

Social AI Agents Too Need to Explain Themselves

Rhea Basappa, Mustafa Tekman, Hong Lu, Benjamin Faught, Sandeep Kakar,
and Ashok K. Goel

Georgia Institute of Technology, Atlanta GA 30332, USA
{rb324,mtekman3,bfaught3,skakar6,ag25}
@gatech.edu, hlu07@tufts.edu

Abstract. Social AI agents interact with members of a community, thereby changing the behavior of the community. For example, in online learning, an AI social assistant may connect learners and thereby enhance social interaction. These social AI assistants too need to explain themselves in order to enhance transparency and trust with the learners. We present a method of self-explanation that uses introspection over a self-model of an AI social assistant. The self-model is captured as a functional model that specifies how the methods of the agent use knowledge to achieve its tasks. The process of generating self-explanations uses Chain of Thought to reflect on the self-model and ChatGPT to provide explanations about its functioning. We evaluate the self-explanation of the AI social assistant for completeness and correctness.

Keywords: Social AI · XAI · Self-Explanation · Self-Models · Generative AI · Combining Knowledge-Based and Generative AI

1 Introduction

Learning at scale, and particularly online learning at scale, offers many well-established benefits such as geographically distributed and self-paced asynchronous learning that meets the reskilling and upskilling needs of working learners and learning workers. However, learning at scale, and again particularly online learning at scale, also have several well-known drawbacks such as lack of social presence, i.e., the ability of learners to establish and maintain a sense of connectedness both with one another and with the instructor [1].

SAMI (Social Agent Mediated Interaction) is an AI social assistant that helps students in large online classes form social connections by introducing them to one another based on shared characteristics and interests [2–4]. This is posited to increase social presence in an online class environment [3]. However, students interacting with SAMI often have questions regarding its inner workings [4]. Knowing how SAMI works internally may help students build trust in its recommendations. Thus, the specific research question for us in this paper becomes: How might an AI social assistant, such as SAMI, provide an explanation of its inner workings to online students?

We present a computational technique for self-explanation in SAMI. Our self-explanation technique consists of several parts. First, we view self-explanation as a process of question answering in which a user provides the AI agent input in natural language (English), the agent then introspects on its knowledge of its own reasoning and then produces an answer back to this question also in natural language (English). Second, this introspection requires the AI agent to have a self-model of its goals, knowledge, and methods. We use the Task, Method and Knowledge (TMK) framework [5–7] for representing this self-model. Third, we replace logical propositions in the traditional TMK models with short descriptions in English while still retaining their task-method-knowledge decomposition. Fourth, we conduct a similarity search on the input question and the English descriptions in TMK model to find the relevant snippets for answering the question. Fifth, we use Chain of Thought [8] to walk step-by-step over the TMK model to generate prompts into ChatGPT to produce an answer from the identified snippets. Thus, the self-explanation technique of SAMI combines the strengths of generative AI (training over a very large corpus and the ability to address a large variety of natural language tasks) with that of knowledge-based AI (knowledge representation and organization at multiple levels of abstraction).

2 Related Work

Self-explanation has re-emerged as an important topic in AI. Muller et al. (2019) [9] provide a fairly comprehensive and a very useful summary of the history of AI research on self-explanation. Confalonieri et al. (2021) [10] present another and more recent take on the history. The need for interpretability of the representations and processing in modern neural networks is one of the main reasons for the resurgence of interest in self-explanation in AI agents [11]. Rudin [12] advocates the construction and use only of AI agents capable of self-interpretation and self-explanation.

It is useful here to distinguish between two kinds of AI assistants: AI assistants that interact with individual humans and AI assistants that enable interaction among humans. In the context of AI in learning and teaching, teaching assistants such as Jill Watson [13, 14] that answer a student’s questions are an example of the former; AI social assistants such as SAMI [2–4] that help foster interactions among students are an example of the latter. The latter class of assistants exemplify the paradigm of “computers are social actors” [15]. It is important to note that self-explanation in social assistants is as important as it is in personal teaching and learning assistants.

