

Evaluating What Others Say: The Effect of Accuracy Assessment in Shaping Mental Models of AI Systems

HYO JIN DO, IBM Research, USA

MICHELLE BRACHMAN, IBM Research, USA

CASEY DUGAN, IBM Research, USA

QIAN PAN, IBM Research, USA

PRIYANSHU RAI, IBM Research, India

JAMES M. JOHNSON, IBM Research, USA

ROSHNI THAWANI, IBM Canada Ltd., Canada

Forming accurate mental models that align with the actual behavior of an AI system is critical for successful user experience and interactions. One way to develop mental models is through information shared by other users. However, this social information can be inaccurate and there is a lack of research examining whether inaccurate social information influences the development of accurate mental models. To address this gap, our study investigates the impact of social information accuracy on mental models, as well as whether prompting users to validate the social information can mitigate the impact. We conducted a between-subject experiment with 39 crowdworkers where each participant interacted with our AI system that automates a workflow given a natural language sentence. We compared participants' mental models between those exposed to social information of how the AI system worked, both correct and incorrect, versus those who formed mental models through their own usage of the system. Specifically, we designed three experimental conditions: 1) *validation condition* that presented the social information followed by an opportunity to validate its accuracy through testing example utterances, 2) *social information condition* that presented the social information only, without the validation opportunity, and 3) *control condition* that allowed users to interact with the system without any social information. Our results revealed that the inclusion of the validation process had a positive impact on the development of accurate mental models, especially around the knowledge distribution aspect of mental models. Furthermore, participants were more willing to share comments with others when they had the chance to validate the social information. The impact of inaccurate social information on altering user mental models was found to be non-significant, while 69.23% of participants incorrectly judged the social information accuracy at least once. We discuss the implications of these findings for designing tools that support the validation of social information and thereby improve human-AI interactions.

CCS Concepts: • **Human-centered computing** → **HCI design and evaluation methods**.

Additional Key Words and Phrases: Mental Model; Social Information; Natural Language Interface; Validation

ACM Reference Format:

Hyo Jin Do, Michelle Brachman, Casey Dugan, Qian Pan, Priyanshu Rai, James M. Johnson, and Roshni Thawani. 2024. Evaluating What Others Say: The Effect of Accuracy Assessment in Shaping Mental Models of AI Systems. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW2, Article 373 (November 2024), 26 pages. <https://doi.org/10.1145/3686912>

Authors' Contact Information: Hyo Jin Do, hjdo@ibm.com, IBM Research, Cambridge, MA, USA; Michelle Brachman, michelle.brachman@ibm.com, IBM Research, Cambridge, MA, USA; Casey Dugan, cadugan@us.ibm.com, IBM Research, Cambridge, MA, USA; Qian Pan, qian.pan@ibm.com, IBM Research, Cambridge, MA, USA; Priyanshu Rai, priyanshu.rai@ibm.com, IBM Research, Pune, India; James M. Johnson, jmjohnson@us.ibm.com, IBM Research, Cambridge, MA, USA; Roshni Thawani, roshni.t@ibm.com, IBM Canada Ltd., Toronto, Ontario, Canada.



This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs International 4.0 License.

© 2024 Copyright held by the owner/author(s).

ACM 2573-0142/2024/11-ART373

<https://doi.org/10.1145/3686912>

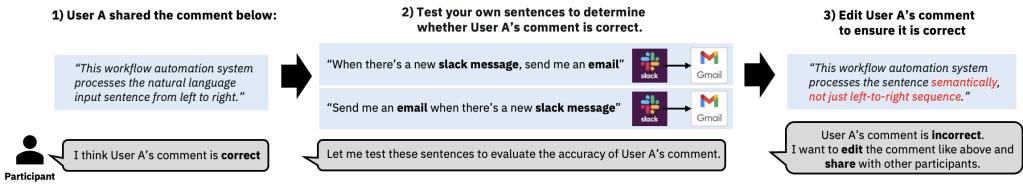


Fig. 1. In the validation condition, we designed a main task containing a series of steps: 1) a participant evaluated the accuracy of a given piece of social information (e.g., User A's comment); 2) validated the social information accuracy by testing various sample utterances; 3) made a final judgment about the social information accuracy, edited the given social information, and rated how much they were willing to share their revised comment.

1 Introduction

With abundant social information streams available, including social media, online communities, and blogs, people are more widely and actively sharing how they use, interact with, and understand AI systems, which essentially represent their mental models. There is evidence that others consuming this social information, or shared mental models, are influenced by it through 'social learning' [32, 63, 70]. Forming accurate mental models that align with the actual system behavior is critical to improving user experience and task performance [25, 27]. Social information, on the other hand, can be inaccurate, due to various reasons, including limited experiences or outdated sources [10, 13]. Consequently, there exists a gap in understanding how this inaccurate social information impacts individuals' mental models.

Previously, there has been research conducted on two important topics: identifying social influence in understanding and using AI systems [63, 65], and investigating ways to foster accurate user mental models [27, 51]. Our research aims to integrate these two research streams by examining the impact of *social information* accuracy in shaping accurate mental models of an AI system. Additionally, whether and how we can support users in *validating* the accuracy of shared mental models is still an open question. There has been an emergence of tools that aid users in interpreting how their input influenced the system output [7, 37, 49, 66] and HCI research understanding their impact [40, 47, 53, 67]. However, no prior research has explored how validating the accuracy of information shared by others affects user mental models. Building on these studies, we investigated the impact of validating social information on user mental models and identify strategies people used for validation.

In this study, we focused on an AI system that generates a task workflow given a natural language utterance. We used a dataset from a prior study that collected social information which describes correct or incorrect shared mental models about how the system works. We conducted a between-subject experiment involving 39 crowdworkers who were randomly assigned to one of the three conditions: 1) a validation condition, where participants received and evaluated social information followed by an opportunity to validate its accuracy through testing various utterances; 2) a social information condition, where participants received and evaluated social information but without the validation opportunity; and 3) a control condition, where participants wrote about their own mental models after completing a goal-oriented task, without receiving social information or having the validation opportunity. For further elaboration, we outline the main tasks in the validation condition in Fig. 1, with additional details provided in Fig. 3.

Using survey responses, we measured participants' mental models of the system in three dimensions [27], global behavior (how the system works overall), local behavior (how the system makes

an individual decision), and knowledge distribution (what the system knows and how it uses the knowledge) types. We also measured their performance on the final task and their editing strategies of the social information through system logs. To preview our key results, we found that 69.23% of participants in treatment conditions incorrectly judged the social information accuracy at least once. When participants were given a chance to validate the social information accuracy, their mental models improved significantly and boosted their confidence in their accuracy judgments. Participants also became more willing to share the comments they had collaboratively written after validation, compared to those who didn't get the validation experience. Overall, our results highlight the importance of empowering individuals to critically and empirically evaluate social information to foster more accurate mental models of AI systems with confidence. This, in turn, enables sharing of accurate social information, benefiting the broader user community as a result.

Our contributions are threefold: first, our empirical findings show that validating social information improves mental models, especially in knowledge distribution type. Second, we found that inaccurate social information had a non-significant impact on mental models. Third, we identified various strategies participants used to collaboratively build shared mental models. We offer design implications on how to assist people to develop and share accurate mental models, and thereby improve the quality of social information and human-AI interactions.

2 Related Work

Our research was primarily informed by prior work in three different topical areas: mental models in human-AI interactions, social information around AI systems, and natural language automation systems.

2.1 Mental Models in Human-AI Interactions

A mental model is defined as a user's *knowledge of the components of a system, their interconnection, and the processes that change the components, the knowledge that forms the basis for users being able to construct reasonable actions, and explanations about why a set of actions is appropriate* [18]. A mental model is different from the conceptual model [55] which is the scientist's or developer's understanding of the system. Gero et al. identified three key components in a conceptual model that are global behavior, local behavior, and knowledge distribution types [27]. An accurate mental model that aligns with the conceptual model of an AI system improves user experience [41], decision-making [9], and task performance [18, 27].

However, misaligned mental models are common [13, 42] and can be hard to change [23, 73]. Norman stated that mental models are usually inaccurate in a number of ways, including contradictory, erroneous, and unnecessary concepts [56]. For example, research has shown that people believe AI agents will perform better than humans [36], including themselves [62]. As users' perceptions change over time [71], with every interaction, users must overcome the evaluation and execution gulfs to create accurate mental models. Further, prior work showed people form mental models of AI systems without personally having interacted with them such as forming mental models about robots seen in the media, without necessarily having interacted with them [8]. Jakesch et al. showed that user mental models about large language models are hindered by flawed heuristics of discerning AI-generated texts and human-generated texts [34], which can lead to risks including over-reliance of misleading information [72]. A mismatch between the human's mental model and the true error boundary can lead to sub-optimal decisions such as: the human trusts the AI even when it makes an erroneous output and the human does not trust the AI even when it makes a correct output. These decisions can lower productivity and/or accuracy [9], which in turn result in frustration and abandonment of the technology.

Researchers have worked to support users in building accurate mental models with explainable AI (XAI) and transparency in AI systems as surveyed in [1, 6, 19, 29]. For example, Lim et al. found that explanations describing why the system behaved a certain way led to better understanding of the system [45]. In a visual question-answering context, showing counterfactual examples effectively supported users' mental models [3]. Nourani et al. suggested that designers of intelligent systems provide guidance for users, as un-directed use of a system can lead to erroneous mental models [57]. Among different types of explanations, many XAI studies have shown the benefits of 'example-based explanation', which shows AI predictions for different input examples [16, 21]. Social transparency in which the system provides insights into how other users interacted with the system have found to be also effective in understanding the AI systems [13].

