Towards a Grounded Dialog Model for Explainable Artificial Intelligence

Prashan Madumal, Tim Miller, Frank Vetere and Liz Sonenberg

School of Computing and Information Systems
University of Melbourne, Australia
pmathugama@student.unimelb.edu.au

Introduction

Explainable Artificial Intelligence (XAI)

Explaining the behaviours, actions and decisions made by AI systems.

Introduction

Explanation as a Process (Miller, 2017)

Cognitive process: process of determining an explanation for a given event.
Social process: transferring knowledge between explainer and explainee.

Motivation

- Trust.
- Transparency and ethics.

Human Explanation

- Explanation as a continuous interaction.
- How humans engage in conversational explanation
- Explanation models influenced by human explanation more likely to be accepted.
- Easier for the AI to emulate human explanations.

Goal

- Introduce a Human explanation model.
- Grounded on data
- Analyze relationships in dialog components.

Related work

- Early work
  - *Kass and Finin (Kass, 1988) and Moore and Paris (Moore, 1991) discussed the requirements a good explanation facility should have, including characteristics like "Naturalness".*
  - *Cawsey's (Cawsey, 1993) EDGE system also focused on user interaction and user knowledge.*

Related work

- Explanation dialog models
  - *Walton's (Walton, 1993) shift model*

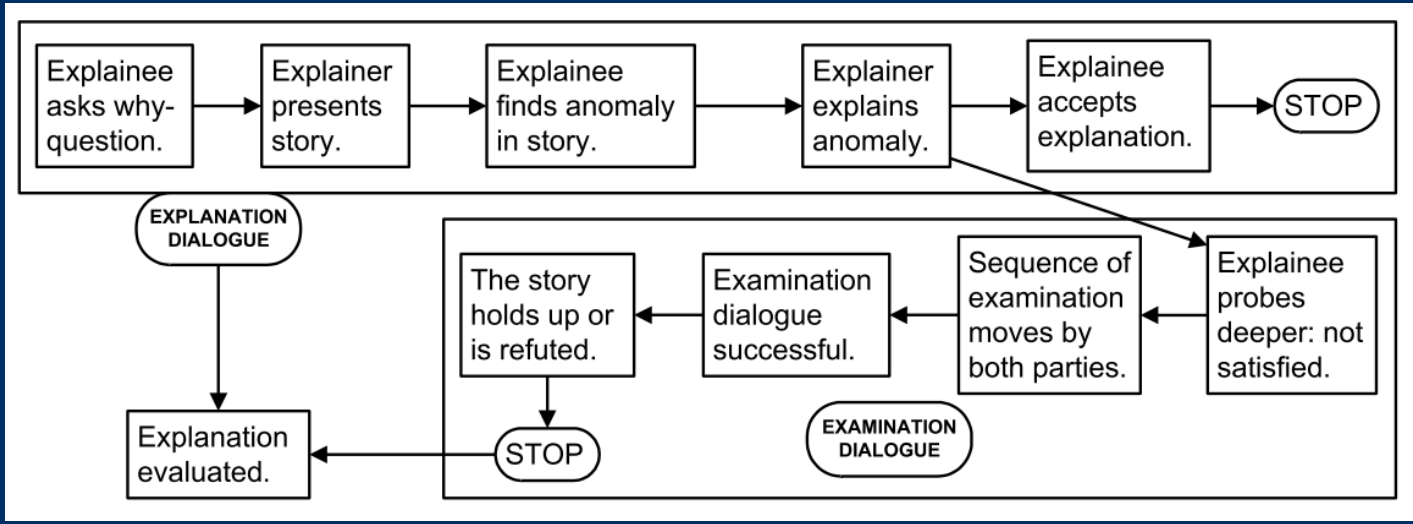


Figure: Argumentation and explanation in dialogue (Walton, 1993)

Research Design and Methodology

- Grounded theory (Glasser, 1967) as the methodology
- Gain insights into three areas:
  1. Key components that makeup an explanation dialog.
  2. Relationships that exist within those components.
  3. Component sequences that occur in an explanation dialog and cycles.

Data

- Six different data sources, Six different types of explanation dialogs.
- Total of 398 explanation dialogs.
- Text based sources, some are transcribed from voice and video-based interviews.

Table: Coded data description.

Explanation Dialog Type	# Dialogs	# Transcripts
1. Human-Human static explainee	88	2
2. Human-Human static explainer	30	3
3. Human-Explainer agent	68	4
4. Human-Explainee agent	17	1
5. Human-Human QnA	50	5
6. Human-Human multiple explainee	145	5

Data cont.

- Different combinations explainee and explainer participants.

Table: Explanation dialog type description.