One of the key ideas to emerge out of this early research on explanation was the importance of explicit representation of knowledge of the design of an AI assistant [16, 17]. An explicit representation of the design knowledge of an AI assistant enables the generation of explanations of the tasks it accomplishes, the domain knowledge it uses, as well as the method that use the knowledge to achieve the tasks. This raised the question of how this design knowledge can be identified, acquired, represented, stored, accessed, and used for generating

explanations [7]. One possible answer was to endow the AI agent with meta-knowledge of its own design [18] and enable the agent to generate explanations through introspection on its meta-knowledge.

3 Computational Architecture and Process for Self-Explanation

3.1 SAMI, the Social AI Agent

SAMI accesses the self-introduction posts of students in an online discussion forum and extracts information such as their location, hobbies and academic interests. Using this, SAMI builds a knowledge graph for each student. It then uses this knowledge graph to 'match' students who share one or more similarities. SAMI communicates its recommendations of matches to the online students who elect to contact the recommended matches [2–4].

3.2 Semantic Representation of SAMI

From its code base, we manually create a symbolic representation of SAMI in the Task-Method-Knowledge (TMK) framework [5–7]. Briefly, the TMK model specifies SAMI's tasks (or goals), methods (or mechanisms for achieving the goals) and the domain knowledge of the environment. The TMK is organized hierarchically. The top-level task specifies SAMI's method for accomplishing it; the method specifies the finite state machine for accomplishing the task in terms of a sequence of information states and state transitions. The state transitions are annotated by either subtasks or domain knowledge. This decomposition continues until all leaf nodes in the TMK model are primitive tasks that can be directly accomplished by the available domain knowledge. Figure 1 illustrates the hierarchical organization of the TMK model of SAMI. Figure 2 illustrates the state-transition specification of a method in the TMK model in detail.

Having built the TMK model of SAMI, we manually translate the logical propositions in the TMK model into brief natural language descriptions to obtain a semantic representation of SAMI. This semantic representation becomes the self-model of SAMI that empowers our self-explanation technique.

3.3 Self-Explanation Technique

The self-explanation technique utilizes the information from SAMI's TMK self-model to provide explanations about its inner workings. As Figure 3 illustrates, the technique has three main stages: Classification, Localization and Reasoning. When a question is asked, the Classifier first analyses the question to determine which of the pre-defined classes the question belongs to. These classes are 'mmodel', 'kmodel', 'multimodel' (along with 'can't answer') and are used to determine which part(s) of the self-model of SAMI are later used in the self-explanation pipeline. A 'kmodel' classification would lead to using information

