Co-Designing with Users the Explanations for a Proactive Auto-Response Messaging Agent

PRANUT JAIN, University of Pittsburgh, USA
ROSTA FARZAN, University of Pittsburgh, USA
ADAM J. LEE, University of Pittsburgh, USA

Explanations of AI Agents' actions are considered to be an important factor in improving users' trust in the decisions made by autonomous AI systems. However, as these autonomous systems evolve from reactive, i.e., acting on user input, to proactive, i.e., acting without requiring user intervention, there is a need to explore how the explanation for the actions of these agents should evolve. In this work, we explore the design of explanations through participatory design methods for a proactive auto-response messaging agent that can reduce perceived obligations and social pressure to respond quickly to incoming messages by providing unavailability-related context. We recruited 14 participants who worked in pairs during collaborative design sessions where they reasoned about the agent's design and actions. We qualitatively analyzed the data collected through these sessions and found that participants' reasoning about agent actions led them to speculate heavily on its design. These speculations significantly influenced participants' desire for explanations and the controls they sought to inform the agents' behavior. Our findings indicate a need to transform users' speculations into accurate mental models of agent design. Further, since the agent acts as a mediator in human-human communication, it is also necessary to account for social norms in its explanation design. Finally, user expertise in understanding their habits and behaviors allows the agent to learn from the user their preferences when justifying its actions.

CCS Concepts: • Human-centered computing → Empirical studies in interaction design.

Additional Key Words and Phrases: proactive agents, explanations, co-design, messaging awareness

ACM Reference Format:
Pranut Jain, Rosta Farzan, and Adam J. Lee. 2023. Co-Designing with Users the Explanations for a Proactive Auto-Response Messaging Agent. Proc. ACM Hum.-Comput. Interact. 7, MHCI, Article 201 (September 2023), 23 pages. https://doi.org/10.1145/3604248

1 INTRODUCTION

Adoption of interactive virtual agents is growing at a tremendous pace, evidenced by their increased deployment and usage in various industries[1]. People frequently interact with these virtual, agent-based systems for a variety of reasons including maintaining their smart-home systems [28] or getting customer service online[2]. Most current interactions with these virtual agents are reactive, i.e., they respond to an individual's command or request. For instance, instructing a smart-home

[1]https://www.prnewswire.com/in/news-releases/chatbots-market-size-to-reach-usd-3892-1-million-by-2028-at-a-cagr-of-20-valuates-reports-841372826.html
[2]https://gettalkative.com/info/virtual-shopping

Authors' addresses: Pranut Jain, University of Pittsburgh, Pittsburgh, PA, USA, pranutjain@pitt.edu; Rosta Farzan, University of Pittsburgh, Pittsburgh, PA, USA, rfarzan@pitt.edu; Adam J. Lee, University of Pittsburgh, Pittsburgh, PA, 15260, USA, adamlee@cs.pitt.edu.

2573-0142/2023/9-ART201 $15.00
https://doi.org/10.1145/3604248

voice assistant to turn off the lights. As the capabilities and intelligence of virtual assistants expand, they are also starting to be utilized in proactive ways [18]. For instance, predictive models have been used to develop virtual assistants that send automated responses to incoming messages on mobile devices [27] or provide real-time recommendations and suggestions for equipment maintenance [9]. The increased research interest in proactive agent design also stems from the recent developments in general-purpose sensing, where sensors can monitor larger contexts rather than be restricted to a single instrumentation [37]. Combining streams from multiple sensors can further enable highly accurate prediction models for a multitude of applications [13, 25, 37]. However, a high level of proactive behavior can also lead users to feel a sense of loss of control due to the agent acting autonomously without their direct input [27, 68], which can subsequently affect their trust in these systems [33].

The traditional black box design of AI systems can make it dificult for people to understand how they work [19]. This, in turn, can impede peoples' formation of accurate mental models—i.e., abstractions of the anticipated mechanisms that a system uses to perform a given task [51]—which are vital to enable proper use of a system [2]. The lack of accurate understanding may result in negative consequences for users, such as developing aversions to a technology [65] and exerting unnecessary effort to use the system functions [27]. It may even harm the users through unexpected disclosure of sensitive information [11]. Explanations have been instrumental in improving user understanding of automated agents' actions and building trust in automated systems [1, 41, 47, 61]. Particularly in recommender systems, various explanation interfaces such as textual [32], visual [62], and interactive [35, 53] have been explored. These explanations usually aim to improve transparency, effectiveness, persuasiveness, scrutiny, trust, satisfaction, and eficiency of recommendations [60].

In this work, we explore how we can design explanations for actions taken by a proactive auto-response messaging agent [27]. The agent auto-responds on behalf of its users when it predicts them to be unavailable to reduce their sense of obligation to respond to any incoming message immediately. The auto-response includes context related to the user's unavailability state to improve situational awareness of the user's state to their contacts. There are several challenges involved in explaining the behavior of such agents: (1) determining what explanations users desire; (2) how these explanations should be presented, and (3) when is the best timing to present these explanations. Furthermore, since the agent is acting proactively as an intermediary in human-human communication, it is essential to ensure that the nuances of human-human interactions are supported and users do not feel an additional burden to justify the agent's behavior to their contacts [27]. Addressing these challenges informs our research questions, as detailed below.

- RQ1: How do users reason about the design and actions of a proactive auto-response messaging agent? As people interact with and reason about technology, they naturally form mental models [17] of how it works [2, 51]. Understanding users' reasoning can help us understand how these mental models are formed and support users in building more accurate understandings of AI-based agent systems. Exploring this question helps us identify gaps in user knowledge related to agent understanding, and what explanations can help fill those gaps. Further, understanding where in their reasoning process users go off target can help identify when to present explanations.
- RQ2: What are users' motivations for desiring explanations of the behaviors of a proactive AI agent? Understanding user motivations to desire explanations can help identify opportunities (when) to proactively present explanations to users, reducing their effort to ask for an explanation. Further, due to the social aspect (intermediary in human communication) associated with the agent use, users may be more critical of some agent actions over others [27]. Thus, understanding the motivation behind desiring explanations of

different agent behaviors can help us design explanation interfaces that precisely address user concerns without overwhelming them with too much information.
- RQ3: In what ways can interactions with the agent create opportunities to learn from and teach the agent? Understanding how we can design interactions supporting improved learning about AI and design feedback mechanisms that can help users teach AI about their preferences is crucial in allowing users to better appropriate the agent for their use [27, 49].

We conducted a design study with 14 participants (paired into seven dyads) in two phases to address our research questions. First, users interacted with the messaging agent for two weeks to become familiar with its capabilities. Then, they participated in a design session to discuss the design of an explanation module for the agent. We used qualitative methods to analyze the data collected through the design sessions. Our findings indicate that participants formed their initial mental models of agent behavior through observations and prior technology experiences. The mismatch between their initial mental models and the actual agent model created a desire for further explanations from the agent. We also observed that dyadic interactions during the design sessions were influential in helping participants refine their mental models. Our participants' discussions were often focused on the agent's decisions, where the agent made decisions without user intervention. Relatedly, emotional responses became heightened because the agent intercedes in an existing interpersonal relationship between the message sender and the recipient. Our participants also recognized that they were uniquely positioned to teach/inform the agent given their ground-truth knowledge of reasons for their own (un)availability.