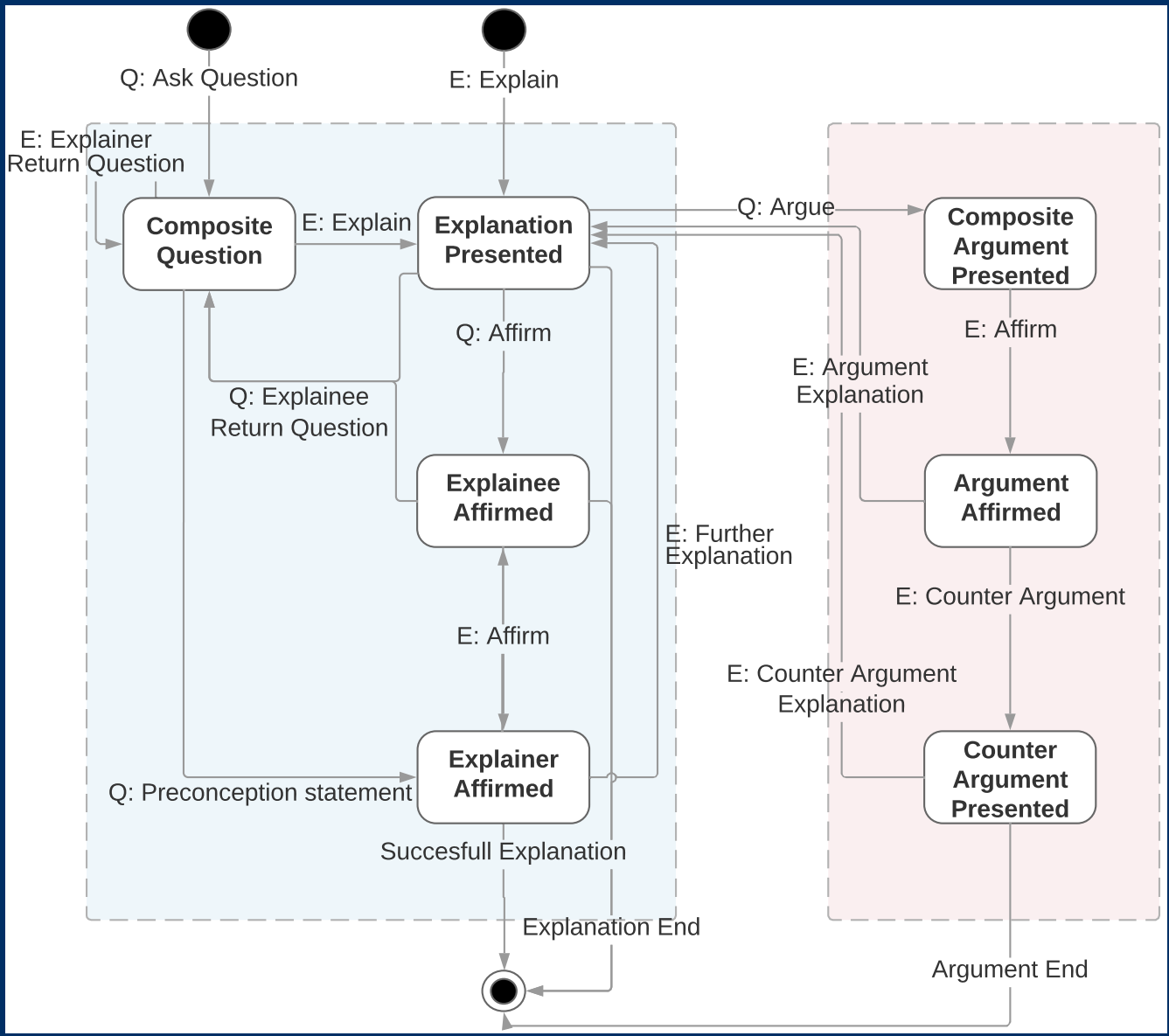
Participants	Number	Medium	Data source
1. Human-Human	One-to-one	Verbal	Journalist Interview transcripts
2. Human-Human	One-to-one	Verbal	Journalist Interview transcripts
3. Human-Agent	One-to-one	Text	Chatbot conversation transcripts
4. Agent-Human	One-to-one	Text	Chatbot conversation transcripts
5. Human-Human	Many-to-many	Text	Reddit AMA records
6. Human-Human	One-to-many	Verbal	Supreme court transcripts

Data cont.

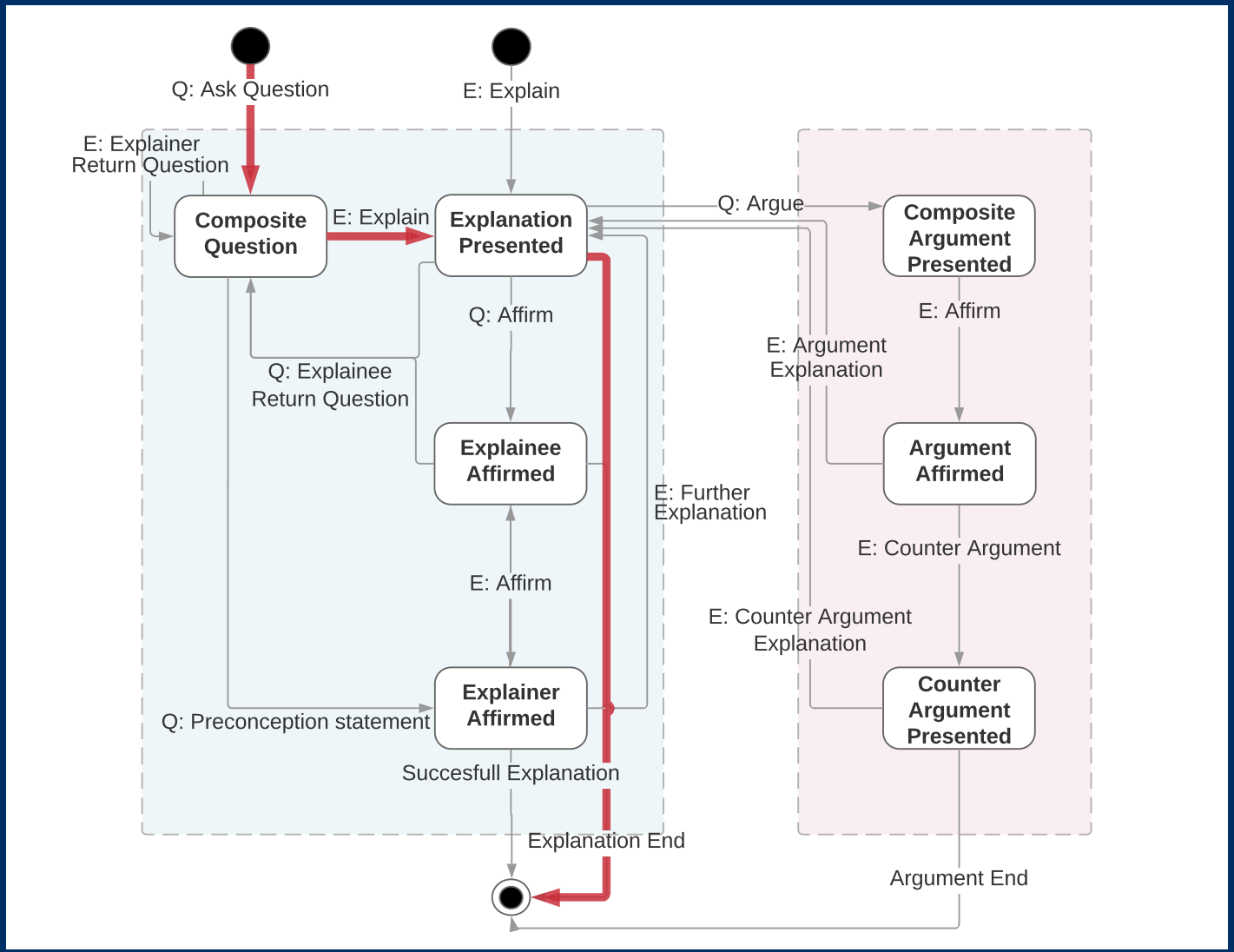
- Coding, categories and their definition.