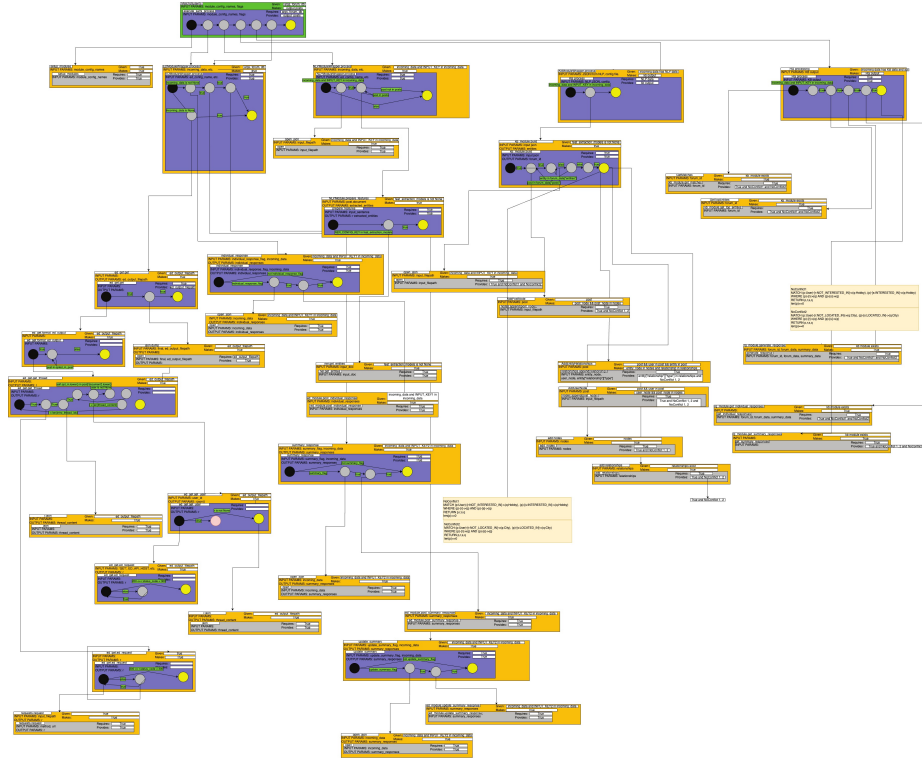


Fig. 1. The organization of the TMK model of SAMI. Outer rectangles represent high-level goals. Inner rectangles represent methods, with the circles within them representing sub-tasks and the arrows between them representing transitions. (We know that the text in this figure is not readable. Our goal here is to convey the hierarchical decomposition of the TMK model of SAMI, along with its size and complexity.

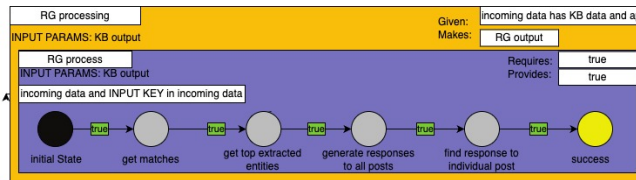


Fig. 2. This figure illustrates the state-transition specification of one method (‘RG process’) in the TMK model of SAMI. The circles within the method represent the individual tasks. The arrows connecting the tasks represent the state by state transitions within this particular method.

only from the domain knowledge part of the self-model. An ‘mmodel’ classification would lead to localizing the relevant task and method information within the self-model. With a ‘multimodel’ classification, a similarity search would be conducted to find the relevant pieces of information from all knowledge, method, and task parts of the self-model. Lastly, any question deemed as not being relevant to SAMI would be classified as ‘can’t answer’. The Classifier employs LangChain¹ to create a prompt that uses pre-written templates describing each of these classifications, along with the question to be answered. This prompt is then sent to ChatGPT², which returns a value for the classification, along with a complexity ‘k’ value which is used to control the verbosity of the final answer in later stages.

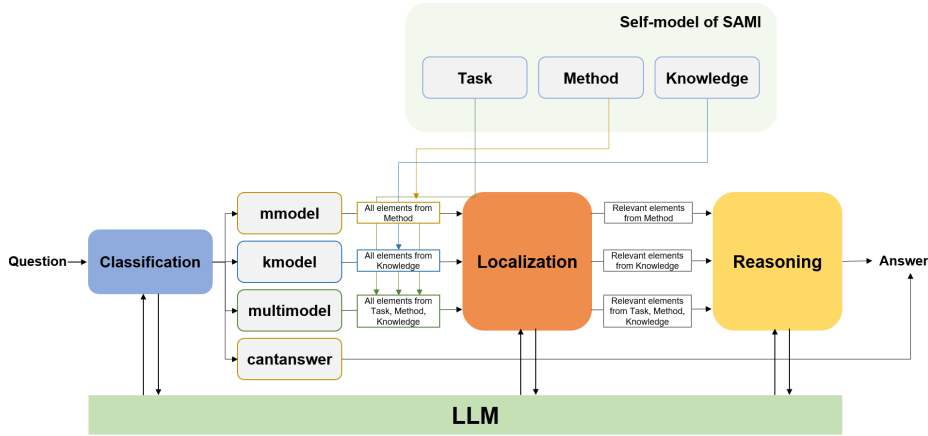


Fig. 3. The computational architecture and process flow of the self-explanation technique.

Next, the Localizer conducts the similarity search to find the most relevant k pieces of information within the sub-model(s) identified as relevant by the classifier. The Localizer uses the FAISS library³ to do a similarity search on the input question and the natural language descriptions in the relevant sub-model(s). The hierarchical organization of the TMK model (see Figure 1) helps in this localization. Like the Classifier, the Localizer employs LangChain and prompts ChatGPT to determine the names of the relevant items from the sub-model(s). These are then passed to the reasoning stage.

