, University of Bologna, Italy

elena.yan@emse.fr

EXTRAAMAS@AAMAS 2025
19 May 2025

Introduction

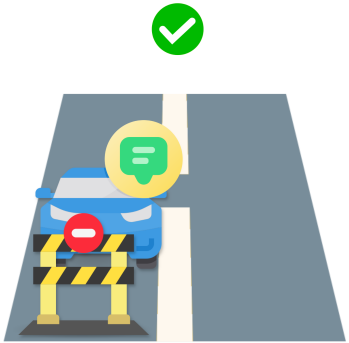
BDI agents provide a natural support for building explanations, thanks to their explicit representation of the agent's mental state (i.e., beliefs, desires, and intentions) [Rao and Georgeff, 1995]

Introduction

BDI agents provide a natural support for building explanations, thanks to their explicit representation of the agent's mental state (i.e., beliefs, desires, and intentions) [Rao and Georgeff, 1995]

Introduction

BDI agents provide a natural support for building explanations, thanks to their explicit representation of the agent's mental state (i.e., beliefs, desires, and intentions) [Rao and Georgeff, 1995]



Introduction

BDI agents provide a natural support for building explanations, thanks to their explicit representation of the agent's mental state (i.e., beliefs, desires, and intentions) [Rao and Georgeff, 1995]

Explainability research mainly targets to *humans*.

Introduction

BDI agents provide a natural support for building explanations, thanks to their explicit representation of the agent's mental state (i.e., beliefs, desires, and intentions) [Rao and Georgeff, 1995]

Explainability research mainly targets to *humans*.

Inter-agent explainability is crucial for improving communication, trust, and knowledge sharing in multi-agent systems.

Introduction

BDI agents provide a natural support for building explanations, thanks to their explicit representation of the agent's mental state (i.e., beliefs, desires, and intentions) [Rao and Georgeff, 1995]

Explainability research mainly targets to *humans*.

Inter-agent explainability is crucial for improving communication, trust, and knowledge sharing in multi-agent systems.

Methodology

1. We review the literature on BDI agent explainability and limited work on inter-agent explainability
2. We frame the explainable agency requirements [Langley, 2019] as *research questions (RQ)* for BDI inter-agent explainability
3. We provide *implementation strategies (IS)* for enabling BDI inter-agent explainability in practice

Request Explanations

RQ1 How are explanations requested?

- User interface [Ahlbrecht, 2023, Yan et al., 2025]
- Why or why not questions [Yan et al., 2025, Winikoff et al., 2021, Winikoff, 2024]
- Dialogue [Dennis and Oren, 2022, Panisson et al., 2021, Espinoza et al., 2019]

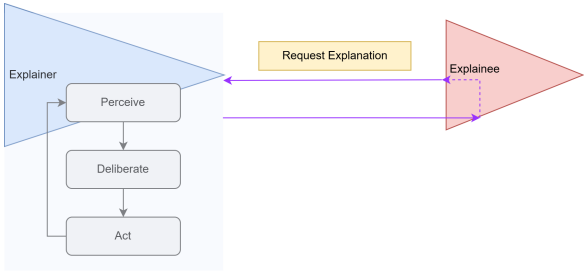
Request Explanations

RQ1 How are explanations requested?

- User interface [Ahlbrecht, 2023, Yan et al., 2025]
- Why or why not questions [Yan et al., 2025, Winikoff et al., 2021, Winikoff, 2024]
- Dialogue [Dennis and Oren, 2022, Panisson et al., 2021, Espinoza et al., 2019]

IS1 Inter-agent explanation protocol to exchange requests

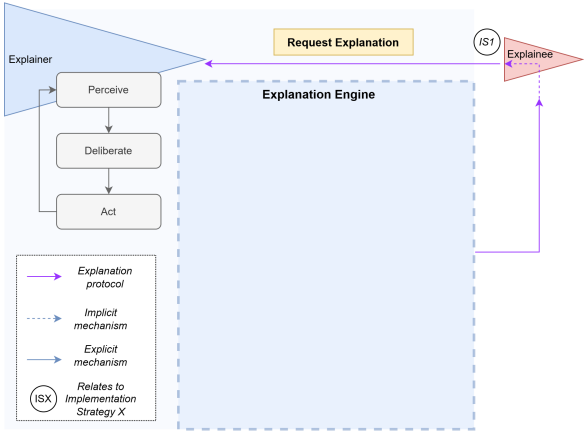
- Interaction protocol to exchange requests and explanations using messages



Generate Explanatory Content

RQ2 How is the explanatory content generated?