Code	Category	Description
QE start	Dialog	Explanation dialog start
QE end	Dialog	Explanation dialog end
How	Question Type	How questions
Why	Question Type	Why questions
What	Question Type	What questions
Explanation	Explanation	Explanation given for questions
Explainee Affirmation	Explanation	Explainee acknowledges explanation
Explainer Affirmation	Explanation	Explainer acknowledges explainee's acknowledgment
Question context	Information	Background to the question provided by the explainee
Preconception	Information	Preconceived idea that the explainee has about some fact
Counterfactual case	Information	Counterfactual case of the how/why question
Argument	Argumentation	Argument presented by explainee or explainer
Argument-s	Argumentation	An argument that starts the Dialog
Argument-a	Argumentation	Argument Affirmation by explainee or explainer
Argument-c	Argumentation	Counter argument
Argument-contrast case	Argumentation	Argumentation contrast case
Explainer Return question	Questions	Clarification question by explainer
Explainee Return question	Questions	Follow up question asked by explainee

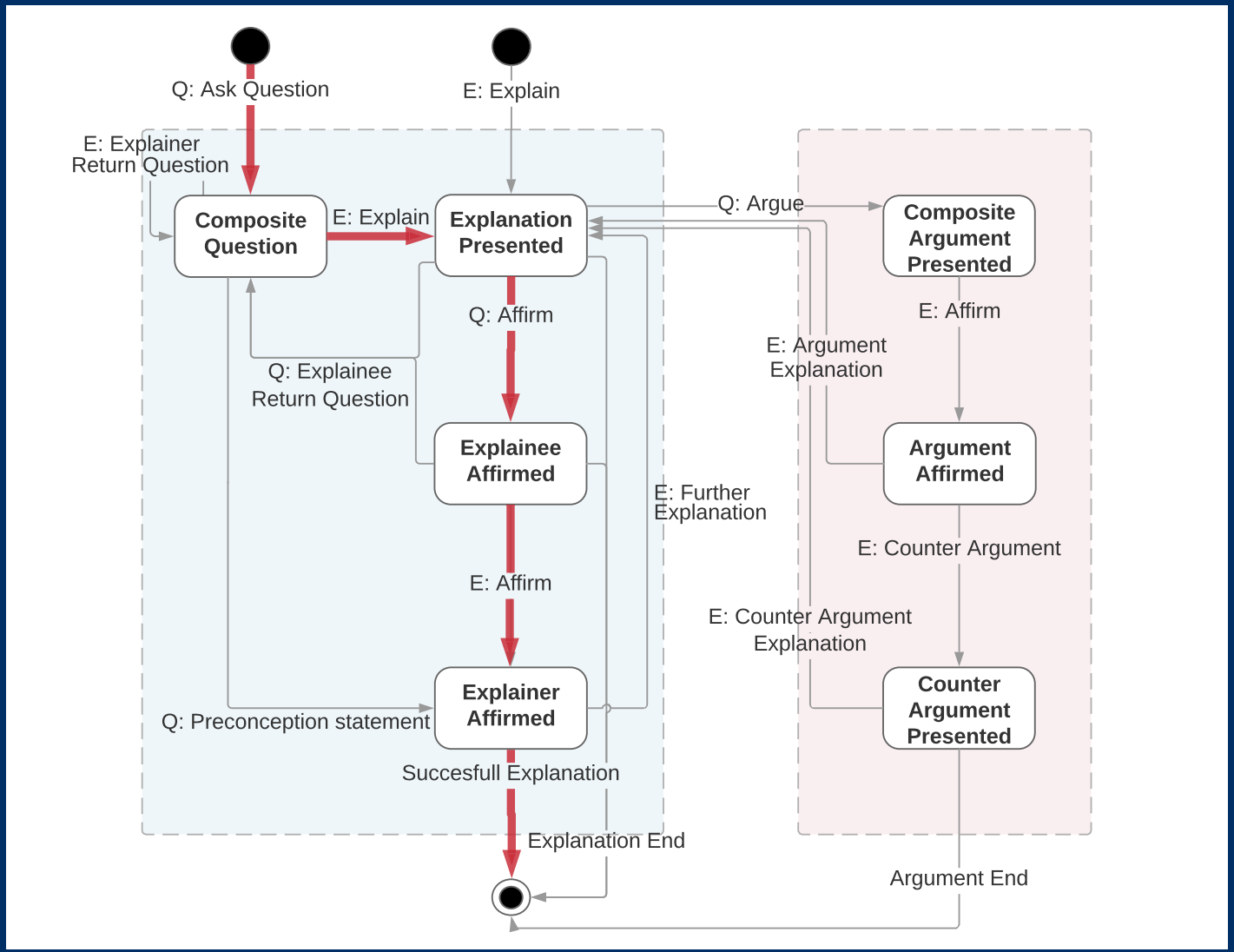
Explanation Dialog Model



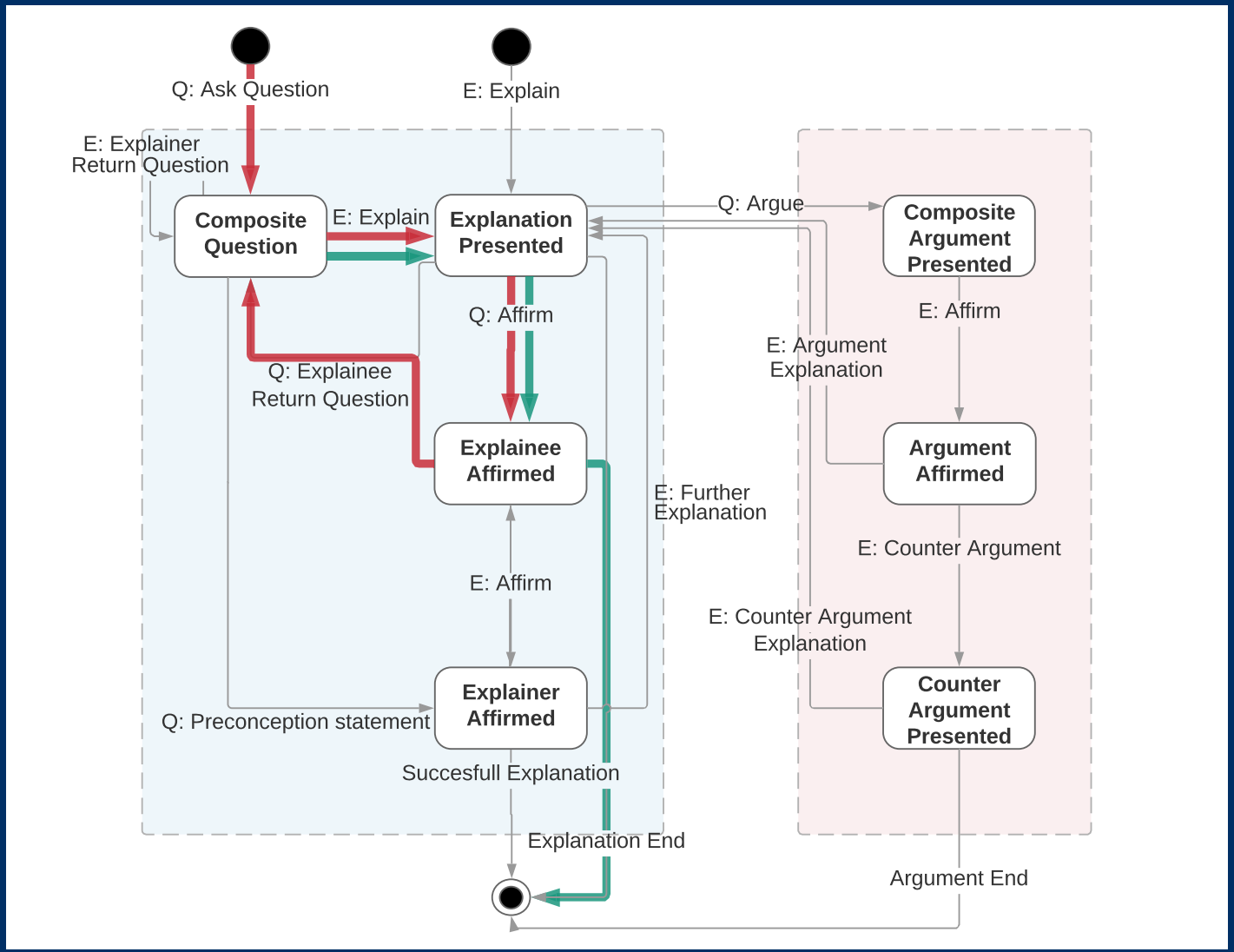
Explanation Dialog Model cont.