In the final stage, if the relevant items identified by the Localizer include a method, the Reasoner uses Chain of Thought to walk step-by-step over the

¹ LangChain documentation

² OpenAI’s gpt3.5-turbo-instruct model has been used

³ Meta’s FAISS documentation

specification of the identified method including the subtasks in the method (see Figure 2). This enables the answer to include descriptions of task annotations on the state transitions within a method that might have led to a particular outcome of SAMI. Like the Classifier and the Localizer before it, the Reasoner once again uses LangChain to construct prompts to ChatGPT to compose the final answer. Thus, the self-model technique makes multiple calls to ChatGPT for different purposes in different stages.

4 Evaluation

Correctness and Completeness Study Design: To evaluate the self-explanation technique, we used high-level, non context-dependent questions taken directly from XAI question banks [19, 20] such as “*What is the source of the data?*” [19], “*How often does the system make mistakes?*” [19] and “*What is the scope of the output data?*” [20]. Additionally, we modified some questions so that they become more relevant to SAMI. For example, “*What are the results of other people using the system?*” [20] was adapted as “*What is the result of other students opting-in to use SAMI?*”.

In total, 57 questions were borrowed and adapted from the question banks. Additionally, 9 questions specific to SAMI were created. These include questions such as “*What is a match?*” and “*How do you find matches for students?*”. The self-explanation technique of SAMI was prompted with each of these 66 questions and SAMI developers assessed each explanation for correctness and completeness. We considered the definition of correctness as “nothing but the truth” [21] and completeness as “the whole truth” [21]. For correctness, three categories - yes, partial and no - were noted, and for completeness, two categories - complete and incomplete - were used.

Result of the Correctness and Completeness Study: Table 1 summarizes the completeness and correctness scores for each question category. The self-explanation technique provided correct answers to 49 out of 66 questions; 37 of these 49 correct answers were also complete. For example, for the question, “*What is a match?*”, the self-explanation method provided the answer “*A match is a student recommended by SAMI to the user who shares one or several similarities with the user. This information is based on the task, method and goals of the Social AI agent provided, which contains information about objects and their properties that SAMI tracks in its system. The task, method and goals of the Social AI agent states that a match is a student recommended by SAMI, and this recommendation is based on shared similarities between the user and the recommended student. These similarities could include location, hobbies, specialization, classes taken, and timezone*”. Since this explanation did not contain any incorrect information nor did it miss any other relevant correct information, it was assessed to be both correct and complete.

The question “*What are the different types of responses you are capable of giving?*” was answered with “*I am capable of giving Individual Responses, which*

Table 1. Results of categorising all 66 questions that were used to evaluate the self-explanation technique, a representative question for each category along with their adaptation and corresponding completeness and correctness results

Category	# of Questions	Example Question(s)	Completeness Scores	Correctness Scores
Input	4	What kind of data does SAMI learn from? What is the source of the data?	Complete: 2/4 Incomplete: 2/4	Correct: 3/4 Partially Correct: 1/4 Incorrect: 0/4
Output	22	How can I best utilise SAMI's output? What kind of output does SAMI give?	Complete: 16/22 Incomplete: 6/22	Correct: 14/22 Partially Correct: 2/22 Incorrect: 6/22
How (global)	17	Is location used or not used in matchmaking? Are hobbies used or not used in matchmaking?	Complete: 12/17 Incomplete: 5/17	Correct: 14/17 Partially Correct: 2/17 Incorrect: 1/17
Why not	1	Why would two students not be matches?	Complete: 1/1 Incomplete: 0/1	Correct: 1/1 Partially Correct: 0/1 Incorrect: 0/1
Others	10	What are the result of other students opting-in to SAMI? How will I be affected if other students use or not use SAMI?	Complete: 10/10 Incomplete: 0/10	Correct: 9/10 Partially Correct: 1/10 Incorrect: 0/10
Others (context)	3	Who is responsible for this system? Who are the authors?	Complete: 2/3 Incomplete: 1/3	Correct: 0/3 Partially Correct: 2/3 Incorrect: 1/3
SAMI Specific Questions	9	Why am I matched with a particular student? What entities do you use in matchmaking?	Complete: 6/9 Incomplete: 3/9	Correct: 8/9 Partially Correct: 0/9 Incorrect: 1/9

