Extended Abstract: Adaptable Social AI Agents

Manuel Preston de Miranda, Mahimul Islam, Rhea Basappa, Travis Taylor, Ashok Goel

Georgia Institute of Technology

mmiranda31@gatech.edu, mahimul@gatech.edu, rb324@gatech.edu, ttaylor99@gatech.edu, ashok.goel@gatech.edu

Abstract

This paper presents enhancements of an AI social agent, SAMI, with episodic self-explanation capabilities, advancing the Theory of its own Mind by allowing for dynamic, context-dependent reasoning about internal decisions. By utilizing an LLM, GPT-4o-mini, and a graphical database, Neo4j, for its knowledge representation and meta-reasoning, the enhanced SAMI is able to promote greater explainable AI (XAI) capabilities and foster trust in human-AI interaction in a personalized education setting.

Introduction

Online learning and especially learning at scale in an online setting has many benefits ranging from increased ease of access to affordability. However, one significant drawback is that it is more difficult for learners to maintain or even initiate connections with other learners (Garrison, Anderson, and Archer 1999). One proposed method to assist with this is Georgia Tech's SAMI (Social Agent Mediated Interactions) AI that aims to connect learners via mutual interest/traits that are obtained from learner posts in an online class discussion forum (Wang et al. 2020; Kakar et al. 2024). An important characteristic of AI is for it to be able to explain its reasoning and innerworkings to help foster trust with users.

Previous work on SAMI aimed to solve this problem by implementing a Task, Method, Knowledge (TMK) framework that revolved around enabling the AI agent to answer static questions about its inner working (Basappa et al. 2024; Goel and Rugaber 2017). The scope of answerable questions was limited to examples such as "What kind of data does SAMI learn from?" and "How often does SAMI make mistakes?" both of which are examples that do not require dynamically changing contextual information. In other words, these questions will always have the same correct answer unless there is some specific update to the inner working of SAMI. While this provides a significant improvement to the Theory of Its Own Mind, it is not able to answer episodic questions. Episodic in this paper is defined as the derivational trace in a given instance of decision making. Examples of episodic questions for SAMI are "Why was I matched with student x?" or "If I said I liked reading would I have been matched differently with student y?". As seen, these questions revolve around the ever-changing interests of learners that may change over time and are specific to a given situation. The proposed work in this paper thus becomes: How can an AI agent be improved so that it is able to provide accurate answers to online learners about its decisions and innerworkings in the context of a dynamically changing environment and input?

Method and Implementation

The SAMI architecture can be thought of as two parts. The first part consists of the initial matchmaking and data collection. To do this a script is run that extracts student information from posts in a discussion forum and stores it in a graph-based knowledge representation, implemented using Neo4j. To represent the data extracted, nodes are used with branches connecting the nodes. Nodes in the database are things such as hobbies, student names, time zones, etc. The links connecting the nodes represent the relation between the nodes such as interested_in or at_time. An example of this relation would be 2 nodes, the first being a student and the second being a hobby and the link between them would be interested_in. Using this setup, the knowledgebase instance is queried and run through a matchmaking algorithm to connect students. This process is manually done a few weeks into a given semester.

The second part of SAMI specifically deals with self-explanation. For this to work an ongoing flask server is instantiated that has access to the knowledgebase instance created by the first part. When students post to the forum and include in their post the text '#samiexplain' the forum sends a request to the SAMI server. This post is then characterized as being either a static or episodic question. If it is static, it

Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

uses the previous TMK method of self-explanation. Alternatively, if it is episodic SAMI uses the new proposed method of self-explanation.

When a student submits a question and it is deemed episodic, the proposed system privatizes the query by anonymizing any mentioned individuals using SpaCy's entity recognition, replacing names with placeholders such as student_name_0. The privatized question is then analyzed by GPT-4o-mini to determine its intent, categorizing it into one of four types: **Personal**, **Relational**, **Other_matches**, or **Private**. These intent types determine what information is needed to fully answer the question. The table below succinctly describes what each intent type represents.

Type	**Description**
Personal	Questions about the student's own attributes or interests not involving any other student.
Relational	Questions about the asking student's relationships or matches.
Other_matches	Questions about nonspecific other student matches and potential matches the asking student could have.
Private	Questions about other students and information about them without any relation to the asking student.

Table 1: Intent type descriptions

Based on the identified intent, relevant information is retrieved from the knowledgebase, such as shared interests between students, user attributes, or names of students that share a certain trait. These results are formatted in a simple natural language representation for later use. Finally, GPT-4o-mini is used to synthesize this data to generate a coherent, context-aware natural language response to the original question which is then posted to the discussion forum. This entire process happens in real time and takes no more than a few seconds, after a question is posted, to provide a response.