A higher level of understanding of intelligent agents can lead to more effective use of these agents [49] and more effective human-AI teaming to achieve users' goals, such as attending to their ongoing tasks rather than worrying about responding to every incoming message. While multiple works have been on developing explanations for intelligent agent systems, our study utilizes the co-design methodology to directly involve users in designing explanations for a messaging agent. Our work contributes and provides insights into users' thought processes and priorities when trying to understand the messaging agent's behavior as it acts as an intermediary in their messaging communications. This understanding can help designers develop explanation interfaces that facilitate user understanding of proactive messaging agents and augment these interfaces with appropriate controls to allow users to tune the agent to their preferences.

2 BACKGROUND AND RELATED WORK

Explanations of intelligent agent system actions have been long studied. Past studies have looked into what explanations users desire [23] and how user characteristics can influence user preferences related to the presentation of these explanations [39]. As we shift towards increasingly proactive and event-driven computing [18], we need to consider user motivations for explanations from these systems. This is important since presenting explanations for each agent action and including details of every factor in the agent's decision may create an information overload [2, 31, 57] and may cause users to ignore explanations altogether. For instance, even though not directly comparable to AI explanations, privacy policies and terms of agreement documents are likely to be ignored by the users due to their complexity, language, and length even though they are usually only presented once during the first interaction with a new platform [4, 46, 55]. To help navigate privacy policies, summarization and concise information about data collection, usage, and sharing in the form of Privacy Nutrition Labels have been effective in helping users find information accurately related to their data while also allowing them to enjoy the information-seeking process [29, 30].

We start by discussing significant breakthroughs in user understanding of the system through explanations of agent behavior. Then we provide background on proactive agents and the challenges of designing explanations for these agents.

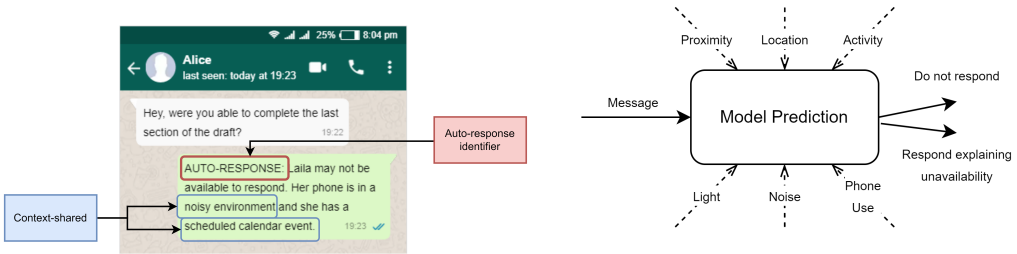
2.1 Explanations and Understanding of agent-based systems

There has been significant work in exploring and designing explanation interfaces to improve the understanding of intelligent systems' decisions. Haynes et al. described the different types of explanations users desired as they interacted with an intelligent agent (SAPS) [23]. The study participants (domain experts and developers) were given tasks familiarizing them with the agent and its controls. They reported that users wanted operational ('how do I use it?'), mechanistic ('how does it work?'), ontological (identity, definitions, and relations) explanations, and design rationale for various agent constructs. With more agents utilizing machine learning (ML), explainable AI (XAI) research has focused on explanations related to what the models have learned (global explanations) and why a particular prediction was made (local explanations). Particularly for classification tasks, the focus is on the top features that the model has learned [22]. The design of these explanations often relies on the researchers' intuition of important constituents of explanations, generally following a more algorithmic view [39, 44], which may not always be appropriate for novice or lay users with limited understanding of AI and ML [10, 39, 57]. These users have been shown to prefer what is known as local explanations, i.e., explanations about a specific model prediction rather than the overall model reasoning (global explanations) [39]. Although, these local explanations can be misleading for novice users [8, 15]. Further, the interpretability of explanations can also depend on their content and presentation [24, 57]. For instance, explanations can be textual [32], visual [62], interactive [35, 53] or a mix of different types [57]. Szymanski et al. reported that while novice users preferred visual explanations, they often drew inaccurate conclusions from them [57].

Compared to these prior works, our work studies information users desire and their motivations for desiring AI explanations. This is important since some agents, particularly proactive agents, can take multiple actions within a short period. Presenting explanations for all those actions can easily overwhelm the user causing an information overload [2, 31] and leading them to draw incorrect conclusions [57]. Further, more agents now have conversational or social aspects associated with them. These agents not only interact with their owners but could also interact with non-owners or bystanders [7]. Depending on the task the agent is assigned, user expertise in that task can vary, affecting how users perceive the agent's utility [27]. For instance, for a messaging agent assisting users in mediating messaging interactions, the user understands their reasons for being unavailable. The agent is tasked with trying to approximate and signal these reasons to their contacts which it may not always be able to do as well as the user might have. By identifying user motivations and reasoning based on their interaction with the system, we identify gaps in their understanding and opportunities to present these explanations, enabling them to form more accurate mental models of the function of these agents. Further, we involve the users directly in the design process, engaging them to think about aspects vital to them in the design of explanation interfaces.

2.2 Proactive agents and systems

The explanations and understanding of the system become even more critical when considering proactive systems [67]. These systems are developed with the ability to act autonomously on behalf of their users [18, 33]. This contrasts reactive systems, which function based on user input. Kraus et al., in their work, highlight the lack of a clear definition of proactive agents [33], with autonomous systems often classified using the levels by Sheridan and Verplank [52]. Level 10 of the classification scale identifies an entirely autonomous system; lower levels tend to shift the responsibility of taking actions to the users with assistance from the system. Zhargham et al. studied



(a) A sample agent response (b) Modeling setup

Fig. 1. An example auto-response from Jain et al. [27] is shown on the left. The agent's modeling, shown on the right, includes environmental and device usage features.

the desirable circumstances for using proactive voice assistants [68]. Their findings suggested that while proactive interactions can be useful, there are situations where these can lead to a loss of agency, and such behavior can be construed as inappropriate and invasive. Yorke-Smith et al., in their work on a proactive task management agent, proposed guidelines for the design of proactive agents [67]. They emphasized the importance of the user being able to understand such a system. Kraus et al. studied the effects of different levels of proactive agent actions. Their findings suggested that users tended to trust systems with low to medium proactivity as they felt more in control, and proactive suggestions from the system also helped reinforce their beliefs [33]. Even systems with high proactive behavior that acted independently but provided an explanation for their action still resulted in greater trust [33]. This is important since, for some tasks, such as signaling unavailability through the virtual assistant, the assumption is that the user cannot get to their messages. In these cases, the agent must act highly proactively to be useful.

We extend prior research in proactive agent systems by exploring the design of explanations to improve user understanding of an auto-response messaging agent. Our work looks into the users' reasoning process as they think about the explanations they desire from this agent and the presentation and methods for accessing these explanations. Further, we look into how users want to align their understanding and mental models with the agent's behavior and how that informs their design discussions to teach the agent.

3 RESEARCH CONTEXT: PROACTIVE AUTO-RESPONSE MESSAGING AGENT

A well-studied issue in mobile messaging relates to the lack of awareness that message senders have of recipient status while initiating conversation [14, 48, 58]. Without accurate availability awareness, it is dificult for message senders to form accurate expectations of recipient response times. Studies have shown that people typically expect fast responses to their messages, despite the asynchronous nature of the mobile messaging medium [43]. These expectations can pressure message recipients to respond quickly to incoming messages, which can cause interruptions to ongoing tasks and over-engagement with messaging applications [6].