are recommendations for matches based on the information provided by the user. These responses are based on the entities I extract from the user's introduction post, such as location, hobbies, specialization, classes taken, and timezone." Whilst this is correct, it is incomplete because SAMI is capable of providing individual, summary and feedback responses. Saying it only provides individual response makes the answer incomplete. 12 of the correct answers were incomplete. Some answers given by the self-explanation method were partially correct. Such partially correct answers were given to 8 of the 66 questions. The question, "What are your limitations?" received an answer as, "I am a software system called SAMI, and my main limitation is that I can only mediate social interactions among students who have opted in to use me. I do this by extracting entities such as location, hobbies, specialization, classes taken, and timezone from their introduction post and connecting them to others based on similar entities. However, I am limited to only students who have opted in and have included the #connectme optin tag in their introduction post in the ED online discussion forum. Additionally, I am limited to only students who are enrolled in one of the five specializations offered by Georgia Institute of Technology for its Computer Science Masters students." This answer is evaluated to be partially correct because even though the first part around the opt-in limitation is correct, the second part around being only limited to students in one of the five specializations is incorrect. On the other hand, the answer is complete as it's not missing any other relevant information.

Incorrect answers were given to only 9 out of the 66 questions. 6 of these 9 incorrect answers were to questions belonging to the ‘Output’ category - this could be attributed to the fact that there’s no output specified in SAMI’s self-model, since it does not have a static output. For example, “*What does the output of SAMI mean?*” received an answer as, “*The output of SAMI refers to the result or outcome of the system’s task, which is to mediate social interactions among students who have opted in to use SAMI. However, according to the given task, method and goals of the Social AI agent, the output for this specific task is not specified, so it is not possible to provide a specific answer. It is recommended to ask other questions that may be answered with the information provided in the task, method and goals of the Social AI agent.*”. This answer was evaluated to be incorrect (and incomplete) as it doesn’t correctly explain what the output of SAMI means.

In total, 48 out of 66 answers were complete (with 18 being incomplete). Questions like “*What does a match mean?*”, “*What type of matches does SAMI give?*” and “*What kind of output does SAMI give?*” received complete answers, whilst questions such as “*What data is the system not using?*” [19] and “*What is the system’s overall logic?*” [19] received incomplete explanations. We expect the completeness of SAMI’s self-explanations to improve with the completeness of its TMK self-model.

5 Conclusions

Our computational technique for self-explanation in AI social assistants combines classical knowledge-based methods with modern generative AI methods. The technique for self-explanation leverages ChatGPT to introspect over a TMK self-model of the AI social assistant to generate explanations about its functioning. Our preliminary analysis of the self-explanation technique showed that it is capable of providing complete and correct explanations about the inner workings of SAMI provided that the question asked was relevant to the TMK self-model of SAMI and the self-model itself was complete and correct. We tentatively conclude that introspection by generative AI on a self-model of the AI social assistant is a promising way of generating self-explanations and thus merits further investigation.

Acknowledgements

This research has been supported by NSF Grants #2112532 and #2247790 to the National AI Institute for Adult Learning and Online Education. We thank members of the Design & Intelligence Laboratory for their contributions to this work.