Transparency approaches privileges 'seeing without knowing' in which seeing inside an AI system does not necessarily mean understanding its behavior [4]. To address this, researchers have argued that dynamically interacting with the system rather than observing is critical to truly understand how systems behave [61]. Many HCI and CSCW researchers have supported this idea and proposed related concepts such as 'interactive explanations' or 'explanatory debugging', which the system explains the reasons for its outputs to the end-user, who can then correct and refine inputs back to the system [40, 47, 53, 67]. For example, Liu et al. tested an interactive interface that enabled users to experiment with counterfactual examples of a given instance and observe changes in AI predictions, which led to an increase in the human perception of the usefulness of AI assistance [47]. Many tools that support users this interactive analysis of AI models have been developed [7, 37, 49, 66]. Motivated by these prior works, we leveraged the concept of testing different input examples of an AI system as a means of enhancing user mental models of AI systems.

2.2 Social Information around AI Systems

People are constantly exposed to other people's experiences and knowledge of AI systems in a variety of ways, such as through social media, news sources, other internet sources, and everyday life. Research has shown that people are influenced by mental models that are from other users, i.e. social information [28, 32, 55, 70]. Another related concept is the Theory of Mind [11], or the idea that humans try to make sense of each other's mental processing. Therefore, social information is essentially a shared mental model that evolves through collaboration with other users [5]. If multiple individuals (e.g., team members) have a shared mental model of their shared task and of each other, then they are able to accurately predict others' needs and behaviors and thereby increase overall performance.

However, social information can be inaccurate or incomplete, due to biases, outdated experiences, limited numbers of interactions, and differences in memory and computation limitations amongst individuals [9, 10, 44]. This inaccurate social information can inhibit a user from developing accurate mental models. Our study investigates the impact of inaccurate social information, as well as whether validation can help users improve their mental models. While prior research has attended to the effects of inaccurate social information from a wide range of domains and angles (e.g., news media [48, 58], general knowledge [52]), there is a significant gap in research exploring the impact of such information on user mental models of AI systems, as well as approaches for mitigating any adverse effects. Our study aims to bridge this gap and suggest a *validation* approach, which empowers users to test different input examples to determine the accuracy of social information.

2.3 Natural Language Systems for Workflow Automation

Our work focuses on a goal-oriented natural language interface that enables users to automate a workflow by integrating data and business applications. Many systems for this purpose have

been developed (e.g., IBM App Connect¹, Zapier², IFTTT³, Microsoft Power Automate⁴, Google AppSheet⁵) and they share a common goal: supporting a low- or no-code interface that a user with little or no technical expertise can create trigger-action or rule-based programs easily. Natural language interfaces have been increasingly developed for this task automation purpose [30, 33, 60] and other similar programming tasks [43, 75], database query processing [2], IoT and mashup configurations [46], data visualizations [64], and web page designs [38]. Natural language interfaces are intuitive, natural, and easier to learn compared to traditional interfaces that require users to use programming languages and software protocols to operate the system [31, 74].

One caveat of natural language interfaces is that it is difficult for the users to formulate accurate mental models such as what types of input sentences the system can reliably parse [35]. One reason is that a natural language system is usually a complex system with many interacting rules and structures. Natural language systems can generate very different outputs with even slight variations in the input. There are almost infinite ways to write a natural language utterance compared to the limited number of interactions a user uses to test the system and develop their mental models. In comparison to GUIs, natural language systems often lack visual signifiers to understand system capabilities. Our work tries to address this problem by giving a user the opportunity to validate the mental models of others through testing sample utterances.

There is a lack of empirical research to incorporate the validation experience in a natural language system and evaluate its impact on building accurate mental models. Through this user study, we provide empirical knowledge on how people validate social information by experimenting with different natural language utterances. Prior work also identified repair strategies [35], which describes how participants repaired their utterances to produce the desired outputs. In our study, we identify strategies participants used for validation and discuss similarities and differences with the repair strategies.

3 Context

3.1 Goal Oriented Flow Assistant (GOFA)

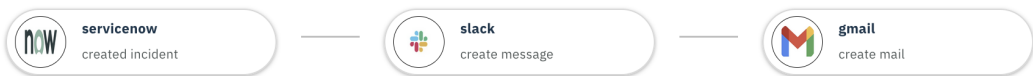


Fig. 2. An example trigger-action program represented by a flow diagram. A natural language utterance “when there’s a new incident on ServiceNow, send a message on Slack, and send an email on Gmail” will generate the program.

This research was conducted in the context of a natural language system that automates trigger-action workflows, titled Goal Oriented Flow Assistant (GOFA). When a user writes a trigger-action program in a natural language utterance, the GOFA system identifies applications (e.g., Gmail), trigger events, actions (e.g., create), and objects (e.g., mail). Then, it creates a short program that is represented by a flow diagram. Fig. 2 illustrates an example flow diagram generated by an utterance ‘when there’s a new incident on ServiceNow, send a message on Slack, and send an email on Gmail’. The system identifies ServiceNow, Slack, and Gmail as applications, a new incident as a

¹<https://www.ibm.com/cloud/app-connect>

²<https://zapier.com>

³<https://ifttt.com/>

⁴<https://powerautomate.microsoft.com/digital-process-automation/>

⁵<https://about.appsheet.com/home/>

trigger event, send as an action, and a message and an email as objects. Technical details of the system can be found in a prior work [12]. We describe example behaviors of GOFA using the following conceptual model [27]. We selected this model because it was flexible and adaptable to other AI systems, including ours, and was similar to well-known XAI types (e.g., global and local explanations), making it easier to understand and base on XAI research.

Global behavior (how the system works such as how it parses utterances and composes a flow in general).

- Divide and conquer: The system splits and parses the utterance into small parts
- Part of speech: The system looks for and uses certain parts of speech (e.g., noun, verb, preposition) to generate the flow.
- Ordering: The ordering of the generated flow components corresponds to the proper semantics of the utterance rather than left-to-right.

Local behavior (how the system makes an individual decision or correct/incorrect workflow within a single interaction).

- Structuring words: The system knows certain words that are used to create the structure of the flow (e.g., comma, 'when')
- Near each other: The application name, object, and operation need to be placed near each other to be recognized as one component in the diagram but can have filler words in between.
- Simple input: The system works best when the right amount of information is given without extraneous details to generate a correct flow.
- Writing style: Active or passive voice should not matter as long as they convey the same meaning. Same with tenses.

Knowledge distribution (what knowledge the system has access to and how the system uses that knowledge).

- Keyword matching: The system attempts to match keywords in the input utterance to a set of applications, operations, and object names. It requires at least one application in the utterance.
- Knowledge graph: The system has knowledge about applications (e.g., Slack), objects that those applications can handle (e.g., message), and operations allowed on those objects (e.g., send).
- Substitution: The system substitutes an application if there's a mismatch between an application that the user wrote and the knowledge base.

3.2 Social Information

The social information we used was randomly sampled from a dataset collected in a prior study [13]. In this prior study, researchers ran a study on MTurk where 252 participants used the GOFA system to generate trigger-action workflows and answered open-ended survey questions about their mental models such as how they think the system translated their text input to the flow diagram, and any words or structures they used in their input sentences that they believe helped them to successfully generate goal flows. Researchers categorized mental models into several themes, which can be mapped to global behavior, local behavior, and knowledge distribution aspects of GOFA. We aimed to capture this diverse range of mental models in our dataset of social information by randomly selecting at least one piece of information from each theme identified in this prior work. As a result, we selected 12 pieces of social information, each corresponding to one of the conceptual model types listed in section 3.1. The full list of social information we used in this experiment is reported

in Appendix A. To improve readability, we fixed grammar mistakes, clarified ambiguous words (e.g., pronouns), or shortened comments that were too long, without changing their meaning.

Each piece of social information was labeled as correct or incorrect. The accuracy of each piece of social information was cross-checked with the system engineers. Another researcher who did not participate in the information sampling process independently annotated the information into one of the mental model types after learning their definitions and achieved a perfect agreement except for one, in which we reached a consensus after a discussion.

4 Formative Study

We conducted a formative study to understand how participants perceive and interact with our system. We recruited 6 individuals from within an international enterprise with diverse roles and locations. Participants completed the user study followed by a semi-structured interview. During the interview, we asked questions about their feedback about the study in general, including tasks and our interface designs. They were asked to think aloud during the tasks. The study including the interview took 66.5 minutes on average ($SD = 18.42$).

Based on the participants' feedback, we iterated on the design of the experiment, survey questions, and interface, such as clarifying the instructions, replacing ambiguous social information, and controlling for timing. Participants wanted more context information in addition to the social information, thus we included the goal flows for the tasks prior participants were working on while writing each piece of social information (e.g., User A's goal flow and User A's comment in Fig. 4). We initially provided five attempts per task but reduced it to three attempts based on participants' feedback, except for the onboarding task in which participants preferred to have all five attempts.

5 Method

In this study, we planned to investigate the impact of accurate and inaccurate social information on mental models. We also aimed to explore the effects of the validation experience where people get opportunities to validate social information accuracy by testing various utterances. Specifically, we had four research questions (RQs):

- **RQ1. Validation:** How does the validation experience impact users' mental models?
- **RQ2. Social information:** How does the accuracy of social information influence users' mental models?
- **RQ3. Task performance:** How does the validation experience impact users' task performance?
- **RQ4. Collaborative comment:** How do users collaborate in constructing and sharing social information?

To answer these research questions, we designed and conducted a between-subject experiment. We designed two treatment conditions, the *validation* condition showed social information followed by the opportunity to validate the information, and the *social information* condition showed social information without the validation experience. Additionally, we designed a *control* condition that does not present social information, eliminating the need for any validation process as well. To avoid the straw man fallacy, which is oversimplifying the control condition to make it easier to find a significant effect, we included an interaction task instead of the validation step where a participant submitted natural language utterances to generate a given goal flow.