To help mitigate this tension, prior work has looked into building user attention models to improve availability awareness in mobile messaging [25, 26, 48]. These models utilize users' phone usage and environmental information to predict whether they will attend to a message within a certain threshold (typically 5-10 minutes). Jain et al. used the personalized modeling approach from [26] to build a messaging agent that, based upon the model's prediction about the user's state, can send an auto-responses in situations where the message recipient is predicted to be unavailable [27].

These auto-responses can share contextual information such as location, activity, device usage, and environmental information (e.g., noise, light), based upon which features significantly impact the model's prediction. Ideally, such auto-responses provide context that can help manage message sender expectations of response and reduce the pressure on message recipients to attend to messages when they are otherwise engaged. A sample auto-response from Jain et al. and the modeling process is shown in Figure 1. Each new messaging session represents a prediction instance for the attentiveness model, which is trained using XGBoost algorithm [12]. Relative feature importance for each local interpretation is estimated through SHAP [40].

In their user study on this messaging agent [27], the authors reported that users found utility in the agent to reduce distractions due to messaging and, indeed, allowed them to engage less with their device. However, a lack of understanding and uncertainty about how the agent worked created some situations where users engaged more with their phones, defeating the agent's purpose. Thus, we aim to explore the design of explanations that would allow the user to benefit from the agent by understanding its behavior and reducing their effort to appropriate it. The highly proactive nature of this agent, along with the rich feature set used in modeling (58 features), makes this agent an interesting test bed for exploring the use of explanations within the context of a proactive agent. Although we chose this agent, we believe that this general interaction might apply to other applications such as proactive recommendations [9], automated schedule planning [18], conversational information retrieval [38], and automated task management [66].

4 METHODS

In this section, we describe our study design and data collection in detail.

4.1 Study Design

There were three parts to the study setup, (1) Briefing, (2) Familiarization, and (3) Design session.

4.1.1 Briefing. We set up a 15-minute video call with the participant to explain the study's purpose and a short description of the messaging agent, stating that the agent can intercept incoming messaging communications, predict availability, and send auto-responses if it predicts the user to be unavailable. Participants were provided access to the study webpage, where they could see details about data collection, the purpose of requested permissions, and the description of agent controls and settings. The researcher briefly reviewed the page with the participants and answered any questions during this session. Participants were not provided details on the machine learning aspects of the agent.

4.1.2 Familiarization. Participants installed the agent on their phones for two weeks in the familiarization phase. In the first week, the agent collected data to learn participants' messaging behavior and build a personalized attentiveness model [26]. From the second week onwards, the agent started sending auto-responses to incoming messages. Participants were alerted via email 24 hours before the agent started sending auto-responses. Participants were also asked to take notes of things that were unclear to them as they used the agent. They were told that the purpose of the notes was to guide the design session, and there were no guidelines on the content, length, or timing of these notes. The app also generated a notification at 9 pm every day from the second week onwards where participants could enter their notes for that day. However, it was not required, and participants could also email their notes before the scheduled design session.

4.1.3 Design session. The design session was typically scheduled within a week after a participant completed two weeks of familiarization. At this point, participant dyads for the design session were formed based on the availability of participants. We used a dyad collaboration in our design

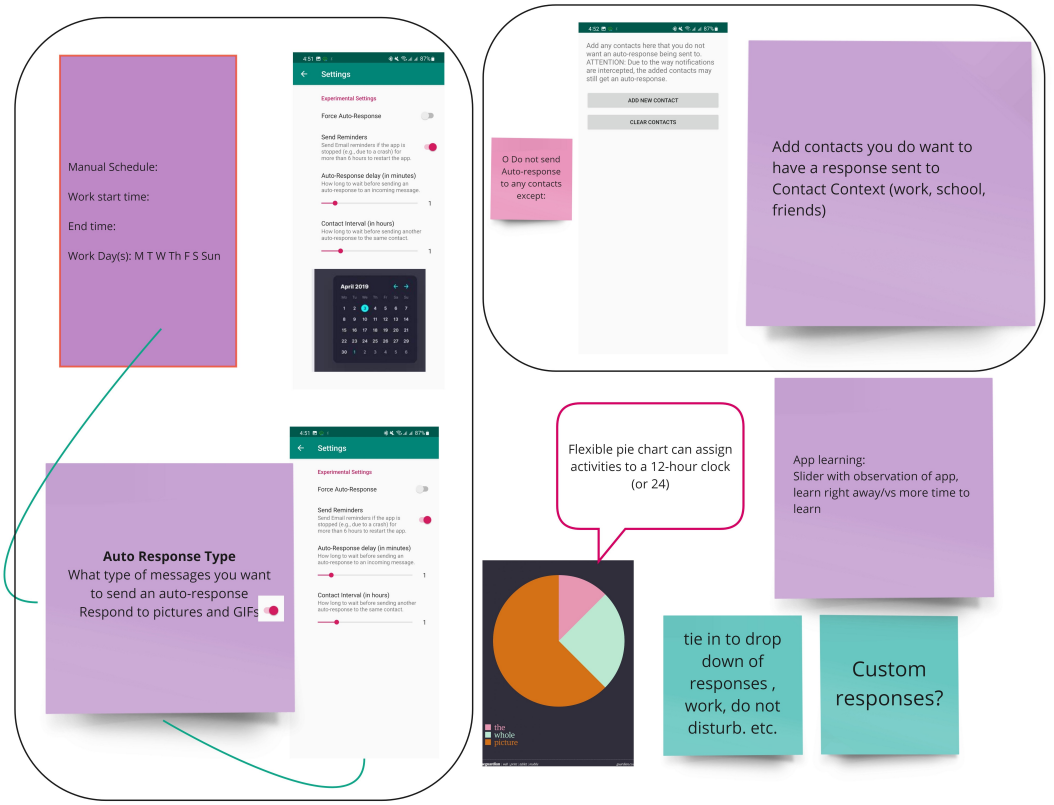


Fig. 2. The design space for Dyad F with design sketches and sticky notes at the end of the session.

sessions since discussion within pairs could bring additional viewpoints into the reasoning process while avoiding suppression of some participants' opinions in ways that might occur in larger group settings [34, 56, 64]. It is important to note that we did not pair participants based on any criteria. However, past research has shown that certain participant pairings (e.g., prior relationship, knowledge level) could affect the results of the dyadic collaboration [34]. At the end of the design session, we sent participants a survey asking them about the collaboration with their partners during the session. They were told that the responses to this survey would be confidential.

We used Miro[3] board to conduct the design session remotely. Participants connected through a Zoom call with the researcher. In this call, the researcher briefly introduced the purpose of the call. The researcher then did a brief tutorial on Miro's basic controls, including how to move around the board and create shapes, sticky notes, and sketches. Participants were also given tasks to get more familiar with Miro and ask the researcher any questions. In the design space, the top section of the board reminded participants of the existing interface and controls of the agent and was used as a reference point if they needed to refer back during the session. The bottom left of the board included the description of their two tasks and the notes they took during the familiarization phase. The board's bottom right side was the space the participants used to discuss their design ideas and thoughts. Figure 2 shows the completed design area for Dyad F for Task 1.