- Implicit collection of explanatory content [Alelaimat et al., 2023, Ahlbrecht, 2023, Broekens et al., 2010] [Dennis and Oren, 2022, Harbers et al., 2010, Winikoff et al., 2021, Yan et al., 2025]



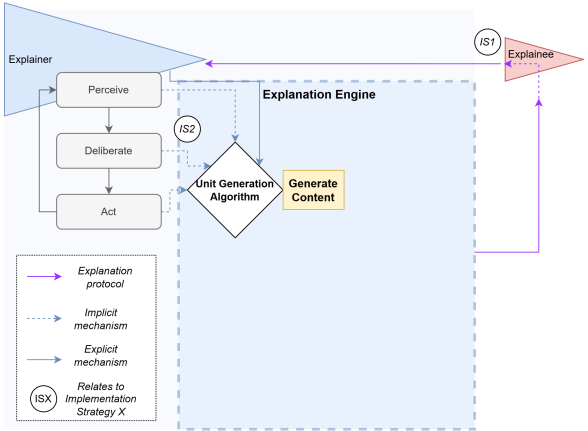
Generate Explanatory Content

RQ2 How is the explanatory content generated?

- Implicit collection of explanatory content [Alelaimat et al., 2023, Ahlbrecht, 2023, Broekens et al., 2010] [Dennis and Oren, 2022, Harbers et al., 2010, Winikoff et al., 2021, Yan et al., 2025]

IS2 Implicit and explicit addition mechanisms of explanatory content

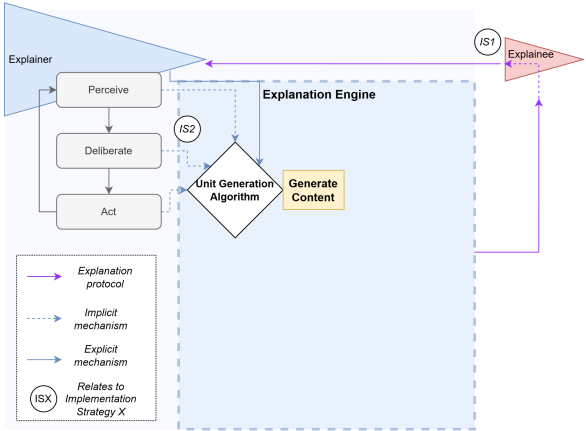
- Automatic implicit addition mechanism
- Configurable explicit addition mechanism with language-level support



Store Explanatory Content

RQ3 How is the explanatory content stored?

- Tree [Ahlbrecht, 2023, Broekens et al., 2010, Harbers et al., 2010] [Winikoff et al., 2021]
- Log/Trace [Dennis and Oren, 2022, Rodriguez et al., 2024, Yan et al., 2025]



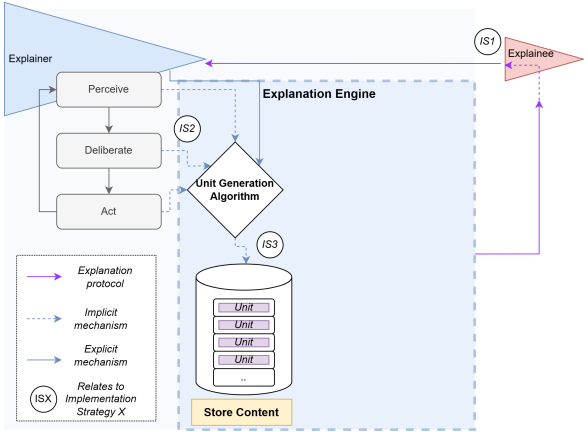
Store Explanatory Content

RQ3 How is the explanatory content stored?