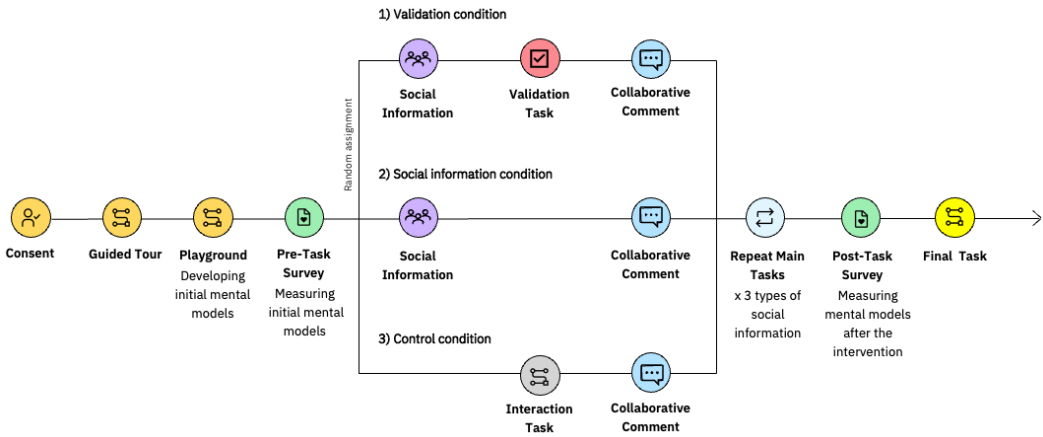


Fig. 3. End-to-end user study flow. Following an onboarding tutorial and a pre-task survey, participants were randomly assigned to one of three experimental conditions for the main tasks: validation, social information, and control conditions. One main task consisted of a series of steps, in which each condition had all or some of the steps. In the validation condition, for example, a participant were given a piece of social information with the corresponding workflow and made an initial judgment about whether the social information is correct or not (social information), validated the social information by testing various sample utterances and made a final judgment (validation task), and revised the given social information to make it more accurate (collaborative comment). Participants completed three main tasks in total, with each showing a unique piece of social information (repeat main tasks). Upon completion, they completed a post-task survey and a final task.

5.1 Study Procedures

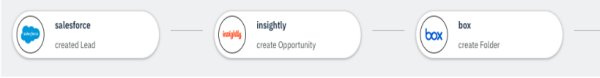
We conducted an online experiment that lasted approximately 30 minutes. The overall study flow is illustrated in Fig. 3. We describe the procedure for the validation condition below, as the social information and control conditions are small variations of the validation condition.

- (1) **Consent form:** A participant entered our website and signed a consent form. In the consent form, we explained that they would be interacting with a natural language system to generate small programs in English. We explained that they did not need to have any experience with programming to participate because all of their responses should be in natural language.
- (2) **Onboarding:** A participant read a guided tour that explained how to interact with our natural language system. Then, they were given a playground interface where they can familiarize themselves with the tool by completing an onboarding interaction task. In this task, a participant was asked to write a sentence that generates a specific visual goal flow within 5 chances.
- (3) **Pre-task survey:** They completed a pre-task survey that included questions about their initial mental models of the system.
- (4) **Main tasks:** A participant completed three main tasks. In each main task, the participant was presented with one piece of social information (e.g., ‘user A’s comment’ in Fig. 4) with an associated goal flow (e.g., ‘user A’s goal flow’), randomly chosen from the dataset. Within a main task, they were asked to complete the following steps (Fig. 4):

Task 1

Previously, a participant (anonymized as User A, B, or C) tried to construct the following goal flow using our system. They subsequently shared a comment about the system behavior such as how they believe the system parsed their sentence, and which specific words or structures were helpful in generating a correct flow.

User A's goal flow:



User A's comment:

The system took the order of the sentence and syntax to create a logical flow from left to right.

Step 1: Initial judgment

What do you think about the comment above?

I think User A's comment is:

- Correct (correctly describing the general behavior of the system)
 Incorrect (incorrectly describing the general behavior of the system)

How confident are you in your judgment?

Not confident at all 1 2 3 4 5 6 7 Extremely confident

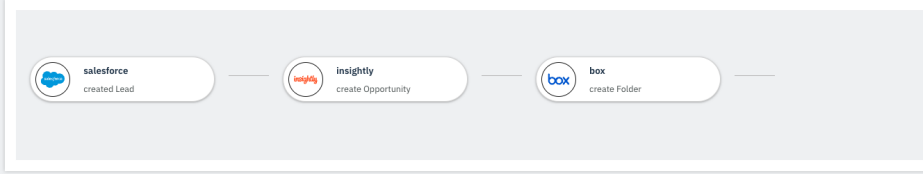
Step 2: Testing User A's comment: correct or incorrect?

Generate 3 different sample flows to test whether User A's comment is correct or not about how our system generally works. This step aims to improve your initial judgement of User A's comment by testing you own sentences in a way that evaluates the accuracy of the comment. Your work will be rejected if you fail to make an honest or reasonable effort to complete the tasks as instructed.

Example: When there is a new email on Gmail, send me a Slack message

When a lead is created in Salesforce, an opportunity is created in Insightly and a folder is created in Box

Submit



Step 3: Final judgment

You have checked the accuracy of the comment. What do you think about the comment now?

I think User A's comment is:

- Correct (correctly describing the general behavior of the system)
 Incorrect (incorrectly describing the general behavior of the system)

How confident are you in your judgment?

Not confident at all 1 2 3 4 5 6 7 Extremely confident

Step 4: Comment

1. Edit User A's comment to ensure it is accurate and valuable for other participants. Researchers will review your edits so please try to make it substantially better than the original comment.

The system took the order of the sentence and syntax to create a logical flow from left to right.

2. Are you willing to share your edited comment above to other participants so that they have more accurate knowledge?

Impossible to share 1 2 3 4 5 6 7 Extremely likely to share

Fig. 4. An abbreviated example of our task interface in the validation condition. For each task, a participant reads a piece of social information (e.g., User A's comment) randomly selected by the system. In the subsequent steps, the participant makes an initial judgment about whether they think the information is accurate or not (step 1), validates the social information by testing different sample flows (step 2), makes a final judgment about the accuracy (step 3), and revises the given social information to make it more accurate (step 4).

- **Step 1. Initial accuracy judgment:** The participant assessed the accuracy of the social information, determining whether they think it correctly or incorrectly described the system's behavior as well as indicating their confidence level regarding their choice.
 - **Step 2. Validation:** The participant generated three different sample flows to test the accuracy of the provided social information, which does not necessarily need to be the same as the given goal flow. For example, a participant may generate a flow that changes the order of the applications to validate social information about ordering. They had to use up all three attempts before proceeding to the next task to avoid a situation where they skip this step.
 - **Step 3. Final accuracy judgment:** We asked the same set of questions as in step 1 to measure any changes in their accuracy judgments after the validation experience.
 - **Step 4. Collaborative comment:** The participant edited the social information in a way that's more accurate and valuable for other users and rated how much they were willing to share their comments. They were unable to proceed if no change was made.
- (5) **Post-task survey:** After completing three main tasks, they submitted a post-task survey including questions about their mental models of the system. We asked an open-ended question about their strategies of validation and repair in the validation and control condition, respectively.
 - (6) **Final interaction task:** the participant completed a final task where they were asked to write an utterance to generate a given goal flow within 3 attempts. We offered a small bonus depending on their task success to increase participants' motivation.

In the social information condition, we excluded validation-related steps (steps 2 and 3) in the main task. This allowed us to assess the effects of the validation experience by comparing the post-task survey results between the validation and the social information conditions. We designed the control condition to represent a condition without social information, allowing us to investigate the effects of social information by comparing treatment conditions and the control condition. Therefore, we excluded steps 1, 2, and 3 in the main tasks for the control condition because there's no social information to judge accuracy or validate. Instead, the control condition had an interaction task in which they were instructed to generate a given goal flow, which were the same workflow presented in other conditions with the social information. This allowed us to avoid the straw man fallacy, which is oversimplifying the control condition compared to treatment conditions. Then, they wrote their own comments to share as in step 4. All three conditions repeated the main task three times with randomized order, each task containing a unique social information (in treatment conditions) or a workflow (in control condition).

We employed a pre-post study design, utilizing pre- and post-task surveys, to analyze the effects of the interventions implemented in the main tasks, while mitigating the influence of confounding factors arising from participants' prior experience with similar AI or automation tools. To ensure effective onboarding, we offered playground interactions, as measuring initial mental models without such interactions would not make sense. We compensated up to \$5, which consists of a base payment of \$4.7 (i.e., \$9.4/hr) to all participants who completed the study, and a bonus (\$0.3) to participants who completed the final task successfully.

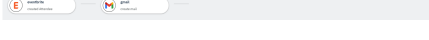



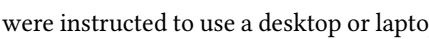
5.2 Participants

We recruited crowdworkers from Amazon Mechanical Turk (MTurk). Participants were eligible if they were 18 years old or older, live in the US, have English as their primary language, and have not participated in our study before. In order to receive quality responses, we invited participants whose number of approved tasks was greater than 1000, whose HIT approval rate was greater than 98, and

Table 1. We found no significant difference in education and AI-related experience of participants across conditions. The number of participants who were associated with each condition is reported. The same goal flows were used across all conditions.

Factors	Range	Validation	Social Information	Control
Education	High school degree or equivalent	0	0	1
	Some college but no degree	2	1	2
	Associates degree	2	1	0
	Bachelors degree	9	8	8
	Graduate degree	2	1	2
AI experience	I have heard about AI in the news, friends, or family	6	7	6
	I closely follow AI-related news	5	2	3
	I have some work experience and/or formal education related to AI	4	2	3
	I have significant work experience related to AI	0	0	1

Table 2. Task designs. We counterbalanced the ordering of mental model types associated with each main task. The goal flows remained consistent across conditions.