[3]https://miro.com/

While some participatory design studies have researchers or external entities designing while the participant discusses their requirements [50], we wanted only the participants to engage in the design activities to avoid researcher design biases in the final designs while also avoiding courtesy bias when the researcher directly interacts with the participants. Further, the researcher's involvement was minimized during the session after the Miro tutorial other than when the participants had questions for the researcher. To achieve this, the researcher turned off their camera and mic feed, but the participants knew that the researcher was on standby.

Participants were then shown the design space, including the existing screens of all the agent's features, such as blocking contacts. Participants were told to use these as reference points in their discussion if needed. The two tasks for the participants were in the middle of the design space. These were for creating designs on how they want the agent to answer (1) why an auto-response was sent; and (2) why it shared certain information in an auto-response. Participants were given 1 hour to work on the two tasks. There were no limits on how much time they could spend on each task, but the researcher did remind the participants of the time if they spent more than 40 minutes on the first task. For each task, selected participant notes sent before the design session were used to guide their discussions.

4.1.4 Pilot. We conducted a pilot session with two participants to assess the study design. Initially, we had one more task besides those mentioned in the previous section. This task was designing explanations for agent data collection practices and permissions. We removed this task to give more time for participants to work on Task 1 and Task 2, as we noticed that in the pilot session, participants already discussed data collection and privacy in the first two tasks. Further, we initially set a hard time limit of 25 minutes for each task. We noticed that interrupting participants in the middle of the session broke off their chain of thoughts, decreased their engagement in the next task, and forced them to rush through the tasks and frequently check the time. We removed this time limit and only reminded participants if they went over 40 minutes into Task 1. This study's results did not include data from the pilot session participants.

4.2 Analysis

All the 90-minute design sessions were audio-video recorded with the participant's consent. We used the built-in transcription of the video conferencing software to transcribe the recorded audio and manually fixed any errors. We performed Inductive Thematic Analysis on the audio/video transcripts [16]. We used Nvivo to structure and categorize all codes[4]. Initially, we identified 77 low-level labels such as 'Creating rules for the agent' and 'Speculating factors for prediction'. Through multiple rounds of discussion between the research team and revising the coding schema, we categorized these initial labels into 36 higher-level codes such as 'Teaching Mechanism: Rules' and 'Teaching Mechanism: Feedback'. These high-level codes informed our four major themes, which we will discuss in detail in the Results section (Section 5).

4.3 Participants

We recruited our participants for this study through a university-maintained registry of participants. Participants were paid 50 USD for completing the study. The screening process for participants involved validating their Android OS version and whether they actively used messaging on their phones. We recruited 17 participants for the study between April to July 2022. Out of the 17, three participants faced technical dificulties and could not complete the study, and their data was not included in our analysis.

[4]https://www.qsrinternational.com/nvivo-qualitative-data-analysis-software/home

Regarding participant demographics, we had four participants who identified as Male, nine as Female, and one as Non-binary. Participant ages ranged from 19 to 63, with an average of 37.5 years and a median of 32 years. Besides the gender imbalance, our participant sample was fairly well distributed in terms of age, education/major, and employment. Although, due to the restrictions associated with qualitative studies, our results may not represent the general population. In the Results section, we will be referring to individual participants using their Dyad (A-G) and number (1 or 2), e.g., F1.

4.4 Ethical considerations

This study was approved by our University's Institutional Review Board (IRB). We were transparent regarding all the permissions and data collection that the Android app required. Participants had access to the data collected by the app on their phones. They were also provided access to the study webpage, which detailed the collected data and how the permissions were used. Information regarding participants' contacts and text message contents was not collected. Information on the participants' contacts was stored locally on their devices for messaging session identification purposes.

5 RESULTS

To assess participants' perceptions of the collaborative design sessions, we asked them to fill out a survey at the end of the session. Participant responses were overall very positive. On the five-point Likert scale, they responded to the following questions, I feel that my opinion mattered and was incorporated into the design: ($\square$ = 4.79, $\square$ = 0.41), I feel that my partner's opinion mattered and was incorporated into the design: ($\square$ = 4.50, $\square$ = 0.63), I feel that the collaboration with my partner improved my designs: ($\square$ = 4.64, $\square$ = 0.72).

Our thematic analysis uncovered four key themes: (1) Exposure and observations of agent actions trigger reasoning about factors in its decisions (Section 5.1); (2) Curiosity about unexpected agent behavior motivated the desire to update initial mental models (Section 5.2); (3) Observations of agent actions and dyad interactions can support learning about the agent (Section 5.3); and (4) Users can strengthen agents' predictive models with rule-based heuristics (Section 5.4). We now explore each of these themes and their interrelations (Section 5.5).

5.1 Exposure and observations of agent actions triggers reasoning about factors in its decisions (RQ1)

As expected based on our study design decision, the two weeks of familiarization and use of the agent inspired participant reasoning about the agent's behavior and speculation about the agent's design. Next, we will discuss some common triggers of these speculations and how participants tried to identify factors that informed the agent's decisions.

5.1.1 Observing the agent and prior experience with technology triggered participants' speculations.

Four dyads recalled their prior experience with other technologies when reasoning about how the messaging agent worked. For instance, G1 incorrectly speculated that the agent might be using the camera or accessing stored pictures since their phone showed a privacy warning of the camera being used, which the Android OS typically shows to improve awareness of when sensor data such as GPS, microphone, and camera are being accessed, "I feel like the app knew when I was taking a picture because I would see a camera icon at the top of my screen. If I'm taking pictures of my kids, I don't want that stuff to be stored somewhere. You just don't want your personal information getting out". Similarly, F1 incorrectly speculated that the agent was using the content of text messages to predict availability because of their prior experience with personalized advertisements based on

past search queries, "I think, maybe it picks up on certain words when we send a text message. Or, you know, any type of message, that's what I'm just thinking, kind of like if you're doing a search on Google".

5.1.2 Participants tried to reason about what factors could influence agent's decision-making. On multiple occasions, participants expressed an understanding of the connection between smartphone sensors and the agent's behavior. For instance, D1 correctly speculated that the light sensor on the phone is being used for determining the ambient light since it is also typically used for adjusting the phone brightness automatically, "it has obviously that sensor where it senses like brightness and everything, so if it senses darkness, it sends that message, your phone is in a darkly lit area, which (it) usually is, so uses that in its explanation". On the other hand, F2 incorrectly inferred, based on the agent's requested permissions and how the agent was using the microphone for noise detection, that the agent could also be accessing the camera to determine the light levels in the surrounding area, "I understand, based on the permissions and knowing that the phone was capturing an audio recording of what the situation was, I'm guessing, similarly, if they're using our phone cameras to see the lit area, (otherwise) how would it know that it's in my pocket?...".

Participants also speculated that agent decisions are based on multiple factors rather than a single feature (Dyads A, E, and F). For instance, F2 stated that the agent is using multiple sensors in the phone, "It's a sensory input of like how much noise, how if it's dark or light, or whatever captured sensory information, data from phone use, to then reuse in auto-response".

5.2 Curiosity about unexpected agent behavior motivated the desire to update initial mental models (RQ2)

The tasks given to participants during the design sessions included two prompts: (1) why did or did not the agent send an auto-response?; and (2) why it shared a specific context as part of the auto-response? In thinking about explanations, participants expressed curiosity, particularly about unexpected agent behavior; and how their behavior affected agent outcomes.