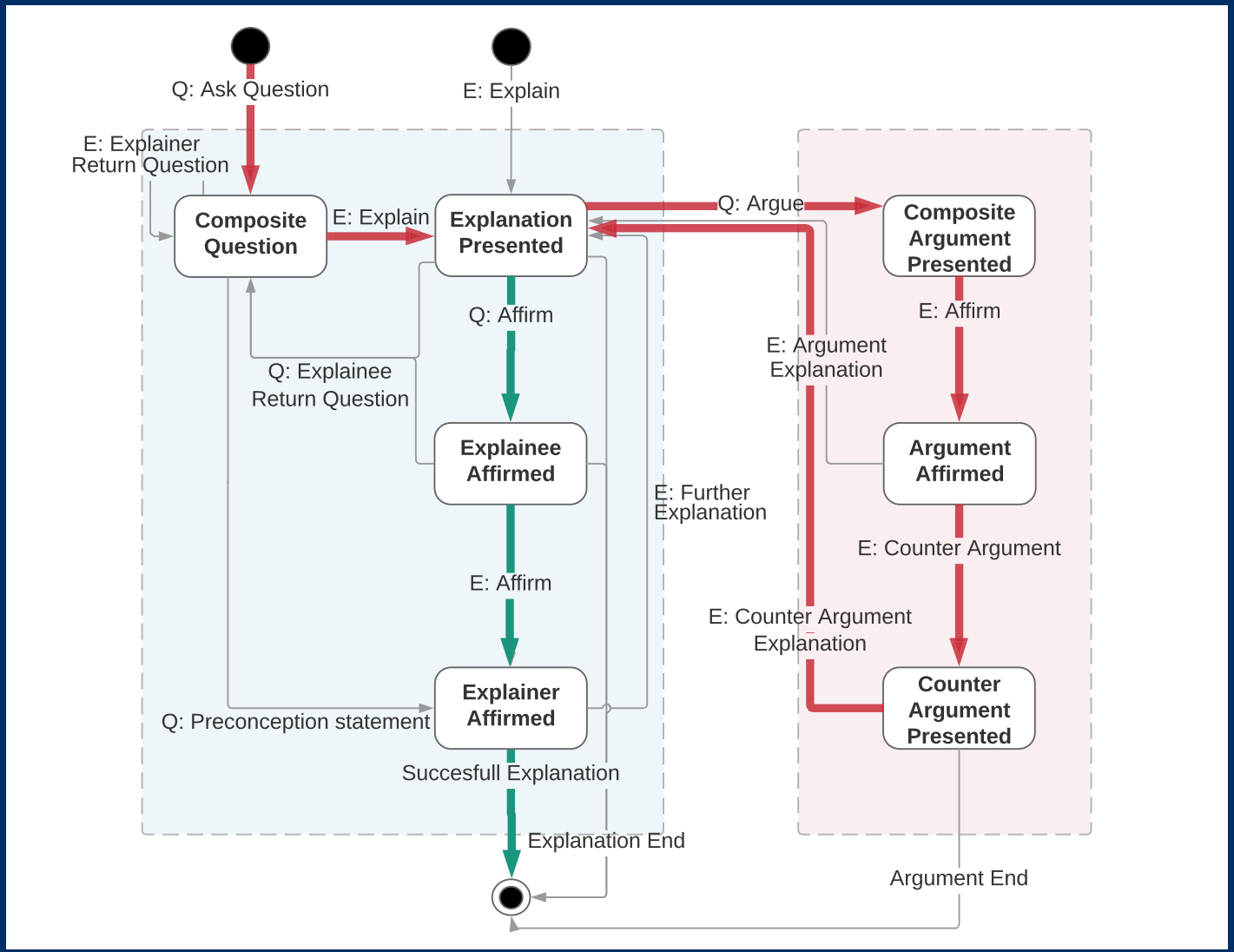
Explanation Dialog Model cont.



Explanation Dialog Model cont.



Explanation Dialog Model cont.



Model Comparison

- Walton's model (Walton, 1993) focus on combining explanation and examination dialogs with argumentation.