Task	Instruction (for the validation condition)	Goal flow
Onboarding interaction task	Imagine you're hosting an event and attendees are reserving their spot through Eventbrite, an event management app. You want to receive email notifications on Gmail whenever someone new registers for the event. This scenario is described by the goal flow. Please enter a sentence in plain English to generate the goal flow.	
Main task 1 (Global behavior)	Previously, a participant tried to construct the following goal flow using our system. They subsequently shared a comment about the system behavior such as how they believe the system parsed their sentence, and which specific words or structures were helpful in generating a correct flow. (followed by social information)	
Main task 2 (Local behavior)	Previously, a participant tried to construct the following goal flow using our system. They subsequently shared a comment about the system behavior such as how they believe the system parsed their sentence, and which specific words or structures were helpful in generating a correct flow. (followed by social information)	
Main task 3 (Knowledge distribution)	Previously, a participant tried to construct the following goal flow using our system. They subsequently shared a comment about the system behavior such as how they believe the system parsed their sentence, and which specific words or structures were helpful in generating a correct flow. (followed by social information)	
Final interaction task	Please enter a sentence in plain English to generate the goal flow.	

who passed our English proficiency test [20]. Participants were instructed to use a desktop or laptop computer for the experiment. We analyzed data from 39 participants in total, which consists of 15 participants in the validation condition, 11 participants in the social information condition, and 13 participants in the control condition, guided by [17]. The number of participants were slightly different across conditions because we filtered out unqualified participants after recruitment, such as those who failed the attention-check questions in the survey or clearly didn't follow the instructions. No significant differences in educational background and AI-related experience were found among the participants across conditions, as we reported in Table 1.

5.3 Tasks

As shown in Table 2, the study consists of an onboarding task, three main tasks, and a final task, each with a goal flow. In the treatment conditions, three pieces of social information were randomly selected from our dataset (see Appendix A), one from each global behavior, local behavior, and knowledge distribution type. For example, the first main task introduced one social information that explains the global behavior of the system, the second main task introduced one social information that explains the local behavior of the system, and the third main task introduced one social information that explains a knowledge distribution behavior of the system. The ordering of the type was counterbalanced.

To provide more context, as was requested in the formative study, we also provided the corresponding goal flow the social information provider was working on in the prior study [13]. Due to the random selection, a participant could read either accurate or inaccurate social information in each task, resulting in a possible range of 0 to 3 encounters of accurate social information per participant. In the control condition, participants were instructed to generate the same goal flows.

5.4 Measures

We used survey questions to measure participants' mental models, accuracy judgments, collaboration to build shared mental models, and strategies to generate workflows. We analyzed system logs to investigate task performance and collect sentences they submitted to generate workflows.

5.4.1 Mental models. A participant rated their agreements on a 7-point Likert scale on ten statements that describe system behavior correctly or incorrectly [13]. We grouped the questions into three types, global behavior, local behavior, and knowledge distribution types, based on the definitions described in a prior work [27]. For example, a participant rated their agreement on the following statement, 'all of the keywords that appear in the goal flow must be in the sentence using the exact same wordings to generate the correct flow.' This statement incorrectly describes the system behavior and was associated with the keyword matching type. We reversed participants' ratings of incorrect statements (i.e. reverse coding) to ensure that higher values consistently indicate better mental models of the system behavior. We calculated the average to assess the overall mental model scores, as well as the scores for each type. A full list of statements is listed in Appendix B.

5.4.2 Accuracy judgments and confidence. As a supplementary measure to explain participants' mental model changes, we measured whether participants felt the social information was correct or incorrect immediately after reading the social information (e.g. Step 1 initial judgment in Fig. 4) in the treatment conditions, as well as after the validation experience (e.g. Step 3 final judgment in Fig. 4) in the validation condition. Along with each judgment, we also asked their confidence level of their judgments on a 7-point Likert scale. Note that we didn't reveal the actual correctness of social information to the participants.

5.4.3 Collaborative comment and sharing intent. A participant edited the given social information from other users in treatment conditions or wrote their own comment in the control condition, namely *collaborative comment*. Through their collaborative comments, we aimed to investigate how people collaborate to write and refine social information and thereby construct shared mental models. After writing a comment, the participant rated how much they were willing to share their comments on a 7-point Likert scale.

5.4.4 Validation and repair strategies. For participants in the validation condition, we asked an open-ended question about their validation strategy in the post-task survey: 'What strategies did you use to generate different flows for testing the accuracy of other users' comments?'. For participants in the control condition, we asked an open-ended question about their repair strategy in the post-task survey: 'What strategies did you use to generate a correct flow?'.

5.4.5 Task performance. We measured participants' performance in the final task using two measures.

- **Number of attempts:** The number of attempts to complete the final task, either until they achieved the correct flow or until they reached the maximum limit of 3 attempts. For instance, if they failed to generate a correct flow, the number of attempts recorded would be 3.
- **Success rate:** We calculated whether a participant generated a correct flow in the final task (success: 1, fail: 0) divided by the number of attempts. This will take into account how successful and efficient the participant has performed in the final task.

5.5 Analysis

5.5.1 Statistical analysis. When analyzing the effect of experimental conditions on the final *task performance*, we built a linear model with the experimental condition as an independent variable and task performance as a dependent variable. When analyzing the effect of the conditions on

mental model scores, we used the mental model score computed from the post-task survey scores as a dependent variable. We then constructed a linear model with the condition as the independent variable and the pre-task mental model score as a covariate to adjust for their initial state. To analyze the effect of *social information accuracy* on mental model scores in treatment conditions, we constructed linear mixed models using the post-task mental model score as a dependent variable, the number of tasks that contained accurate social information as a fixed-effect factor (ranges 0-3), the pre-task mental model score as a covariate, and the experimental condition as a random-effect factor. When analyzing the effects of other variables that were measured after each task (e.g., *accuracy judgment success, confidence, sharing intent*), we used linear mixed models with the condition as a fixed-effect factor. The task ID and Participant ID (PID) were random-effect factors to account for intraclass correlation (i.e., between-cluster variance to the total variance controlled by random intercept models [59]). In cases where we found a significant or marginally significant effect, we conducted post hoc tests with Bonferroni corrections. We summarized statistical tests in Table 3.

Table 3. Statistical tests summary. We summarized primary statistical tests we conducted, listing variables (DV: dependent variable, IV: independent variable, covariate, fixed/random-effect factors), model (LM: linear model, LMM: linear mixed-effect model), analysis type (between/within-subjects analysis), data used, corresponding research questions and sections.

RQ	Section	DV	IV/Fixed	Covariate	Random	Model	Analysis	Data used
1	6.1.1	Post-task mental model	Condition	Pre-task mental model	N/A	LM	Between	All conditions
1	6.1.2	Judgment success	Time of judgment (step 1/3)	N/A	Task ID, PID	LMM	Within	Validation condition
1	6.1.2	Judgment confidence	Time of judgment (step 1/3)	N/A	Task ID, PID	LMM	Within	Validation condition
2	6.2	Post-task mental model	Num of accurate social info	Pre-task mental model	Condition	LMM	Within	Validation, Social information
3	6.3	Num of attempts	Condition	N/A	N/A	LM	Between	All conditions
3	6.3	Success rate	Condition	N/A	N/A	LM	Between	All conditions
4	6.4	Sharing intent	Condition	N/A	Task ID, PID	LMM	Between	All conditions

5.5.2 Utterance analysis. Two researchers conducted the thematic analysis [14] to analyze utterances submitted during the main tasks. First, researchers independently coded the utterances that participants in the validation condition submitted during the validation step (i.e., validation) and the utterances that participants in the control condition submitted to achieve a correct goal flow during the interaction tasks (i.e., repair). During the analysis, we focused on how their utterances changed from the previous attempt to understand their strategy of refining each utterance. Researchers iteratively generated high-level themes, discussed the themes until a consensus was reached, and created and refined a final coding schema. Researchers coded sample data (about 49%) and achieved a Krippendorff's alpha of 0.75, indicating a substantial agreement [39]. Researchers labeled the rest of the data using the established schema and counted the number of participants who mentioned each theme.

5.5.3 Validation and repair strategies analysis. Two researchers analyzed open-ended responses in the post-task survey in which participants in the validation condition explained their validation strategies and participants in the interaction condition explained their repair strategies. Following the procedure of thematic analysis [14], researchers independently coded the data and generated high-level themes. Researchers discussed their themes until a consensus was reached and created a coding schema. Researchers iterated this process until they reached 100% agreement. Researchers simply aimed for 100% agreement rather than reaching a good inter-rater reliability score in this analysis due to the small size of the dataset [50]. Most of the themes we found from the utterance analysis and the open-ended survey responses were similar, thus we report the aggregated list of themes in Table 5.

Table 4. Descriptive statistics of pre-task and post-task mental model scores overall and for each type for all conditions. We reported the means and standard deviations in parentheses. Our findings indicated a significant improvement in overall mental model scores in the validation condition, particularly regarding the knowledge distribution type of the mental model.

Mental model score	Validation condition		Social information condition		Control condition	
	Pre-task	Post-task	Pre-task	Post-task	Pre-task	Post-task
Overall	4.51 (\pm 0.3)	4.99 (\pm 0.59)	4.83 (\pm 0.26)	4.66 (\pm 0.38)	4.58 (\pm 0.42)	4.68 (\pm 0.39)
Global behavior	3.83 (\pm 0.77)	4.63 (\pm 1.42)	4.45 (\pm 0.52)	4.32 (\pm 0.9)	3.96 (\pm 0.78)	4.58 (\pm 0.98)
Local behavior	4.58 (\pm 0.48)	4.88 (\pm 0.81)	4.73 (\pm 0.39)	4.59 (\pm 0.64)	4.69 (\pm 0.55)	4.73 (\pm 0.62)
Knowledge distribution	4.77 (\pm 0.64)	5.28 (\pm 0.67)	5.11 (\pm 0.70)	4.91 (\pm 0.52)	4.77 (\pm 0.73)	4.67 (\pm 0.58)

5.5.4 Collaborative comments analysis. We followed the thematic analysis method [14] to analyze the open-ended collaborative comments, i.e. edits of others' shared social information, obtained from the treatment conditions. Two researchers independently coded and extracted high-level themes from the entire dataset. They then iteratively discussed the themes and updated a coding schema. Using the final schema, researchers achieved substantial agreement [39] (Krippendorff's alpha of 0.84) We counted the frequency of each theme and reported the results in Table 9. Additionally, two researchers independently annotated the accuracy of each collaborative comment and made a substantial agreement [39] (Krippendorff's alpha = 0.78).