5.2.1 Agent action. Multiple participants (six dyads) expressed curiosity about how the agent decided what information to share in an auto-response. For instance, A2 mentioned that the correlation between their unavailability and the context shared by the agent was unclear to them, "To me, reading the messages, I understood why it sent the message (auto-response) because obviously, it explains it very specifically in there, but not why it chose to send it because of that". Participants also desired clarifications for the agent's logic in classifying their state shared in the auto-response. For instance, F1 questioned why the agent thought they were 'not receptive to communication', "To people that I normally talk to, and I respond back to them within probably, I don't know, five or six minutes, and it came up with a response, saying that I'm not receptive to communications".

5.2.2 Agent inaction. Participants also wanted clarifications when the agent did not send an expected auto-response. C1 noted that even though they labeled their work location in the app, the agent did not auto-respond when they were at work, "I've been having this problem throughout the whole experience. Auto-responses were not being sent out, even though I was at work, and it kind of ignored my location". This led to privacy concerns and a lack of trust later on in the session, where they questioned the collection of location data, "if I was just someone who was using the app for the first time and I had to put down my location, and they said well, this is to specify your location, and then I get no messages specifying the location, it's kind of shady". Similarly, E2 wanted clarification on how the agent was factoring in contact information in its decisions, as they noticed auto-responses only being sent frequently to a select few contacts, "Mine just some friends and

family it responded to, and others it didn't, and I don't know, maybe it was the time of day. How it determined I happen to be available, I can't figure that out".

5.2.3 Effect of user action. Participants also wanted to know how their actions affected the agent's behavior. For instance, B1 mentioned that they lowered the delay setting in the app to increase the frequency of auto-responses but without success. C1 mentioned a similar experience, "...I even ended up changing my settings too. I lowered the (delay), I set it to 0, and then also the what was the other one I forgot, oh the interval. But it didn't make a big difference. It was still not sending the auto-responses, even when I was at practice or at work for a couple of hours". In addition to trying to understand the impact of adjusting settings, participants were also curious about how their device usage might affect agent actions. For instance, B1 mentioned that they would have liked to know how using their phone affects the agent's decision whether to auto-respond, "I was just curious if I'm doing something on my phone, will it still send out a message?". Similarly, E2 wanted to know after how long of not using their phone the agent would send an auto-response, "How long must I not be using the phone for [the agent] to generate auto-responses? For instance, if I have not used my phone in an hour, 2 hours, 5 hours, 24 hours".

5.3 Observations of agent actions and dyad interactions can support learning about the agent (RQ3)

We observed multiple instances where participants indicated an improved understanding of how the agent worked through either repeated interactions with the agent or by interacting with their partner during the design session.

5.3.1 Learning through repeated observations of agent behavior. As participants interacted more with the agent, they showed an increased understanding of how it worked. For instance, C2 noted that agent responses started to improve over time, "As it is gathering more data, I guess it became more clear, and it provided some information as to why I may not respond. At first, it was just saying she might not respond. Okay, but then it would say, because she's not usually active on the phone at this time of day, or because she's in a silent environment, which I thought was funny, or the phone is in my pocket or something. It's started to make more sense the more data it gathered". It is worth noting that the participants were not informed that the agent model was retrained every day. In another similar case, repeated observations led D1 to infer that location was not a major factor in any of the agent's decisions, "I put down my dance studio for the locations so that it gives an explanation for when I'm at practice, but no auto-responses were sent when I received texts at the studio". They recalled another instance when they were at their work location, "[redacted] and [redacted] both texted me while I was at work, and no auto-response was sent. The location doesn't seem to influence the auto-sender...".

5.3.2 Learning about the agent through dyad interactions. Participants often exchanged knowledge when discussing their experience with the agent during design sessions. In some cases, a participant expressed an issue with the agent's behavior, and their partner suggested a solution. For instance, C2 recalled an issue with a high frequency of auto-responses being sent for them even when they were available to respond. Their partner, C1, asked them whether they changed the agent settings (i.e., delay and interval), to which they responded that they did not and agreed that it might have helped. Dyad D had an exchange where D2 discussed wanting controls to prevent an auto-response. D1 shared their experience with D2 that opening the incoming message can prevent an auto-response, "The agent doesn't respond when it sees that you opened up a message. If it's a message that you read, it won't respond to it. If you haven't read it, no matter what the platform is, it's going to respond".

Since the agent could share multiple categories of auto-responses depending on what it learned, some participants did not experience all auto-response types. There were multiple cases where one participant learned about a particular category of auto-response from their partner. For instance, C1 learned that the agent could also share ambient noise from their partner and discussed potential reasons why it wasn't shared for them, "I didn't know (about noisy environment auto-response), I usually have a silent environment at work, so that didn't always work". In another case, Dyad G disagreed about specific information that the agent shared. They had two exchanges regarding two different shared contexts. In the first one, G1 presented a scenario to G2 where environmental information such as surrounding noise level could be useful.

> G2: "Right and not share information on your environment at all, like, you know, the low light area."
> G1: "I don't know, somebody's phone is in a noisy environment, I mean, I think that that's okay, what if you were at a concert or something like that and obviously, you're not really going to respond if you are seeing the live concert, so I think that's a good response."
> G2: "See, that wouldn't be my choice because if I was the person receiving that, I would be like, So what's that got to do with responding to my text? Why am I getting this text response?"

In the second one, G1 again reasoned how sharing proximity sensor value (device in bag or pocket) could be a valid reason for unavailability. G1 seemingly convinced G2 to the second scenario but not the first one.

> G2: "I think my friends will be like, well, what does that mean it's in a bag, or it's in your pocket. Good! it's in your pocket, (now) respond."
> G1: "I guess it's, you know, look, my phone's usually on vibrate or silent, so like I can't hear it anyway."
> G2: "yeah, good point."

5.4 Users can strengthen agents' predictive models with rule-based heuristics (RQ3)

Participants discussed various situations where they could teach the agent their preferences. C1 emphasized the importance of incorporating user feedback in the agent design as they felt that past behavior is not always indicative of their future actions, "I don't think the past, maybe past ways of using the app, are a good way of predicting what the future actions will be because people's schedules change, and it happens pretty quickly". Towards that, they also discussed the need for a supporting interface to provide them with control options to be able to teach the agent their preferences as a set of rules.

All seven dyads had discussions where they felt that predicting the user's state in certain situations was unnecessary. For instance, Dyad D discussed that the agent did not need to use prediction at certain times during the day and could have a fixed behavior at those times, "D2: (Add a) sleeping option, like, I guess, if we were to set those general parameters and say between 11 pm and 6 am if anybody sends me a message like we can make it as quirky or funny as you want, and say something along the lines of [redacted]'s catching her z's". Similarly, A1 mentioned wanting an emergency mode for their specific work-related situations, which required them to answer texts on their phone during certain times. They indicated that the agent does not need to auto-respond when this mode is turned on, "A1: Maybe something like emergency mode? In which we press that, and then all the messages of the agent (are) stopped".