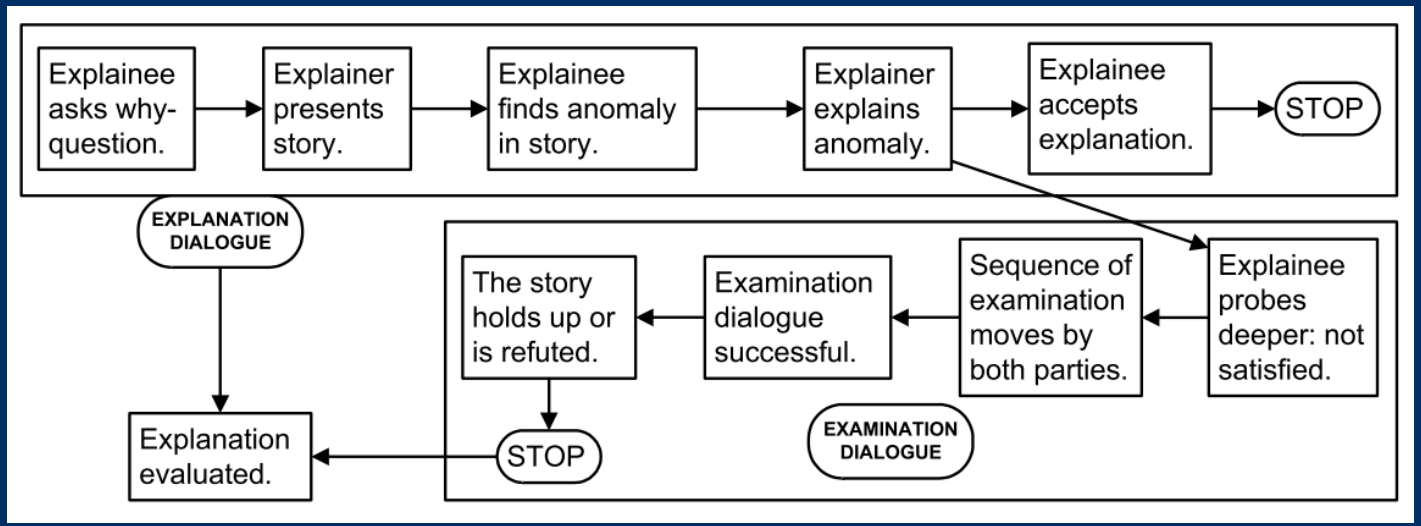


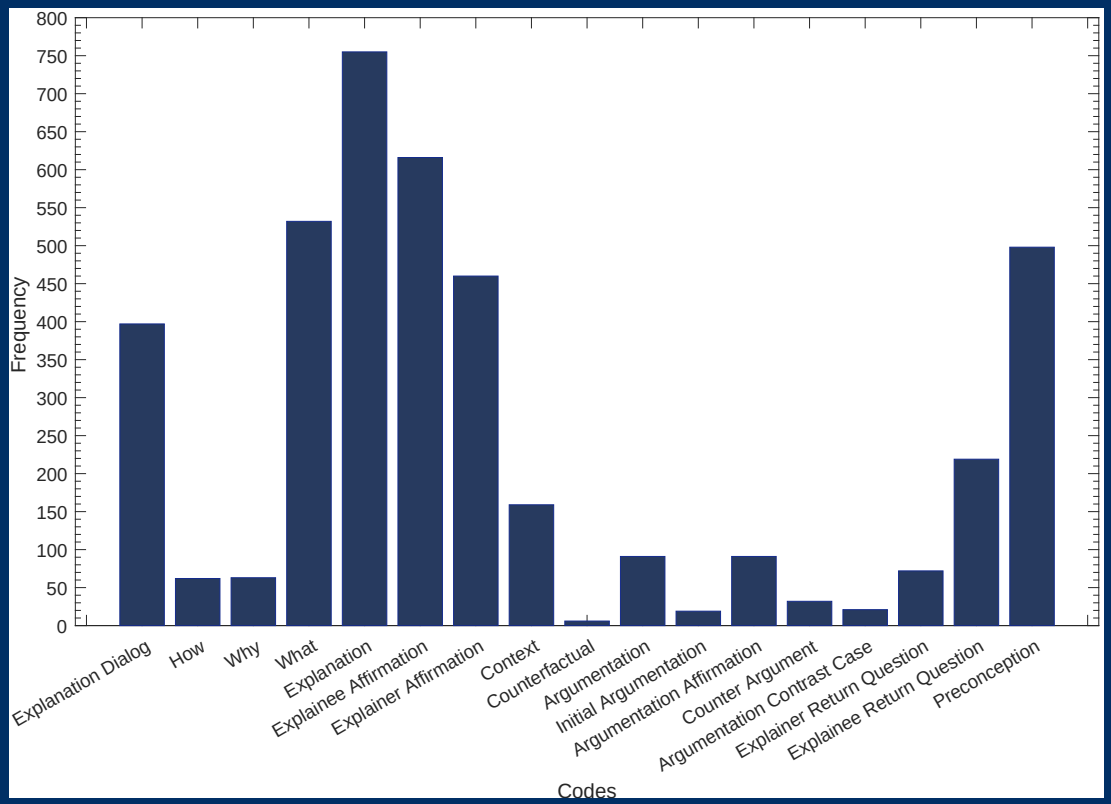
Figure: Argumentation and explanation in dialogue (Walton, 1993)

Model Comparison

- Two differences,
  - The lack of *examination* dialog shift in our model
  - Walton's model focus on the evaluation of the successfulness of an explanation.
- The differences are at a more detailed level than at the high-level.
- Our model captures the subtleties.

Analysis and Evaluation

- Analysis on three areas:
  1. Key components of an Explanation Dialog.
  2. Relationships between these components and their variations between different dialog types.
  3. The sequence of components that can successfully carry out an explanation dialog.