6 Results

In this section, we explain the effects of validation (RQ1) and the accuracy of social information (RQ2) on user mental models. Additionally, we investigated the effects on task performance (RQ3) and explore how people collaborate to build and share more accurate social information (RQ4).

6.1 Validation (RQ1)

6.1.1 Mental models. The validation experience significantly improved the overall user mental models. We found a significant effect of the condition on the changes in mental model scores ($p < .001$). Post hoc tests demonstrated that participants who had chances to validate the accuracy of social information significantly improved their mental models compared to the participants in the social information who had no chance to validate them ($p < .05$) and compared to the control condition with marginal significance ($p < .1$).

Among the three types of mental models, we found that the knowledge distribution type showed statistically significant differences across conditions ($p < .01$), while the other types were non-significant. Post hoc tests revealed the participants who had chances to validate the accuracy of social information significantly improved their knowledge distribution type of mental models compared to the participants in the social information condition ($p < .05$) and the control condition ($p < .05$). The descriptive statistics are summarized in Table 4.

6.1.2 Accuracy judgments and confidence. To learn more about the effects of the validation experience, we investigated how participants' accuracy judgments and confidence changed before and after the validation step in the validation condition only. Mental models may improve not only when participants successfully identify inaccurate social information but also when they have increased confidence in accurate social information. We created an 'judgment success' variable, where a participant's successful judgment of the social information's accuracy resulted in a 'success' (1) outcome, and an incorrect judgment led to a 'fail' (0) outcome. We discovered that their accuracy judgment success were not significantly different before and after the validation in which only 4 participants out of 15 participants changed their judgments after validation. However, confidence

Table 5. Validation and repair strategies mentioned in the post-task survey or identified in the utterance analysis. The number of participants who mentioned each theme in the post-task survey and the number of utterances that were refined using each theme are noted in subsequent columns for the validation condition (V) and the control condition (C).

Theme	Description	Example Survey Response	Example Utterance	Survey		Utterance	
				V	C	V	C
Describing the goal flow	Described the goal flow using the provided keywords, ordering, and structure as is with subjective adjustments.	I simply used the application name and the action below it (the goal flow) to generate the flow.	"servicenow incident, jira issue, domino mailmessage" (1st attempt)	2	6	48	39
Variations	Tested alternate variations of wordings and phrases such as synonyms but the overall meaning was similar	I tried a couple of different descriptions until one style got it right the first time	"When a service now incident is created, create an issue on jira and then create a mail message on domino." (1st attempt) -> "Whenever an incident is created on servicenow, create an issue on jira and then create a mail message on domino." (2nd attempt)	3	0	41	7
Simplifying text	Shortening the utterance by removing words	I just tried to use natural-sounding language with a clear sequence of steps, minimizing the amount of unnecessary text used.	"When a message is created on slack, then create a file in dropbox" (1st attempt) -> "a message in slack, create file in dropbox" (2nd attempt)	1	1	11	12
Reordering	Rearranged parts of the utterance and words	I would rearrange how I said specific phrases that the comments would give.	"when a raw message created in slack creates files in dropbox" (1st attempt) -> "create files in dropbox when a rawmessage created in slack" (2nd attempt)	2	1	14	9
Polishing	Cosmetic edits such as fixing typos, grammar, adding spaces, capitalization	N/A	"Whwn raw messagw created in slack, dropbox create files" (1st attempt) -> "When raw message created in slack, dropbox create files" (2nd attempt)	N/A	N/A	5	2
Adding words	Added new keywords including new application names, verbs, transition words	N/A	"When there is an incident create issue and mail" (1st attempt) -> "When the incident created in servicenow create an issue in jira and create a mail in domino" (2nd attempt)	N/A	N/A	27	23
Confirmation or counterexample	Follow/replicated what is described in the social information and attempted opposite variations	I tried to use the same thought process as other users if I agreed with them and hence same/similar syntax or commands. If I disagreed with a user. I would use the opposite of what they considered better to prove them wrong.	N/A	3	0	N/A	N/A
Following an example flow	Used the example utterance that we provided when describing the goal flow	I followed the sentence structure in the given example to generate an input sentence describing the events in the goal flow.	N/A	3	2	N/A	N/A
Others	No sufficient detail, used commonsense, don't know, no strategy	I don't know how to answer this question	N/A	1	3	N/A	N/A

ratings in their accuracy judgments significantly increased after the validation (average confidence ratings of the initial judgment: 5.31 (± 1.79) vs. final judgment: 5.98 (± 0.48), $p < .05$).

6.1.3 Validation utterances and strategies. We found similar themes of strategies between the validation and the control conditions, which we listed in Table 5. In the post-task survey, one unique strategy that participants reported to validate social information was testing confirmation and counterexamples that aligns with or opposes the provided social information. Most participants in the control condition (46.15%) reported that they focused on describing the given goal flow by utilizing the provided keywords and structure, assuming that these were the most effective methods for writing an utterance.

6.2 Social information (RQ2)

As shown in Table 4, we found a small negative effect on user mental models in the social information condition, likely because the incorrect social information negatively undermined participants' understanding of the system behavior. However, the effect was not statistically significant; The number of accurate social information had no significant impact on the mental models of participants in the social information and the validation conditions ($p = .50$).

To further investigate the possible reasons contributing to this finding, we compared the actual accuracy of the social information and the participants' accuracy judgments after reading the social information. As shown in Table 6 and Table 7, we found that the majority of the social information accuracy was correctly judged. For example, participants in the social information condition were

Table 6. Contingency table between accuracy and people’s judgment in the social information condition.

Social information	Accurate	Inaccurate
Judged Accurate	14	8
Judged Inaccurate	3	8

Table 7. Contingency table between accuracy and people’s initial judgment in the validation condition.

Social information	Accurate	Inaccurate
Judged Accurate	25	10
Judged Inaccurate	2	8

able to judge the accuracy of the social information in 66.67% of the cases (22 out of 33 judgments). Participants in the validation condition were able to judge the accuracy of the social information in 73.33% of the cases (33 out of 45 initial judgments). However, analyzing the data on a per-user basis indicated that 69.23% of participants in treatment conditions incorrectly judged the social information accuracy at least once (10/15 participants in the validation condition, 8/11 participants in the social information condition). Drawing from this observation, we can anticipate inaccurate social information exerts a smaller influence on their mental models than we anticipated. This aligns with our findings in section 6.1 in which we found that the validation opportunity didn’t change their accuracy judgment but rather their confidence about their judgment. Other methods besides validation need to be devised as well to assist users in judging social information accuracy.

6.3 Task Performance (RQ3)

80% of participants in the validation condition successfully completed the final task, taking an average of 1.8 (± 0.86) attempts out of 3 available attempts. In comparison, 64% of participants in the social information condition succeeded and it took them an average of 2.36 (± 0.81) attempts. In the control condition, 54% of participants succeeded, and they took an average of 2.15 (± 0.99) attempts. However, we did not find statistically significant differences in the success rate ($p = .32$) as well as in the number of attempts ($p = .27$) across the conditions. One possible explanation for this finding is that the validation experience primarily improved the knowledge distribution type of mental models as we explained in section 6.1, thus it might not have enhanced all aspects of mental models that lead to significant improvements in task performance.

6.4 Collaborative Comments (RQ4)

In this section, we examined the collaborative comments provided by participants after each task. Through qualitative analysis, we found that participants in treatment conditions employed a set of strategies when they had a chance to edit the social information in a way that is more accurate and valuable for other users. As listed in Table 9, we found that paraphrasing the social information was the most frequent strategy employed by participants overall, particularly in the social information condition (55%). Participants who had a chance to validate the social information showed more diverse approaches to editing the social information including paraphrasing (24%), opposing the social information (20%), appending new information (24%), or adding more examples (16%).

The experimental conditions significantly affected participants’ willingness to share their comments with other users ($\chi^2(2) = 8.92, p < .05$). Post hoc tests revealed that participants in the social information condition (4.21 ± 1.76) rated their sharing intention significantly lower than participants in the control condition (6.05 ± 1.39) ($p < .01$). This finding indicates that participants who read social information were reluctant in sharing their comments, possibly due to the fact that they encountered different viewpoints from other users and became less confident about their system understanding. Other comparisons didn’t show significant differences; for example, the sharing intention between participants in the validation condition (5.29 ± 1.94) and the control condition (6.05 ± 1.39) was not significantly different ($p = .4$). This could be interpreted that

Table 8. The count of accurate, inaccurate, and N/A labels for the collaborative comments across or within individual conditions.

Accuracy	Total count	Validation condition	Social information condition	Control condition
Correct	70	26	23	21
Incorrect	39	17	10	12
N/A	8	2	0	6

Table 9. Editing strategies for collaborative comments. Edited parts in the example comments are italicized. In the frequency column, we calculated the number of themes that occurred in treatment conditions overall and within a specific condition such as in the validation condition (V) and the social information condition (S).