Results and Evaluation

In order to validate the answers from SAMI, a set of certified XAI questions were slightly modified and tested on a sandbox instance of Neo4j (Liao, Gruen, and Miller 2020; Sipos et al. 2023). This sandbox instance of the knowledgebase was created as described above but for the learners it uses posts and information in a discussion thread used by other members of the research team rather than a full classroom of students. The validation questions consisted of 18 modified questions from a XAI database and 7 questions that were deemed relevant but not present in the XAI database. To test these questions the answers were evaluated based on completeness and correctness. Correctness being if the answer generated was correct and completeness being if the answer fully answered the question and any needed elements. For example, given a student question such as "Why was I matched with person x?", a correct but incomplete answer would be one such as "You were matched with x because of reason y" versus a correct and complete answer would be "You were matched with x because of thing y and thing z.". The score values were then totaled for each answer to each question with a score of 2 being that the answer was correct and complete and 0 being incorrect and incomplete. Of the 25 tested questions, 100% of the answers generated were deemed correct and complete. Future work for SAMI involves deploying SAMI with enhanced Theory of Its Own Mind in on going classes and determining student reception and feelings as well as deploying surveys to evaluate student opinion on sample answers generated.

Discussion and Conclusion

The importance of AI agents possessing self-explanation capabilities cannot be overstated, especially in the context of education and online learning. Enabling AI agents to self-explain bridges the gap between opaque "black box" algorithms and user understanding, leading to far more transparent interactions. When AI agents can articulate the reasoning behind their choices, it empowers users to more fully engage with and understand the technology they are using, fostering greater trust between the user and the AI.

By enhancing SAMI's self-Theory of Mind to allow for episodic self-explanation, we address the dynamic and constantly changing nature of interactions between humans and AI agents. In a world where change is inevitable, it becomes paramount for AI agents to adapt accordingly. By enabling dynamic reasoning over past decisions, the enhanced SAMI can account for the unique context with each of its users and the specific situations between the users. In an educational setting, particularly when AI is used to match students, such transparency is crucial because AI has the potential to substantially influence a learner's experience in a class.

The evaluation of the enhanced SAMI demonstrates its capability to provide correct and complete answers to episodic questions, thus validating the effectiveness of the new self-explanation features. Future work involves deploying SAMI in active classrooms to collect student data and feedback to assess its ongoing effects.

In conclusion, enabling AI agents to self-explain through the use of a graphical database and leveraging GPT-4o-mini

for reasoning capabilities allows meta reasoning. This advancement can not only improve transparency and trust but may also enhance the overall user experience in an online learning environment. As AI continues to evolve in education, self-reasoning agents like SAMI will be essential in building meaningful connections and trust.

References

Basappa, R.; Tekman, M.; Lu, H.; Faught, B.; Kakar, S.; Goel, A. K. (2024). *Social AI Agents Too Need to Explain Themselves*. In **Intelligent Tutoring Systems: 17th International Conference, ITS 2024, Proceedings, Part I** (pp. 351–360). Cham: Springer.

Garrison, D. R.; Anderson, T.; Archer, W. (1999). *Critical Inquiry in a Text-Based Environment: Computer Conferencing in Higher Education*. **The Internet and Higher Education**, 2(2–3), 87–105. doi.org/10.1016/S1096-7516(00)00016-6

Goel, A. K.; Rugaber, S. (2017). *GAIA: A CAD-Like Environment for Designing Game-Playing Agents*. **IEEE Intelligent Systems**, 32(3), 60–67. doi.org/10.1109/MIS.2017.44

Kakar, S.; Basappa, R.; Camacho, I.; Griswold, C.; Houk, A.; Leung, C.; Tekman, M.; Westervelt, P.; Wang, Q.; Goel, A. K. (2024). *SAMI: An AI Actor for Fostering Social Interactions in Online Classrooms*. In A. Sifaleras & F. Lin (Eds.), **Generative Intelligence and Intelligent Tutoring Systems** (Lecture Notes in Computer Science, vol. 14798, pp. 149–161). Cham: Springer. doi.org/10.1007/978-3-031-63028-6_12

Liao, Q.; Gruen, D.; Miller, S. (2020). *Questioning the AI: Informing Design Practices for Explainable AI User Experiences*. In **Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems** (pp. 1–15). New York, NY: Association for Computing Machinery. doi.org/10.1145/3313831.3376590

Sipos, L.; Schäfer, U.; Glinka, K.; Müller-Birn, C. (2023). *Identifying Explanation Needs of End-users: Applying and Extending the XAI Question Bank*. In **Proceedings of Mensch und Computer 2023** (pp. 492–497). New York, NY: Association for Computing Machinery. doi.org/10.1145/3603555.3608551

Wang, Q.; Jing, S.; Camacho, I.; Joyner, D.; Goel, A. (2020). *Jill Watson SA: Design and Evaluation of a Virtual Agent to Build Communities among Online Learners*. In **Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems** (pp. 1–8). New York, NY: Association for Computing Machinery. doi.org/10.1145/3334480.3382878