Instead of completely turning the agent function on or off, participants also discussed teaching the agent to account for specific user context to determine the best course of action. For instance, D1 discussed wanting the agent to always auto-respond when they were at work, "If I was at the

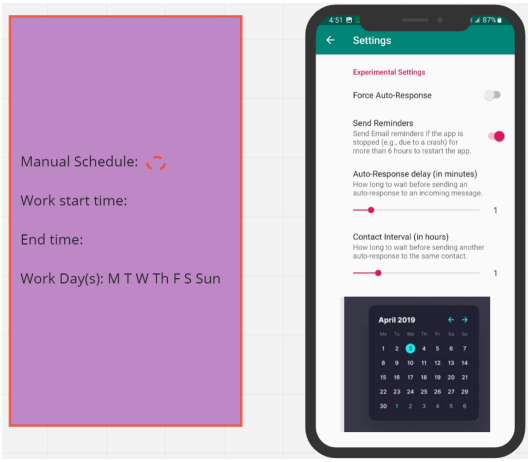


Fig. 3. The design sketch by Dyad E to manually enter the user schedule to assist the agent in its predictions. Selecting a date on the calendar opens a new screen where the user can set their schedule.

hospital, I want it to learn that when I'm at the hospital or at this location that I've labeled hospital, I want you to respond that I'm working or busy. That seven-day learning period would be the time to teach it the locations and where you're usually at and allow you to check or uncheck certain phrases at different locations, just kind of get (it) to know your routine a little bit". Similarly, Dyad E discussed wanting the agent to learn their schedule and account for it in its decisions for what responses it shared and designed an interface as shown in Figure 3, "Put in one's work schedule and have maybe a way to differentiate how other responses are done during work versus non-work times".

In addition to teaching the agent about schedules and locations, six of the seven dyads emphasized wanting to teach the agent how to handle different contact types. For instance, Dyad A discussed wanting to have different agent behavior based on the type of contact, "A2: We could have different categories of responses that they could send out, you could send to my kid, less formal language, less specific language and to my employer, more formal and more specific messages". They suggested categorizing contacts during the initial training phase of the agent, "...whenever you're setting them up in the beginning, you can categorize as a specific thing like personal or business, then that way, you don't have to feel compelled to customize each individual person right off the bat unless you want to". Dyad B discussed wanting the agent to instead automatically gain additional context about how frequently they interact with different contacts and use that information to determine how much information to share, "If you send one text message to one person a day, then you probably just get the response of, "[redacted] is not available at this time". But maybe if the system's able to see that it is your mom or somebody like that and you message this person 100 times a day, they get a more in-depth response in terms of, "[redacted] is not available. He hasn't been on his phone in a while". The more frequency of text messages, the more in-depth it is, (the) less frequent, the less in-depth the auto-response will be".

5.5 Interaction with the agent and speculations about agent design create pathways towards learning about and teaching the agent (RQ1, RQ2, RQ3)

Figure 4 shows the four key concepts extracted from our five major themes in our qualitative data and how the participants transitioned between these concepts in the design sessions. These concepts particularly represent the pathway to learning from and teaching the agent. We created

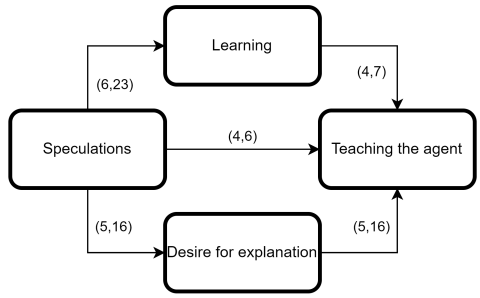


Fig. 4. Four main concepts in the findings (Learning, Speculations, Desire for Explanations, and Teaching the agent) and how they are connected in the analysis. The first number represents unique dyads that transitioned from the source concept to the target concept. The second number represents the total number of times that transition happened in any discussion.

this transition visualization by coding these concepts in the participants' discussion and what followed each concept as they continued their discussions. Visualizing the transitions provides insights into how each concept relates to the other and can inform the design of agents in facilitating the initiation of each concept and transition to the desired outcome of learning and teaching. We observed that participants always started the discussion (start node) with speculations (Section 5.1), and teaching to the agent (Section 5.4) was always the end point of the discussion (end node). Below, we explain each transition in more detail.

5.5.1 From speculation to learning, desire for explanation, and teaching. As seen in Figure 4, speculations were often followed by participants learning about the agent's behavior. However, it also leads to a desire for further explanation and an opportunity to teach the agent.

Speculations to Learning (6 dyads, 23 references): As discussed in Section 5.1, speculations emerged when participants tried to guess various agent behavior. These speculations were confirmed or rejected with continued agent use during the familiarization phase. This helped participants to transition from speculation to a learning experience as they tried to confirm or reject the agent's behavior. For instance, E2 mentioned that they initially noticed that the agent did not respond to messages on Google Voice and thought it was unsupported. However, later, when the agent did eventually respond to a Google Voice message, they concluded that there could have been another reason for the lack of earlier auto-response, "I wasn't sure if it was going to (respond), for whatever reason, the first day it didn't with Google voice, and the second day it did. I guess it was just it thought I was available after what it had learned over the seven days".

In another exchange, E1 described experimenting with the agent to understand how the agent learns their availability and context to share in the auto-response, "I was doing work activities from a sort of novel location, and I did mark those in the app as, this town work that town work, so I think it got an idea from that oh, it's the middle of the day, I'm usually working. (It sent) I may not be able to respond, she is usually less responsive this time of day". To confirm whether the agent has learned this schedule, they tried to replicate this behavior, "I asked my partner to message me to see what would happen, and the agent did respond with commentary that I'm usually busy at that time of day, so it had learned the time I was often working".

Speculations to Desire for explanation (5 dyads, 16 references): There were multiple instances where participants speculated about the agent's behavior and then transitioned to wanting an explanation to assess their speculation. For instance, C2 speculated that the agent detected that their phone is connected to their car's Bluetooth and sent an auto-response due to it, and wanted

to confirm if that is the case, "...I wasn't busy, but maybe it thought I was because I was connected to Bluetooth to the car. I don't know how I would know this".

Speculating about the agents' design and actions also created expectations of a particular behavior. When these expectations were not met, participants expressed a desire for an explanation from the agent. For instance, E2 recalled expecting the agent to share their ambient noise level as it previously had, "It was a quiet environment (day before), but then last night I was at a concert, the auto-response was sent, but it didn't say anything about being in a loud environment where I didn't hear the phone, why didn't it say I was in a loud environment?".

Speculations to Teaching the agent (4 dyads, 6 references): Multiple dyads discussed wanting to influence the agent's behavior based on their speculations of how it worked. For instance, B1 incorrectly speculated that the agent does not send auto-responses to every contact; instead, there might be an order for how many and to whom the auto-responses are sent. They then suggested that the agent could prompt the user about contacts and how frequently auto-responses are sent, "...ask the question, like, do you want this auto-response to go to every message? Or every contact? Or do you want this to go out to every third contact? Every fifth contact? Does that make sense? I guess the frequency in which it is being sent out". Speculations about factors used by the agent to determine availability also transitioned to participants desiring control to influence agents' decision-making based on those factors. E1 incorrectly speculated that the app uses the content of the messages when deciding whether to send an auto-response and wanted control to overturn that agent's behavior, "So my boyfriend went for a hike, and he texted me some pretty pictures of nature, and I wasn't paying attention to my phone, and it (agent) didn't say anything to (the) pictures which to me is not that big a deal, but unless he was really trying to get in touch with me, it might be so".