Code Frequency Analysis

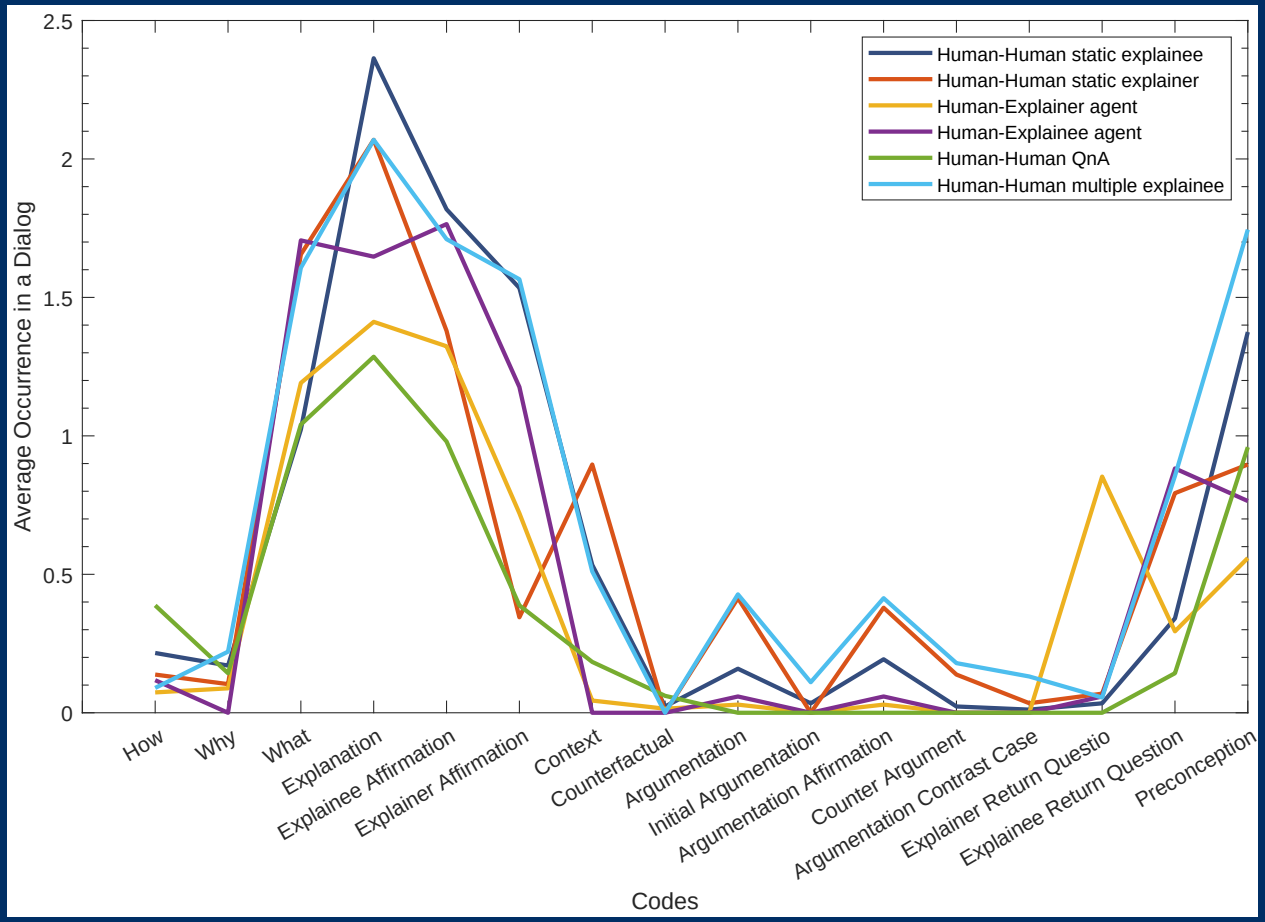


Figure: Average code occurrence in different explanation dialog types

Code Occurrence Analysis per Dialog

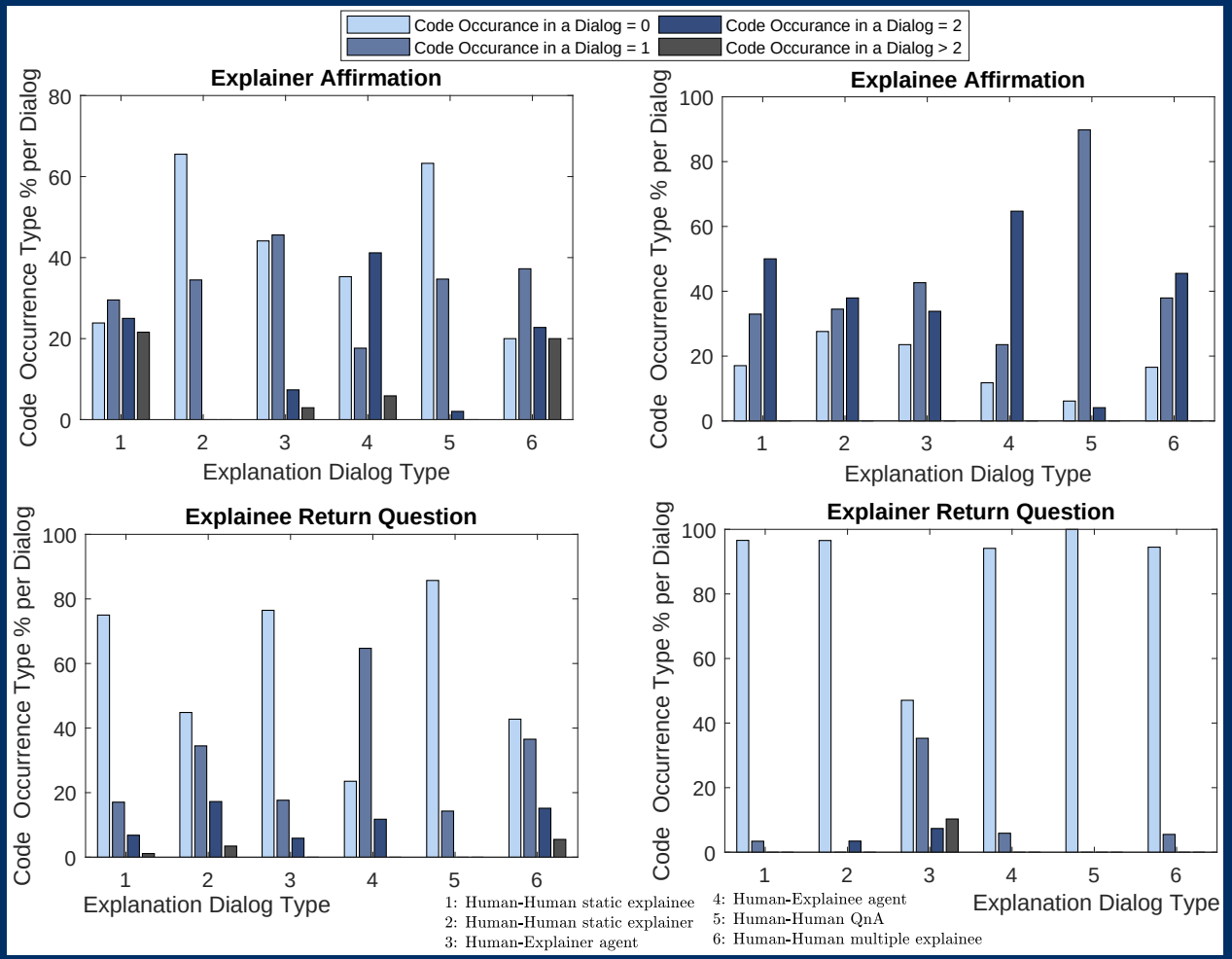


Figure: Average code occurrence per dialog in different dialog types

Explanation Dialog Ending Sequence Analysis

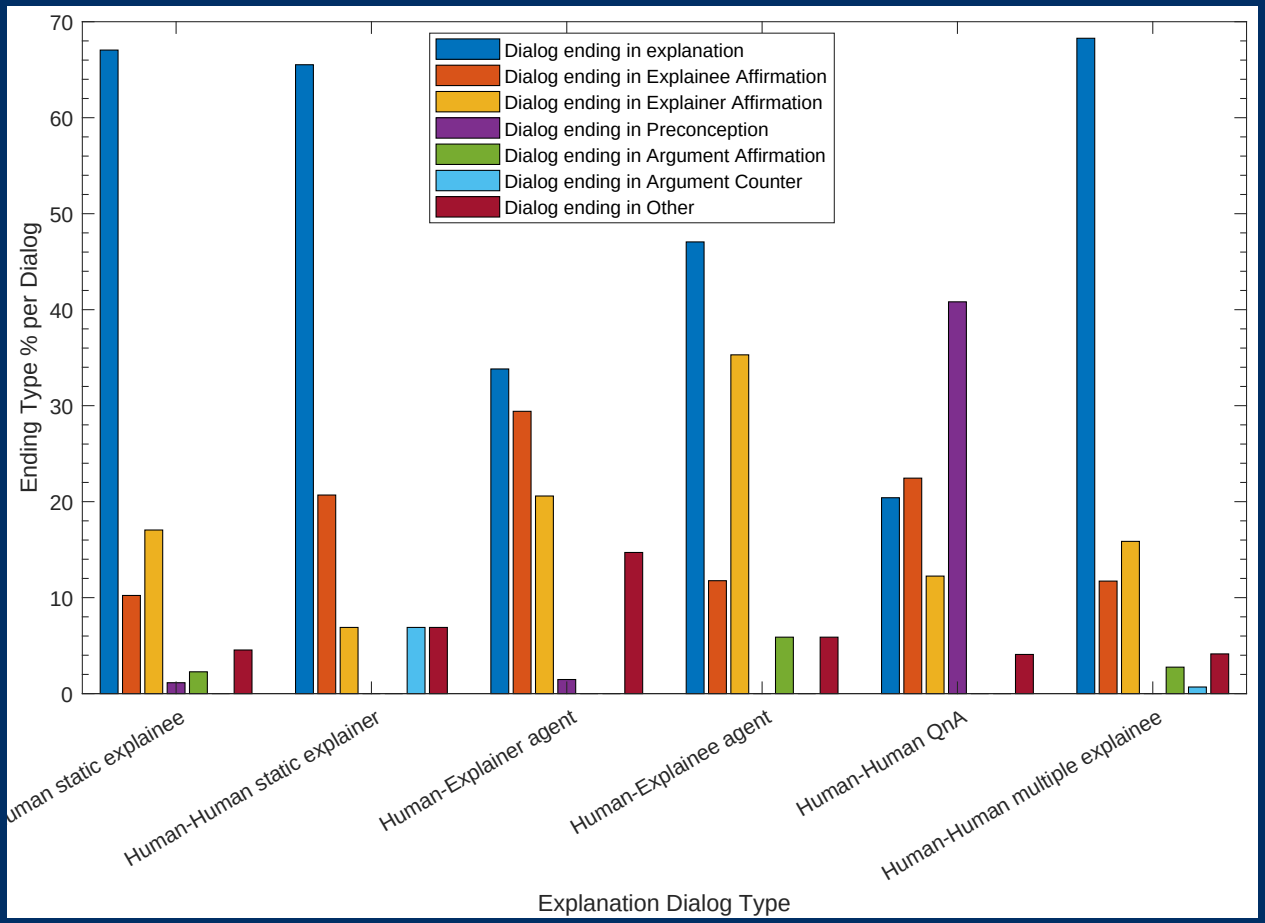


Figure: Average ending code percentage in different dialog types

Discussion

- Using the proposed model in Explainable AI systems.
- Providing interactive explanations, having the freedom for questioning and arguments
- Limitations: *Inability to evaluate effectiveness of the delivered explanation.*

Conclusion and Future Work

- Explanation dialog model is derived from different types of natural conversations between humans as well as humans and agents.
- Formulate the model by analysing the frequency of occurrences of patterns to identify the key components that makeup an explanation dialog.
- relationships between components and their sequence of occurrence inside a dialog.
- XAI systems can build on top of this explanation dialog model to provide better explanations to the intended user.
- Future work: *Evaluate the model in a Human-Agent setting.*

References I