- Tree [Ahlbrecht, 2023, Broekens et al., 2010, Harbers et al., 2010] [Winikoff et al., 2021]
- Log/Trace [Dennis and Oren, 2022, Rodriguez et al., 2024, Yan et al., 2025]

IS3 Runtime state identifiers and state inspection mechanisms

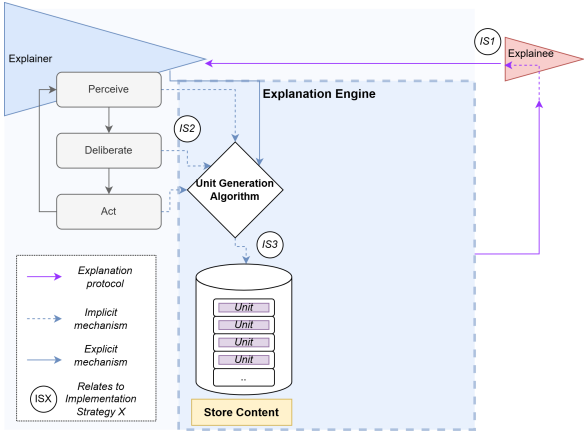
- Language-level identifiers for the core elements and their relationships of an agent runtime state



Generate Explanations

RQ4 How is explanatory content retrieved and used to generate explanations?

1. Retrieve content from the explanation store
2. Process it to generate the final explanation

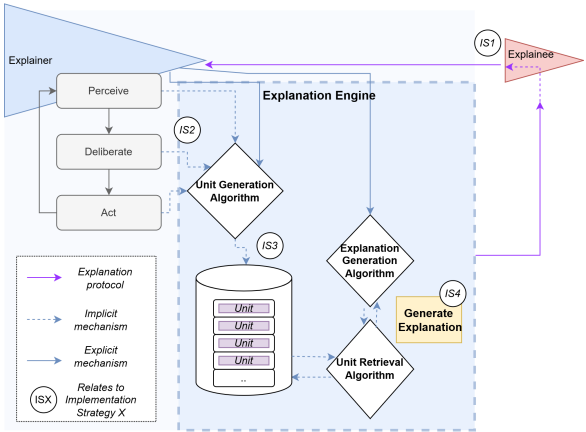


Generate Explanations

RQ4 How is explanatory content retrieved and used to generate explanations?

1. Retrieve content from the explanation store
2. Process it to generate the final explanation

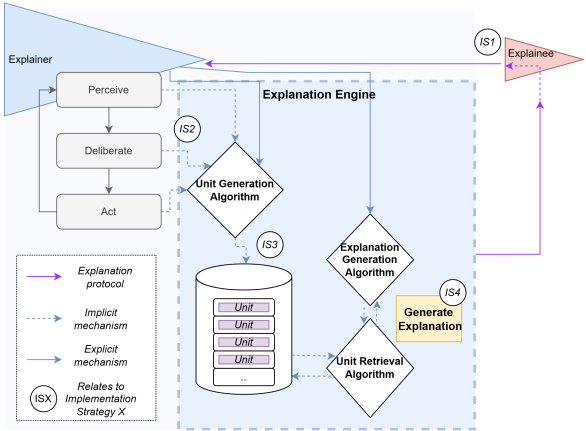
IS4 Configurable retrieval and explanation generation mechanisms



Communicate Explanations

RQ5 How are explanations communicated?

- Narrative [Broekens et al., 2010, Harbers et al., 2010, Yan et al., 2025] [Winikoff et al., 2021]
- Dialogue [Dennis and Oren, 2022]
- User interface [Ahlbrecht, 2023, Yan et al., 2025]

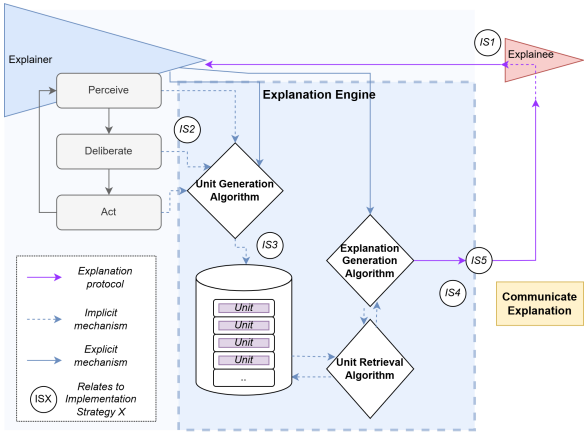


Communicate Explanations

RQ5 How are explanations communicated?