Participants also discussed methods to improve the context sharing of the agent based on their speculations. Dyad A correctly speculated that the agents' prediction would be approximate. They described wanting to set up rules to be able to alter the decision on what context to share based on how confident the agent was for that prediction, "it's like you said, 70% confidence (for a prediction) you'll maybe alter that to say, well only send this part (context) out if it's, you know, 90% confident or something like that". In Section 5.5.1, E2 discussed that their speculation about the agent prioritizing noise levels in its context sharing did not hold. E1 speculated that it is possible that the agent did not have correct calibration for detecting louder noises and suggested controls to teach the agent about different noise levels, "I'm thinking, the app has been learning in the background without us interacting with it so maybe there's a place where we actively try to teach it like go stand next to something noisy. Have a mode where you manually teach it something like, this is too loud that I wouldn't want to converse there".

5.5.2 From Learning about the agent to Teaching the agent (4 dyads, 7 references). Increased understanding of the agent, either through repeated observations or from interacting with the partner, not only helped participants learn about the agent's behavior but also resulted in participants desiring more appropriate controls to teach the agent their desired behavior than those based on early speculations. D1, through repeated observations of agent behavior, concluded that the agent was not factoring location into its decisions. They wanted to teach the agent to emphasize location in its decisions and context sharing, "I really don't have a lot of time to look at my phone (at work). Just having it recognize my location and saying that specifically".

G2 discussed their experience with auto-responses being sent out even when they were actively using their phone. This was in contrast to the agent's behavior for G1, for whom the agent did not respond when they were using their phone. G2's conclusion from this conversation was that the agent learned it from observing their behavior of purposefully being unresponsive to some messages, "I guess it determined that there are times when I'm on my phone that I don't respond to text

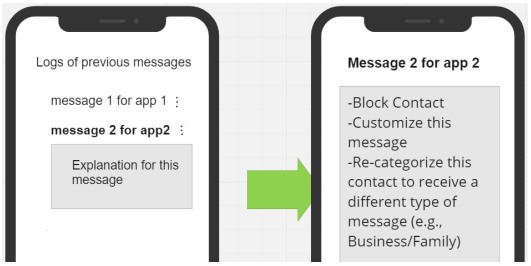


Fig. 5. Design suggestion for actionable explanations by Dyad A.

messages, but what it doesn't know, the app doesn't know is, I'm not responding to that text message because it was a spam or it was a solicitation for funds for some political campaign or whatever the case may be, and that's why I'm not responding to the text". This prompted G2 to desire controls to overwrite what the agent had learned about their responsiveness when actively using their phone, "If I'm on my phone watching a video, maybe I should be able to say you can do that anytime except for when I'm watching a video. Don't send an auto-response unless my phone's inactive".

5.5.3 From Desiring explanation to Teaching the agent (5 dyads, 16 references). While participants discussed wanting explanations for unexpected agent behavior (Section 5.2), their end goal with these explanations was to make the agent conform better to their expectations. For instance, A1 indicated that just getting an explanation (knowing the "why") is not enough; instead, they would also like to have controls to appropriate the agent. "I always want to know "why" because I think knowing "why" would help me make the decision. Knowing "why" it said those things is helpful, but from a user standpoint, knowing "why" doesn't necessarily change; it's not going to affect me, as far as the end result is concerned. I could know "why" all day, but if I don't want it to do that, how do I make it stop doing that?".

Regarding how participants wanted to affect change following an explanation, Dyad C discussed wanting explanations of the factors the agent used in its decisions and altering how those factors are used, ""She has not been receptive to communication for two hours", What kind of communication is that? If I'm only checking my texts and I'm not checking my Facebook, or whatever, what's communication, I guess? I feel like if you're actively using the messaging app, that should override any kind of previous data, maybe that it had collected about your habits or patterns".

Further, regarding how participants wanted to teach the agent following an explanation, Dyad D suggested feedback to the agent should be part of the explanation, "So if you click thumbs down for that (explanation), the next thing it would say is, okay, what would you like me to respond with, or how would you like me to respond when the phone brightness is low in a room or something? It should give you an option to improve or a way to improve on how it's responding". Dyad A had a similar discussion of providing controls within the explanation to alter agent behavior. Their design suggestion is shown in Figure 5 where upon selecting an explanation from the list, the agent shows a list of actions that the user can take for that specific instance of auto-response, such as blocking that contact or customizing the text for future auto-responses to that contact.

6 DISCUSSION

Using a participatory design approach, we studied how dyads discussed their desired explanations from a messaging agent. We observed that participants tried to collect evidence by observing the agent's action and often linked it to their prior experience to build their initial mental models. Participants were motivated to update their understanding of the agent's actions when its actual

behavior did not reflect their mental model. Dyadic interactions and repeated agent observations supported participants in reflecting on and learning more about the agent's behavior. This learning helped participants build confidence about the agents' behavior and led them to propose additional controls to manage agent outcomes better.

The design objectives of this agent make it unique compared to other intelligent agents (e.g., recommender systems or smart voice assistants) that individuals may interact with regularly in three ways: (1) the proactive nature of the agent means that it takes action without user intervention; (2) the agent acts as an intermediary in human-human communication as opposed to human-agent interaction in cases such as voice assistants, which can potentially affect existing interpersonal relationships; and (3) users have the ground truth to evaluate the accuracy of the agent's behavior as the agent's objective is to ascribe a reason for their unavailability that they are well aware of. We contextualize our findings through these three dimensions to explore design implications for future messaging agent systems.

6.1 Adaptable proactive agent design

We designed the messaging agent to be fully autonomous to reduce distractions associated with mobile messaging [27]. Our participants described multiple situations where they desired a specific behavior from the agent based on particular contexts (Section 5.4). These contexts were, for instance, time of day, location, and contact type. This suggests that depending on the agent's task and user-specific context, the agent's autonomy level can be made to vary. Proactive agents can start at a lower level of autonomy, such as proactive suggestions (level 5 autonomy [33, 52]), where instead of acting on predictions, they provide suggestions to the user while also supporting the user's inputs. For instance, the messaging agent could generate auto-responses and prompt the user to rate the responses instead of sending them or support one-click responses to be sent by users (Section 5.5.3). As the agent learns user preferences over time, its level of autonomy can increase, where it can send auto-responses automatically. Another approach towards adaptable proactivity could be based on the agent's confidence in its predictions (Section 5.5.1).

6.2 Understanding and augmenting social norms in agent-mediated interactions

In human-human conversations, people follow social norms such as being cooperative and polite [59]. For conversations to be natural and easy to follow, Grice described four categories under the Cooperative principle – quantity (making your contributions informative without excess information), quality (contributions should be true), relation (be relevant), and manner (avoid obscurity, ambiguity, prolixity; and be orderly) [21]. As the messaging agent acts as an intermediary in human-human communication, it can potentially disrupt the social norms of human-human communication. We observed indications of the four elements of the aforementioned principle in our participants' discussion as their desirable behavior of the agent. For instance, participants questioned the quantity – e.g., multiple dyads discussed how the amount of information the agent shares should be adopted based on the specific relationship with their contacts (Section 5.4). In terms of quality, Dyad A, for example, discussed that the agent should be confident in its prediction before sharing a context to avoid inaccurate disclosure (Section 5.5.1). Concerning relation or relevance, multiple dyads questioned the relevance of context the agent shared, such as why ambient light or phone proximity qualifies as relevant information about their availability to respond (Section 5.3.2). Manner was particularly highlighted in terms of avoiding ambiguity, for example, how an auto-response of 'not receptive to communication' is unclear and can be very vague to the recipient (Section 5.2.1).