- Cawsey, Alison. "Planning interactive explanations". In: *International Journal of Man-Machine Studies* 38.2 (1993), pp. 169 –199. DOI: http://dx.doi.org/10.1006/imms.1993.1009.
- Glaser, Barney G and Anselm L Strauss. *The Discovery of Grounded Theory: Strategies for Qualitative Research*. Vol. 1. 4. 1967, p. 271. ISBN: 0202302601. DOI: 10.2307/2575405. arXiv: 9809069v1 [arXiv:gr-qc]. URL: http://www.amazon.com/dp/0202302601.
- Kass, Robert and Tim Finin. *The Need for User Models in Generating Expert System Explanations*. Tech. rep. University of Pennsylvania, 1988, pp. 1–32. URL: http://repository.upenn.edu/cis_reports/585.

References II

- Miller, Tim. "Explanation in Artificial Intelligence: Insights from the Social Sciences". In: (2017). DOI: arXiv:1706.07269v1. arXiv: 1706.07269. URL: http://arxiv.org/abs/1706.07269.
- Moore, Johanna D. and Cecile L. Paris. "Requirements for an expert system explanation facility". In: *Computational Intelligence* 7.4 (1991), pp. 367–370. ISSN: 14678640. DOI: 10.1111/j.1467-8640.1991.tb00409.x.
- Walton, Douglas and Floris Bex. "Combining explanation and argumentation in dialogue". In: *Argument and Computation* 7.1 (2016), pp. 55–68. ISSN: 19462174. DOI: 10.3233/AAC-160001.

Q&A

Prashan Madumal: pmathugama@student.unimelb.edu.au