Strategy	Description	Original Comment	Collaborative Comment	Frequency
Paraphrasing	Rewording, clarifying, or simplifying the comment while retaining the original meaning.	The system took the order of the sentence and syntax to create a logical flow from left to right.	The system uses <i>the order of the flow</i> from left to right.	29 (37.18%) [V:11, S:18]
Opposing	Revising, removing, or correcting inaccurate parts.	When the sentences were constructed in the passive voice, it seemed to work the best.	The sentences can be constructed in an <i>active or passive voice. Both can lead to good results.</i>	15 (19.24%) [V:9, S:6]
Adding information	Adding new information such as what they did, how they dealt with the problem, and what would happen otherwise.	Words that implied an order were useful, like 'when' or 'then'. It helped the program generate a correct sequence.	Words that implied an order could be useful, like 'when' or 'then'. <i>However, the order of flow separated by a comma also worked.</i>	14 (17.95%) [V:11, S:3]
Elaborating with examples	Adding specific keywords, utterances, or output examples	The system used the verbs to create actions and the nouns as the subjects.	The system used the verbs to create actions and the nouns as the subjects. <i>Use a preposition to indicate the app. For example, 'create a lead in salesforce.'</i>	8 (10.26%) [V:7, S:1]
Adjusting the scope	Changing the scope of the comment to be more or less assertive.	At the end, I was starting to learn to simplify the sentence, and thinking fewer words worked better even if the sentence did not look right.	At the end, I was starting to learn to simplify the sentence, and thinking fewer words worked better even if the sentence did not look right <i>as long as it included keywords in the correct order.</i>	5 (6.42%) [V:2, S:3]
Summarizing	Conclude with their takeaways or interpretations in addition to the comment.	For instance, for Domino, after I changed 'mailmessage' to 'mail', it generated the right flow. So correct wording was very important.	For instance, for Domino, after I changed 'mailmessage' to 'mail', it generated the right flow. So correct wording was very important. <i>I think it is important to understand the proper phrasing for certain processes, especially considering the first part of the workflow.</i>	3 (3.85%) [V:1, S:2]
New opinion	Rewriting the comment on a new topic that is different from the original comment.	The system added extra flow for no reason. The AI is just making it's own assumptions.	<i>The system mixed up the flow order of my tasks.</i>	2 (2.57%) [V:2, S:0]
Others	Other miscellaneous comments that are not meaningful	I had to alter some of my words so that the system would understand it. For instance, for Domino, after I changed 'mailmessage' to 'mail', it generated the right flow. So correct wording was very important.	<i>My solution didn't work, so I have no comment to improve on this.</i>	2 (2.57%) [V:2, S:0]

participants who had chances to validate social information showed a higher intention to share their comments, resulting in a non-significant difference compared to the control group.

Participants wrote accurate comments more than inaccurate comments overall, as shown in Table 8. In contrast to our expectations that participants will write and share more accurate comments when they had an opportunity to validate or interact with the system, we didn't find significant differences across conditions in the accuracy of comments ($\chi^2(4) = 0.22, p = .99$). These findings indicate that participants demonstrated an ability to write accurate comments regardless of the opportunity for validation or interaction. For example, participants selectively preserved or wrote the information they deemed accurate rather than including the information that they think is uncertain or questionable (e.g., 'adjusting the scope theme' in Table 9).

7 Discussion

7.1 Findings and Implications

To review our findings, we showed the positive effects of validation experience on mental models, particularly in the knowledge distribution type of mental models. The validation experience increased participants' confidence in their understanding of how the system behaves. We also

found that the validation experience encourages people to share their mental models with other users. Participants exhibited a strong ability to discern accurate and inaccurate social information. Therefore, inaccurate social information didn't significantly harm their mental models. Through qualitative analysis, we identified strategies that participants utilized to validate social information as well as to collaboratively build more accurate shared mental models.

7.1.1 Validation. The experience of validating social information showed a positive impact on the development of mental models that align with the actual system behavior compared to the social information condition that had no validation process or control condition. Dual-process theories argue that human cognition and reasoning can be partitioned into two types of processes, the fast, effortless, and automatic process that often relies on heuristics (type 1) and the slow, reflective, and deliberative process that involves critical thinking (type 2) [22]. The validation experience encouraged individuals to engage in the type 2 process through cognitive forcing functions (e.g., instructions, visual feedback), fostering more logical thinking and analysis regarding the behavior of the AI system and thereby leading to a more accurate understanding of mental models. However, in reality, individuals might be more inclined to rely on the type 1 process by trusting what others say about an AI system due to time and effort constraints. It would be beneficial to explore mechanisms and interface designs that encourage users to engage in validation practices, similar to cognitive forcing functions [15]. For instance, incorporating a playground feature within an online community or encouraging posters to include their utterances or code snippets can facilitate validation practices while reading social information.

While we found positive effects of validation, the effect was limited to the knowledge distribution type of mental models. This is somewhat different from prior work in a word game where those who won and lost did not have significant differences in knowledge distribution [27]. One possible reason is that in our context, users find it easiest to validate the system's knowledge by adding, removing, and varying keywords to observe whether the system's knowledge base can detect the changes. This implies that depending on the user ability to construct utterance examples to prove or disapprove social information within a short time can affect how much and what they learn through the validation process. More research needs to be done to support various validation strategies that would impact all types of mental models regardless of users' ability of generating effective inputs. For instance, using large language models to generate example validation utterances can be a potential direction of research to assist users in the validation step. The global behavior of an AI system such as whether the system splits and parses the sentence into small parts is difficult to validate using input-output pairs identified in the validation phase. Rather, visualizing a parse tree or explaining how the system internally breaks down an utterance can be more useful to understand the global behavior of the system.

We found that participants increased their confidence in the accuracy judgment of the social information, which may have positively contributed to the improvement in mental models after validation experience. However, we found that users' binary accuracy judgments, either correct or incorrect, remained largely unchanged, with only a small fraction of participants altering their judgments after validation. As to why we see this result, it is possible that people had cognitive biases to their initial mental models such as anchoring bias. According to Tversky and Kahneman's theory, when people make judgments, they often start with an initial value and then adjust it, but these adjustments are typically not significant enough [68]. Although the validation experience had strengthened their confidence in their initial accuracy judgments, it may not have been sufficient to alter the direction of their judgments from incorrect to correct or vice versa. Future work can explore ways to mitigate these biases such as offering validation task before they form initial mental models or providing alternative mental models to reduce the impact of the anchor.

7.1.2 Social information. We did not find that inaccurate social information had a significant impact on shaping mental models. According to belief formation theories [26], beliefs formed through direct observation or experience (i.e., descriptive belief) are stronger than beliefs formed through information by some sources including social information (i.e., informational belief) because people rarely question their own senses from direct experience. In our study, participants were given a chance to form their initial mental models during the onboarding phase, and it is possible that this experience might have overridden the effects of the social information they encountered subsequently. Future research could explore the effects of social information on individuals who have not yet formed any mental models. Moreover, individuals might have felt it difficult to fully understand others' mental models through a single piece of social information from an unknown user, which can result in limited social influence to change their initial opinions. Prior works have found that social influence could become stronger if the information is provided by an expert (the expert effect) or if there is a presence of clusters of individuals sharing similar opinions (the majority effect) [52]. Therefore, we encourage future work to explore the effects of social information by varying the credibility and the number of people who mention the information in a longer-term experiment.

When participants were given the opportunity to revise social information through the collaborative comment step, we found that most participants across conditions were able to write correct comments by removing parts of the comments that they were not confident with. A small portion of participants still ended up writing inaccurate comments after the validation process, and one possible reason for this is that they might have failed to validate the social information correctly such as only creating examples that confirms the incorrect social information.

7.1.3 Validation and repair strategies. We identified strategies that participants utilized to validate social information and the repair strategies they used in the interaction task in the control condition (see Table 5). While there's a significant improvement of mental models in the validation condition compared to the control condition, we found that the strategies participants used in validation were similar to the strategies they used during the interaction task. One potential difference could lie in the motivation underlying these activities. One is driven by an accuracy motivation to obtain more accurate mental models of the system through testing their hypothesis related to social information, while the other is characterized by a goal-oriented motivation to achieve a correct outcome. When participants were queried about their mental models, those who had engaged in validation activities with accuracy motivation felt more confident in their mental models compared to those who focused solely on completing the tasks. This highlights the importance of fostering accuracy motivation when designing AI systems that require accurate mental models. For instance, AI systems can incorporate activities or training materials that encourage users to accurately understand the system's behavior before using the system for their goals.

7.2 Generalizability

Our research may have broader implications beyond the specific AI system we investigated, which generates task workflows from natural language utterances. Like our system, many AI systems are black box models, lacking transparency in their decision-making processes, which increases the amount of inaccurate mental models created and shared by people. For example, large language models have several characteristics that pose challenges for transparency, such as their complexity, proprietary nature, and massive size [44]. While many users actively share their mental models around these technologies, they can easily become obsolete and flawed as the technology is constantly and rapidly evolving [44]. Moreover, users may interact with algorithms through applications such as social media or news recommendation platforms that are built on many

interacting algorithms, making it even more difficult to understand or explain its underlying system [24, 54].

To supplement the limitations of the transparency approach, our study highlights the importance of a user's role to critical think and validate the social information they receive. In comparison to the example-based explanations in XAI research [16, 69], which provide example cases as explanations, validation experience encourages users to make and test their own input examples to deeply understand the system. While AI systems, including large language models, are becoming increasingly versatile, it is challenging to provide universal and comprehensive examples to explain the system; therefore, allowing users to freely formulate their input examples to validate specific mental models is recommended. Developers and practitioners can support this role by providing playground interfaces or sandbox environment where users can experiment with different inputs, parameters, and data to understand how the system actually works and shape their mental models accordingly. Building an online community where people share their validation experiences can also be useful. However, we want to acknowledge that validation approach has its own limitations such as the difficulty of coming up with effective inputs and the interpreting the outputs generated. To support this, AI models may present systematic comparisons between the models' outputs for different user-generated inputs and explain why there are differences, which may enhance users' mental models. We encourage future work to devise advanced methods to help users validate social information and gain better understanding of the AI systems.