- Narrative [Broekens et al., 2010, Harbers et al., 2010, Yan et al., 2025] [Winikoff et al., 2021]
- Dialogue [Dennis and Oren, 2022]
- User interface [Ahlbrecht, 2023, Yan et al., 2025]

IS5 Uniform representations for machine-understandable explanations

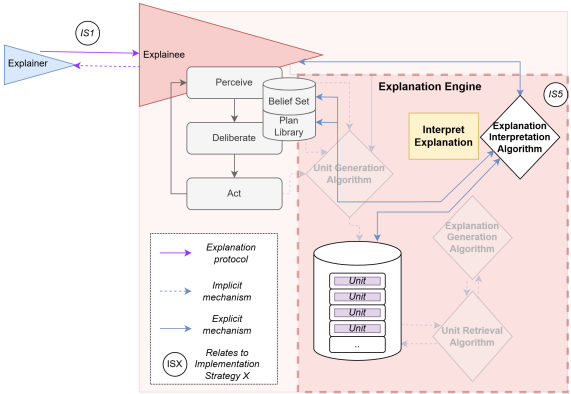


Communicate Explanations

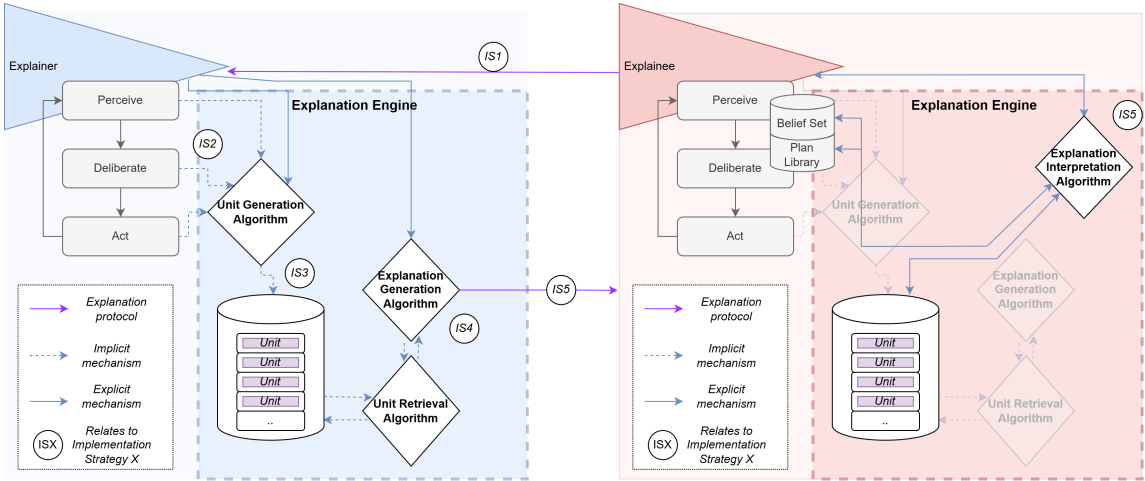
RQ5 How are explanations communicated?

- Narrative [Broekens et al., 2010, Harbers et al., 2010, Yan et al., 2025] [Winikoff et al., 2021]
- Dialogue [Dennis and Oren, 2022]
- User interface [Ahlbrecht, 2023, Yan et al., 2025]

IS5 Uniform representations for machine-understandable explanations



BDI Inter-Agent Explainability



Opportunities and Applications

- **Semantic Web** *(e.g., facilitate the generation of explanatory content and mutual understanding)*
- **Norms** *(e.g., explaining norm deviations)*
- **Organizations and Institutions** *(e.g., explaining the organization's structure, functioning)*
- **Argumentation and Negotiation** *(e.g., arguments can be used in the explanation)*
- **Trust and Reputation** *(e.g., explaining agent behavior or the reason behind the trust or reputation evaluations)*
- **Recommendations** *(e.g., enhance the sharing of recommendations)*
- **Cooperative Learning** *(e.g., enhance the sharing of knowledge)*
- ...

Conclusions and Future Work

We present mechanisms for engineering *inter-agent explainability in BDI agents*, highlighting promising research directions, opportunities, and applications.

Future work will implement these mechanisms in a specific BDI agent technology to demonstrate their practicality.

Thank you for your attention!

For further information:

Katharine Beaumont, Elena Yan, Samuele Burattini, and Rem Collier.
Engineering Inter-Agent Explainability in BDI Agents.
International Workshop on EXplainable, Trustworthy, and Responsible AI and Multi-Agent Systems (EXTRAAMAS 2025), May 2025.