References

1. Garrison, D., Anderson, T. & Archer, W. Critical inquiry in a text-based environment: Computer conferencing in higher education. *The Internet And Higher Education*. **2**. 87–105 (1999).
2. Wang, Q., Jing, S., Camacho, I., Joyner, D. & Goel, A. Jill Watson SA: Design and evaluation of a virtual agent to build communities among online learners. *Extended Abstracts Of The 2020 CHI Conference On Human Factors In Computing Systems*. pp. 1-8 (2020)
3. Goel, A. AI-powered learning: making education accessible, affordable, and achievable. *ArXiv Preprint ArXiv:2006.01908*. (2020)
4. Kakar, S., Basappa, R., Camacho, I., Griswold, C., Houk, A., Leung, C., Tekman, M., Westervelt, P., Wang, Q., Goel, A.: SAMI: An AI Actor for Fostering Social Interactions in Online Classrooms. Accepted for publication. *Proceedings of 20th International Conference, ITS 2024, Springer, Thessaloniki, Greece*. (2024)
5. Murdock, J. & Goel, A. Meta-case-based reasoning: self-improvement through self-understanding. *Journal Of Experimental & Theoretical Artificial Intelligence*. **20**, 1-36 (2008)
6. Goel, A. & Rugaber, S. GAIA: A CAD-like environment for designing game-playing agents. *IEEE Intelligent Systems*. **32**, 60-67 (2017)
7. Goel, A., Sikka, H., Nandan, V., Lee, J., Lisle, M. & Rugaber, S. Explanation as Question Answering based on a Task Model of the Agent’s Design. *ArXiv Preprint ArXiv:2206.05030*. (2022)
8. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q., Zhou, D. & Others Chain-of-thought prompting elicits reasoning in large language models. *Advances In Neural Information Processing Systems*. **35** pp. 24824-24837 (2022)
9. Mueller, S., Hoffman, R., Clancey, W., Emrey, A. & Klein, G. Explanation in human-AI systems: A literature meta-review, synopsis of key ideas and publications, and bibliography for explainable AI. *ArXiv Preprint ArXiv:1902.01876*. (2019)
10. Confalonieri, R., Coba, L., Wagner, B. & Besold, T. A historical perspective of explainable Artificial Intelligence. *Wiley Interdisciplinary Reviews: Data Mining And Knowledge Discovery*. **11**, e1391 (2021)
11. Gilpin, L., Bau, D., Yuan, B., Bajwa, A., Specter, M. & Kagal, L. Explaining explanations: An overview of interpretability of machine learning. *2018 IEEE 5th International Conference On Data Science And Advanced Analytics (DSAA)*. pp. 80-89 (2018)
12. Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*. **1**, 206-215 (2019)
13. Goel, A. & Polepeddi, L. Jill Watson: A Virtual Teaching Assistant. *Learning Engineering For Online Education: Theoretical Contexts And Design-based Examples*. Routledge. (2018)
14. Eicher, B., Polepeddi, L. & Goel, A. Jill Watson doesn’t care if you’re pregnant: Grounding AI ethics in empirical studies. *Proceedings Of The 2018 AAAI/ACM Conference On AI, Ethics, And Society*. pp. 88-94 (2018)
15. Lee, J. & Nass, C. Trust in computers: The computers-are-social-actors (CASA) paradigm and trustworthiness perception in human-computer communication. *Trust And Technology In A Ubiquitous Modern Environment: Theoretical And Methodological Perspectives*. pp. 1-15 (2010)

16. Chandrasekaran, B., Tanner, M. & Josephson, J. Explaining control strategies in problem solving. *IEEE Intelligent Systems*. **4**, 9-15 (1989)
17. Chandrasekaran, B. & Swartout, W. Explanations in knowledge systems: the role of explicit representation of design knowledge. *IEEE Expert*. **6**, 47-49 (1991)
18. Goel, A., Silver Garza, A., Grué, N., Murdock, J., Recker, M. & Govindaraj, T. Explanatory interface in interactive design environments. *Artificial Intelligence In Design'96*. pp. 387-405 (1996)
19. Liao, Q., Gruen, D. & Miller, S. Questioning the AI: informing design practices for explainable AI user experiences. *Proceedings Of The 2020 CHI Conference On Human Factors In Computing Systems*. pp. 1-15 (2020)
20. Sipos, L., Schäfer, U., Glinka, K. & Müller-Birn, C. Identifying Explanation Needs of End-users: Applying and Extending the XAI Question Bank. *Proceedings Of Mensch Und Computer 2023*. pp. 492-497 (2023)
21. Nauta, M., Trienes, J., Pathak, S., Nguyen, E., Peters, M., Schmitt, Y., Schlötterer, J., Keulen, M. & Seifert, C. From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable AI. *ACM Computing Surveys*. **55**, 1-42 (2023)