7.3 Limitations

There are several limitations to this study. First, we tested one specific natural language system and a specific task (i.e., trigger-action program) to answer our research questions and our findings may not generalize to all other AI systems and tasks. We encourage future research to explore other types of AI systems in diverse systems and tasks. Second, the social information used in this study consisted of only short and easily understandable text-based information, which was collected from a preliminary experiment. While the information was effective to answer our research question, real-world social information may consist of ambiguous, long, or contain a variety of other contextual information, such as images or references. We recommend experimenting with more diverse types and formats of social information in future studies. Third, our participants have at least heard about AI in the news or in their social circle and they might already have formed some knowledge around AI systems. We recruited participants from MTurk and our participants were likely unfamiliar with workflow automation tools, which might have posed challenges in completing the tasks. We recommend conducting further research with a larger and more diverse population with varying AI-related backgrounds or experience in automation tools.

8 Conclusion

We investigated the impact of social information accuracy and validation experience on mental models in the context of a natural language system that generates workflows. We designed a between-subjects online experiment with 39 crowdworkers. Participants were randomly assigned to one of the three conditions; the validation condition included both social information followed by the validation step, the social information condition involved social information only, and the control condition had neither social information nor validation. The results indicated the positive impacts of the validation experience on mental models, specifically in the knowledge distribution type. Participants demonstrated the capacity to differentiate between accurate and inaccurate social information, thus mitigating the effects of inaccurate information on their mental models. We discussed design implications and future research on how to improve validation practices in social platforms and AI systems.

References

- [1] Amina Adadi and Mohammed Berrada. 2018. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE access* 6 (2018), 52138–52160.
- [2] Katrin Affolter, Kurt Stockinger, and Abraham Bernstein. 2019. A comparative survey of recent natural language interfaces for databases. *The VLDB Journal* 28, 5 (2019), 793–819.
- [3] Kamran Alipour, Arijit Ray, Xiao Lin, Michael Cogswell, Jurgen P Schulze, Yi Yao, and Giedrius T Burachas. 2021. Improving users' mental model with attention-directed counterfactual edits. *Applied AI Letters* 2, 4 (2021), e47.
- [4] Mike Ananny and Kate Crawford. 2018. Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *new media & society* 20, 3 (2018), 973–989.
- [5] Robert W Andrews, J Mason Lilly, Divya Srivastava, and Karen M Feigh. 2023. The role of shared mental models in human-AI teams: a theoretical review. *Theoretical Issues in Ergonomics Science* 24, 2 (2023), 129–175.
- [6] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bannetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information fusion* 58 (2020), 82–115.
- [7] Zahra Ashktorab, Benjamin Hoover, Mayank Agarwal, Casey Dugan, Werner Geyer, Hao Bang Yang, and Mikhail Yurochkin. 2023. Fairness Evaluation in Text Classification: Machine Learning Practitioner Perspectives of Individual and Group Fairness. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–20.
- [8] Jaime Banks. 2020. Optimus primed: Media cultivation of robot mental models and social judgments. *Frontiers in Robotics and AI* 7 (2020), 62.
- [9] Gagan Bansal, Besmira Nushi, Ece Kamar, Walter S Lasecki, Daniel S Weld, and Eric Horvitz. 2019. Beyond accuracy: The role of mental models in human-AI team performance. In *Proceedings of the AAI conference on human computation and crowdsourcing*, Vol. 7. 2–11.
- [10] Gagan Bansal, Besmira Nushi, Ece Kamar, Daniel S Weld, Walter S Lasecki, and Eric Horvitz. 2019. Updates in human-ai teams: Understanding and addressing the performance/compatibility tradeoff. In *Proceedings of the AAI Conference on Artificial Intelligence*, Vol. 33. 2429–2437.
- [11] Simon Baron-Cohen. 1997. *Mindblindness: An essay on autism and theory of mind*. MIT press.
- [12] Michelle Brachman, Christopher Bygrave, Tathagata Chakraborti, Arunima Chaudhary, Zhining Ding, Casey Dugan, David Gros, Thomas Gschwind, James Johnson, Jim Laredo, et al. 2022. A Goal-Driven Natural Language Interface for Creating Application Integration Workflows. (2022).
- [13] Michelle Brachman, Qian Pan, Hyo Jin Do, Casey Dugan, Arunima Chaudhary, James M Johnson, Priyanshu Rai, Tathagata Chakraborti, Thomas Gschwind, Jim A Laredo, et al. 2023. Follow the Successful Herd: Towards Explanations for Improved Use and Mental Models of Natural Language Systems. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*. 220–239.
- [14] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.
- [15] Zana Bućinca, Maja Barbara Malaya, and Krzysztof Z Gajos. 2021. To trust or to think: cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–21.
- [16] Carrie J Cai, Jonas Jongejan, and Jess Holbrook. 2019. The effects of example-based explanations in a machine learning interface. In *Proceedings of the 24th international conference on intelligent user interfaces*. 258–262.
- [17] Kelly Caine. 2016. Local standards for sample size at CHI. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. 981–992.
- [18] John M Carroll and Judith Reitman Olson. 1988. Mental models in human-computer interaction. *Handbook of human-computer interaction* (1988), 45–65.
- [19] Diogo V Carvalho, Eduardo M Pereira, and Jaime S Cardoso. 2019. Machine learning interpretability: A survey on methods and metrics. *Electronics* 8, 8 (2019), 832.
- [20] Jesse Chandler, Cheskie Rosenzweig, Aaron J Moss, Jonathan Robinson, and Leib Litman. 2019. Online panels in social science research: Expanding sampling methods beyond Mechanical Turk. *Behavior research methods* 51, 5 (2019), 2022–2038.
- [21] Valerie Chen, Q Vera Liao, Jennifer Wortman Vaughan, and Gagan Bansal. 2023. Understanding the role of human intuition on reliance in human-AI decision-making with explanations. *Proceedings of the ACM on Human-computer Interaction* 7, CSCW2 (2023), 1–32.
- [22] Wim De Neys. 2017. *Dual process theory 2.0*. Routledge.
- [23] Jeff Druce, James Niehaus, Vanessa Moody, David Jensen, and Michael L Littman. 2021. Brittle AI, causal confusion, and bad mental models: challenges and successes in the XAI program. *arXiv preprint arXiv:2106.05506* (2021).
- [24] Motahhare Eslami, Karrie Karahalios, Christian Sandvig, Kristen Vaccaro, Aimee Rickman, Kevin Hamilton, and Alex Kirlik. 2016. First I "like" it, then I hide it: Folk Theories of Social Feeds. In *Proceedings of the 2016 CHI conference on*

- human factors in computing systems*. 2371–2382.
- [25] Robert M Fein, Gary M Olson, and Judith S Olson. 1993. A mental model can help with learning to operate a complex device. In *INTERACT'93 and CHI'93 conference companion on Human factors in computing systems*. 157–158.
- [26] Martin Fishbein and Icek Ajzen. 1977. Belief, attitude, intention, and behavior: An introduction to theory and research. *Philosophy and Rhetoric* 10, 2 (1977).
- [27] Katy Ilonka Gero, Zahra Ashktorab, Casey Dugan, Qian Pan, James Johnson, Werner Geyer, Maria Ruiz, Sarah Miller, David R Millen, Murray Campbell, et al. 2020. Mental models of AI agents in a cooperative game setting. In *Proceedings of the 2020 chi conference on human factors in computing systems*. 1–12.
- [28] Rosanna E Guadagno, Daniel M Rempala, Shannon Murphy, and Bradley M Okdie. 2013. What makes a video go viral? An analysis of emotional contagion and Internet memes. *Computers in Human Behavior* 29, 6 (2013), 2312–2319.
- [29] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2018. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)* 51, 5 (2018), 1–42.
- [30] Sumit Gulwani and Mark Marron. 2014. Nlyze: Interactive programming by natural language for spreadsheet data analysis and manipulation. In *Proceedings of the 2014 ACM SIGMOD international conference on Management of data*. 803–814.
- [31] Gary G Hendrix. 1982. Natural-language interface. *American Journal of Computational Linguistics* 8, 2 (1982), 56–61.
- [32] Cecilia Heyes. 2012. What's social about social learning? *Journal of comparative psychology* 126, 2 (2012), 193.
- [33] Ting-Hao Kenneth Huang, Amos Azaria, and Jeffrey P Bigham. 2016. Instructablecrowd: Creating if-then rules via conversations with the crowd. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. 1555–1562.
- [34] Maurice Jakesch, Jeffrey T Hancock, and Mor Naaman. 2023. Human heuristics for AI-generated language are flawed. *Proceedings of the National Academy of Sciences* 120, 11 (2023), e2208839120.
- [35] Ellen Jiang, Edwin Toh, Alejandra Molina, Kristen Olson, Claire Kayacik, Aaron Donsbach, Carrie J Cai, and Michael Terry. 2022. Discovering the Syntax and Strategies of Natural Language Programming with Generative Language Models. In *CHI Conference on Human Factors in Computing Systems*. 1–19.
- [36] Markelle Kelly, Aakriti Kumar, Padhraic Smyth, and Mark Steyvers. 2023. Capturing Humans' Mental Models of AI: An Item Response Theory Approach. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 1723–1734.
- [37] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. 2018. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*. PMLR, 2668–2677.
- [38] Tae Soo Kim, DaEun Choi, Yoonseo Choi, and Juho Kim. 2022. Stylette: Styling the Web with Natural Language. In *CHI Conference on Human Factors in Computing Systems*. 1–17.
- [39] Klaus Krippendorff. 2004. Reliability in content analysis: Some common misconceptions and recommendations. *Human communication research* 30, 3 (2004), 411–433.
- [40] Todd Kulesza, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf. 2015. Principles of explanatory debugging to personalize interactive machine learning. In *Proceedings of the 20th international conference on intelligent user interfaces*. 126–137.
- [41] Todd Kulesza, Simone Stumpf, Margaret Burnett, and Irwin Kwan. 2012. Tell me more? The effects of mental model soundness on personalizing an intelligent agent. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 1–10.
- [42] Todd Kulesza, Simone Stumpf, Margaret Burnett, Sherry Yang, Irwin Kwan, and Weng-Keen Wong. 2013. Too much, too little, or just right? Ways explanations impact end users' mental models. In *2013 IEEE Symposium on visual languages and human centric computing*. IEEE, 3–10.
- [43] Toby Jia-Jun Li, Marissa Radensky, Justin Jia, Kirielle Singarajah, Tom M Mitchell, and Brad A Myers. 2019. Pumice: A multi-modal agent that learns concepts and conditionals from natural language and demonstrations. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology*. 577–589.
- [44] Q Vera Liao and Jennifer Wortman Vaughan. 2023. AI Transparency in the Age of LLMs: A Human-Centered Research Roadmap. *arXiv preprint arXiv:2306.01941* (2023).
- [45] Brian Y Lim, Anind K Dey, and Daniel Avrahami. 2009. Why and why not explanations improve the intelligibility of context-aware intelligent systems. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 2119–2128.
- [46] James Lin, Jeffrey Wong, Jeffrey Nichols, Allen Cypher, and Tessa A Lau. 2009. End-user programming of mashups with vegemite. In *Proceedings of the 14th international conference on Intelligent user interfaces*. 97–106.
- [47] Han Liu, Vivian Lai, and Chenhao Tan. 2021. Understanding the effect of out-of-distribution examples and interactive explanations on human-ai decision making. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–45.