Engineering Inter-Agent Explainability in BDI Agents

Katharine Beaumont* Elena Yan** Samuele Burattini*** Rem Collier*

*UCD School of Computer Science, University College Dublin, Belfield, Dublin 4, Ireland
**Mines Saint-Etienne, Univ Clermont Auvergne, INP Clermont Auvergne, CNRS, UMR 6158 LIMOS, F-42023 Saint-Etienne France
***Department of Computer Science and Engineering, Alma Mater Studiorum, University of Bologna, Italy

elena.yan@emse.fr

EXTRAAMAS@AAMAS 2025
19 May 2025

References I

[Ahlbrecht, 2023] Ahlbrecht, T. (2023).
An algorithmic debugging approach for belief-desire-intention agents.
Annals of Mathematics and Artificial Intelligence, pages 1–18.

[Alelaimat et al., 2023] Alelaimat, A., Ghose, A., and Dam, H. K. (2023).
Mining and validating belief-based agent explanations.
In Calvaresi, D., Najjar, A., Omicini, A., Aydogan, R., Carli, R., Ciatto, G., Mualla, Y., and Främling, K., editors, *Explainable and Transparent AI and Multi-Agent Systems*, pages 3–17, Cham. Springer Nature Switzerland.

[Broekens et al., 2010] Broekens, J., Harbers, M., Hindriks, K., Van Den Bosch, K., Jonker, C., and Meyer, J.-J. (2010).
Do you get it? user-evaluated explainable BDI agents.
In *German Conference on Multiagent System Technologies*, pages 28–39. Springer.

[Dennis and Oren, 2022] Dennis, L. A. and Oren, N. (2022).
Explaining BDI agent behaviour through dialogue.
Autonomous Agents and Multi-Agent Systems, 36(2):29.

References II

[Espinoza et al., 2019] Espinoza, M. M., Possebom, A. T., and Tacla, C. A. (2019).

Argumentation-based agents that explain their decisions.

In *8th Brazilian Conference on Intelligent Systems, BRACIS 2019, Salvador, Brazil, October 15-18, 2019*, pages 467–472. IEEE.

[Harbers et al., 2010] Harbers, M., van den Bosch, K., and Meyer, J.-J. (2010).

Design and evaluation of explainable BDI agents.

In *2010 IEEE/WIC/ACM international conference on web intelligence and intelligent agent technology*, volume 2, pages 125–132. IEEE.

[Langley, 2019] Langley, P. (2019).

Explainable, normative, and justified agency.

In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 9775–9779. AAAI Press.

References III

[Panisson et al., 2021] Panisson, A. R., Engelmann, D. C., and Bordini, R. H. (2021).
Engineering explainable agents: An argumentation-based approach.
In *Engineering Multi-Agent Systems - 9th International Workshop, EMAS 2021, Virtual Event, May 3-4, 2021, Revised Selected Papers*, volume 13190 of *Lecture Notes in Computer Science*, pages 273–291. Springer.

[Rao and Georgeff, 1995] Rao, A. S. and Georgeff, M. P. (1995).
BDI agents: From theory to practice.
In Lesser, V. R. and Gasser, L., editors, *Proceedings of the First International Conference on Multiagent Systems, June 12-14, 1995, San Francisco, California, USA*, pages 312–319. The MIT Press.

[Rodriguez et al., 2024] Rodriguez, S., Thangarajah, J., and Davey, A. (2024).
Design patterns for explainable agents (XAg).
In Dastani, M., Sichman, J. S., Alechina, N., and Dignum, V., editors, *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2024, Auckland, New Zealand, May 6-10, 2024*, pages 1621–1629. International Foundation for Autonomous Agents and Multiagent Systems / ACM.

References IV

[Winikoff, 2024] Winikoff, M. (2024).
Towards engineering explainable autonomous systems.
In *International Workshop on Engineering Multi-Agent Systems*, pages 144–155. Springer.

[Winikoff et al., 2021] Winikoff, M., Sidorenko, G., Dignum, V., and Dignum, F. (2021).
Why bad coffee? explaining BDI agent behaviour with valuings.
Artificial Intelligence, 300:103554.

[Yan et al., 2025] Yan, E., Burattini, S., Hübner, J. F., and Ricci, A. (2025).
A multi-level explainability framework for engineering and understanding BDI agents.
Autonomous Agents and Multi-Agent Systems, 39(1):9.