- [48] Zhuoran Lu, Patrick Li, Weilong Wang, and Ming Yin. 2022. The Effects of AI-based Credibility Indicators on the Detection and Spread of Misinformation under Social Influence. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (2022), 1–27.
- [49] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems* 30 (2017).
- [50] Nora McDonald, Sarita Schoenebeck, and Andrea Forte. 2019. Reliability and inter-rater reliability in qualitative research: Norms and guidelines for CSCW and HCI practice. *Proceedings of the ACM on human-computer interaction* 3, CSCW (2019), 1–23.
- [51] Omid Mohaddesi, Noah Chicoine, Min Gong, Ozlem Ergun, Jacqueline Griffin, David Kaeli, Stacy Marsella, and Casper Hartevelt. 2023. Thought Bubbles: A Proxy into Players' Mental Model Development. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–21.
- [52] Mehdi Moussaid, Juliane E Kämmer, Pantelis P Analytis, and Hansjörg Neth. 2013. Social influence and the collective dynamics of opinion formation. *PLoS one* 8, 11 (2013), e78433.
- [53] Yuri Nakao, Simone Stumpf, Subeida Ahmed, Aisha Naseer, and Lorenzo Strappelli. 2022. Toward involving end-users in interactive human-in-the-loop AI fairness. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 12, 3 (2022), 1–30.
- [54] Thao Ngo and Nicole Krämer. 2022. Exploring folk theories of algorithmic news curation for explainable design. *Behaviour & Information Technology* 41, 15 (2022), 3346–3359.
- [55] Don Norman. 2013. *The design of everyday things: Revised and expanded edition*. Basic books.
- [56] Donald A Norman. 2014. Some observations on mental models. In *Mental models*. Psychology Press, 15–22.
- [57] Mahsan Nourani, Chiradeep Roy, Jeremy E Block, Donald R Honeycutt, Tahrira Rahman, Eric Ragan, and Vibhav Gogate. 2021. Anchoring bias affects mental model formation and user reliance in explainable AI systems. In *26th International Conference on Intelligent User Interfaces*. 340–350.
- [58] Gordon Pennycook and David G Rand. 2021. The psychology of fake news. *Trends in cognitive sciences* 25, 5 (2021), 388–402.
- [59] José Pinheiro and Douglas Bates. 2006. *Mixed-effects models in S and S-PLUS*. Springer Science & Business Media.
- [60] Chris Quirk, Raymond Mooney, and Michel Galley. 2015. Language to code: Learning semantic parsers for if-this-then-that recipes. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 878–888.
- [61] Mitchel Resnick, Robbie Berg, and Michael Eisenberg. 2000. Beyond black boxes: Bringing transparency and aesthetics back to scientific investigation. *The Journal of the Learning Sciences* 9, 1 (2000), 7–30.
- [62] Beau G Schelble, Christopher Flathmann, Nathan J McNeese, Guo Freeman, and Rohit Mallick. 2022. Let's think together! Assessing shared mental models, performance, and trust in human-agent teams. *Proceedings of the ACM on Human-Computer Interaction* 6, GROUP (2022), 1–29.
- [63] Subhasree Sengupta and Caroline Haythornthwaite. 2020. Learning with comments: An analysis of comments and community on Stack Overflow. (2020).
- [64] Vidya Setlur, Sarah E Battersby, Melanie Tory, Rich Gossweiler, and Angel X Chang. 2016. Eviza: A natural language interface for visual analysis. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*. 365–377.
- [65] Abhay Sukumaran and Clifford Nass. 2010. Socially cued mental models. In *CHI'10 Extended Abstracts on Human Factors in Computing Systems*. 3379–3384.
- [66] Ian Tenney, James Wexler, Jasmijn Bastings, Tolga Bolukbasi, Andy Coenen, Sebastian Gehrmann, Ellen Jiang, Mahima Pushkarna, Carey Radebaugh, Emily Reif, et al. 2020. The language interpretability tool: Extensible, interactive visualizations and analysis for NLP models. *arXiv preprint arXiv:2008.05122* (2020).
- [67] Stefano Teso and Kristian Kersting. 2019. Explanatory interactive machine learning. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 239–245.
- [68] Amos Tversky and Daniel Kahneman. 1974. Judgment under Uncertainty: Heuristics and Biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *science* 185, 4157 (1974), 1124–1131.
- [69] Jasper van der Waa, Elisabeth Nieuwburg, Anita Cremers, and Mark Neerinx. 2021. Evaluating XAI: A comparison of rule-based and example-based explanations. *Artificial intelligence* 291 (2021), 103404.
- [70] Sandra A Vannoy and Prashant Palvia. 2010. The social influence model of technology adoption. *Commun. ACM* 53, 6 (2010), 149–153.
- [71] Qiaosi Wang, Koustuv Saha, Eric Gregori, David Joyner, and Ashok Goel. 2021. Towards mutual theory of mind in human-ai interaction: How language reflects what students perceive about a virtual teaching assistant. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–14.
- [72] Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, et al. 2021. Ethical and social risks of harm from language models. *arXiv preprint*

arXiv:2112.04359 (2021).

- [73] Jeffery D Wilfong. 2006. Computer anxiety and anger: The impact of computer use, computer experience, and self-efficacy beliefs. *Computers in human behavior* 22, 6 (2006), 1001–1011.
- [74] Terry Winograd. 1971. *Procedures as a representation for data in a computer program for understanding natural language*. Technical Report. MASSACHUSETTS INST OF TECH CAMBRIDGE PROJECT MAC.
- [75] Frank F Xu, Bogdan Vasilescu, and Graham Neubig. 2022. In-ide code generation from natural language: Promise and challenges. *ACM Transactions on Software Engineering and Methodology (TOSEM)* 31, 2 (2022), 1–47.

A Social Information Data

We sampled social information from survey responses collected in the prior study. In the preliminary study, researchers asked open-ended questions to inquire about their mental models.

Mental Model	Social Information	Accuracy
Global behavior	The system detects transitions in my words and breaks my texts down into different parts, then connects different parts as separate small tasks.	Correct
	The system used the verbs to create actions and the nouns as the subjects.	Correct
	The system used my words and did not pay attention to the order in which I typed them in. It jumbled up the order.	Incorrect
	The system took the order of the sentence and syntax to create a logical flow from left to right.	Incorrect
Local behavior	Words that implied an order were useful, like 'when' or 'then'. It helped the program generate a correct sequence.	Correct
	I used the name of the object and action keywords close to the service names so that the system could provide me with the most accurate result.	Correct
	At the end, I was starting to learn to simplify the sentence and thinking less words worked better even if the sentence did not look right.	Incorrect
	When the sentences were constructed in the passive voice, it seemed to work the best.	Incorrect
Knowledge distribution	I had to alter some of my words so that the system would understand it. For instance, for Domino, after I changed 'mailmessage' to 'mail', it generated the right flow. So correct wording was very important.	Correct
	The system looked for keywords such as app names to know which app is being referenced and then also recognized keywords like create for making new objects in those apps.	Correct
	Keeping the sentence logical while sticking to the syntax. If the goal flow said 'create', then use 'create'. When I tried to use 'write', it messed up.	Incorrect
	The system added extra flow for no reason. The AI is just making its own assumptions.	Incorrect

B Mental model survey questions

We summarize survey questions used to measure mental models, grouped by the types of mental models each item is associated with. Note that in the surveys, the questions were not labeled with the headings and the correctness. The same set of questions was used in pre-task and post-task surveys.

Mental Model Type	Question (<i>Rate the following statements describing our system that you interacted with, using a scale of 1: Strongly disagree – 7: Strongly agree</i>)	Accuracy
Global behavior	The system breaks down the input sentence into separate words and uses certain parts of speech (e.g., nouns, verbs, or prepositions) to generate a flow.	Correct
	The ordering of phrases in the input sentence is always going to relate to the ordering of the generated flow components.	Incorrect
Local behavior	The system detects certain trigger words to create the structure of the flow (e.g., 'if').	Correct
	The system works best when an application, object, and operation words for one flow component are placed near each other in the input sentence.	Correct
	The system works best with the shortest sentence, even if it is grammatically incorrect such as lacking verbs.	Incorrect
	The system requires verb tenses to be consistent in the sentence to generate a correct flow.	Incorrect
Knowledge distribution	The system attempts to match words in the input sentence with application, operations, and object names that the system knows.	Correct
	All of the keywords that appear in the goal flow must be in the sentence using the exact same wordings to generate the correct flow.	Incorrect
	The system has knowledge about frequently used flows, so it may suggest an alternative flow that differs from the one described in the input sentence.	Incorrect
	The system requires application names (e.g., 'Gmail') to be in the input sentence to generate the correct flow.	Correct

Received July 2023; revised April 2024; accepted July 2024