This disruption in social norms of communication by the agent can increase the users' effort to justify agent actions to their contacts [27]. During the design sessions, participants described

various ways such as providing feedback to the agent (Section 5.5.3) concerning the relevance of the shared context in a given situation, setting up rules to better control agent outcomes to limit the quantity and improving the quality of information shared by the agent (Section 5.4) and finally the participants also discussed and negotiated appropriate agent behavior concerning information the agent shared trying to reduce the ambiguity associated with agent responses (manner) (Sections 5.3.2 and 5.5.3). Future guided co-design sessions focusing on understanding social norms and dynamics related to agent information sharing could be helpful for agent designers to more effectively design agent behavior and phrase agent responses to adhere to the socially acceptable behavior for a virtual intermediary in conversations. These sessions could also help establish knowledge for the agent for using certain justifications such as 'not receptive to communication' (Section 5.5.3), where participants indicated a mismatch between their expectation of what constitutes 'not being receptive to communication' compared to what the agent was coded with.

6.3 Leveraging user expertise towards desired behavior

6.3.1 *Opportunities for users to learn.* Multiple participants indicated an understanding and knowledge of sensor data and its uses which they acquired through the experience of using their different smartphone applications (section 5.1.2). Participants used permissions the agent asked for at the time of installation and their experience with smartphone apps to make informed guesses about the agent. While this knowledge inspires speculations, we also observed instances where participants incorrectly speculated about the agent's functions based on these prior experiences (Section 5.5.1). These misinterpreted speculations sometimes lead to desiring unnecessary controls for agent behavior. Thus, guiding users' speculations is necessary to avoid unintended consequences and user disappointment. One way to direct speculations into accurate mental models is to provide mechanistic explanations [23] early in the use of the agent. These explanations can focus on describing the agent's decision-making engine instead of the agents' actions. For example, describe how the agent identifies a new messaging session or samples various sensor data.

6.3.2 *Opportunities to learn from the user.* The agent learns from user interaction and sensor data to form a user model based on patterns of the users' attentiveness to messaging [25]. The agent then uses this model to predict user behavior and construct an auto-response to share with the users' contacts. Unlike many predictive models (e.g., recommender systems) that assist individuals in gaining information, in this case, the users of the agent have the ground truth about what the agent is attempting to predict. They are well aware of the reasons for not responding to a message. In other words, they are the "experts" on their messaging behavior. Therefore, the agent's justification may not always match user expectations. The agent is correlating the inattentive state with the features used in the user model instead of identifying the cause of unavailability. Further, the agent's information is limited to environmental and usage data that can be captured through smartphone sensors and user interaction. For instance, the agent cannot detect the sleeping state with complete certainty [36, 45]. Finally, even if the agent's explanation is perfectly accurate, the user may not find it appropriate to share with specific contacts (Section 5.3.2). This user expertise in their messaging behavior provides an opportunity for the agent to learn user preferences. Our participants indicated a willingness to provide feedback to the agent as part of explanation interfaces. For instance, building quick feedback mechanisms such as thumbs up or down into the explanation (Section 5.5.3), customizing the content of the auto-response, and automatically triggering auto-responses based on specific contexts (Section 5.4). While participants indicated wanting to give feedback, it is unclear how frequently and for how long they would be willing to do so. Further research is needed to discover effective interface designs to adapt better couple human and agent inputs.

6.3.3 Community-based knowledge exchange. In Section 5.3.2, we discussed that dyadic interactions promoted improved learning about the agent. Through exchanging different experiences, dyads were able to discuss and discover more about the agent's behavior. Thus, we posit that integrating user-user interaction into the explanation interface can promote further learning about the agent. However, it is also vital that these discussions do not reinforce misinterpretations of the agent's behavior. Therefore, designers should consider including guided community discussions within the scope of agent applications, allowing users to share their experiences and engage in knowledge exchange without falling into misinterpretation pits. Further research is needed to understand how we can create designs to facilitate user-user interactions related to experience-based knowledge exchange.

6.3.4 Engage with User curiosity. In terms of user-agent interaction, prior work has emphasized that explanations must be interactive to be more engaging [42, 44]. Allowing users to ask follow-up questions is one way of driving more natural interactions between humans and AI [39]. We also observed in Section 5.2 that curiosity was a motivating factor in the participants' desire for explanations. Indeed, prior work has reported that human curiosity is a powerful motivator for exploration to reduce uncertainty and lead to learning [5]. Designing agents which can spark user curiosity can enable interactions from which the user can ask questions from the AI and learn more about it. Further, other factors, such as anthropomorphic agent features, may allow users to perceive AI more as an entity with which they can have conversations [20]. Future work is needed to identify what factors could effectively enable human-like interactions between humans and AI to allow learning about the agent.

6.4 Limitations and Future Work

There is a tendency to passively accept others' opinions in group-based discussions [63]. While we attempted to minimize this by designing our process to form dyads instead of bigger groups, this may have still affected the study participants' designs and discussions. Further, typical participatory design research limitations also apply to this study. Only a small group of participants were involved in the design process, which may not reflect the broader population's opinions [54]. It is also important to note that participant pairings were random and based on their availability, and the participants did not share any relationship. Prior research has shown that dyad pairing with a prior relationship can lead to a broader exploration of topics and could be more effective in terms of information exploration than pairing strangers [34]. This can be particularly relevant for a messaging agent, where as discussed earlier, social norms in interpersonal communication are important to consider in designing this type of agent. Finally, since the interface design suggestions discussed by our participants are early stage, further research is needed to evaluate user engagement with these interfaces and their effectiveness in improving user understanding of agent functions.

6.4.1 Investigating user-agent interaction from the perspective of a non-agent owner. User-agent interactions are not always isolated. It may involve other people in the vicinity of the agent, like in cases where the agent uses sensors or actuators in its functional environment [7]. In the case of the messaging agent, there is the direct involvement of the non-agent owner, where they are on the receiving end of the agent's explanation or justification for the agent owner's unavailability. Prior work on the auto-response messaging agent reported that agent interactions with non-agent owners affected agent owner perceptions of agent utility in certain situations and agent owner's behavior and engagement with agent controls [27]. With more smart-home systems and intelligent agents integrated into our daily lives, bystander privacy has recently been an active area of research [3, 7]. While in this work, we did not explore the non-owners perspective related to agent explanations; there is potential for further exploration related to how we can adapt the

agent models to account for non-owners understanding and interpretations of agent explanations. Investigating these perspectives can help designers tailor agent interactions to be more considerate of non-owners preferences while maintaining utility for agent owners.

7 CONCLUSION

Virtual assistants are being increasingly integrated into technologies we interact with in our daily lives. From automated temperature controls in HVACs to the automated ordering of groceries, these agents are taking on more proactive tasks. The convenience these systems hope to add for their users can come at the cost of perceived control over the automated actions of these agents. Through a co-design study, we explored how users reasoned about the agent function based on their observations and prior experience with data-driven technologies to form initial mental models. These user mental models created an expectation of specific behavior from the agent, which, if not met, created a desire for explanations and influenced their need for controls to inform the outcomes of the agent. These results reinforce the importance of delivering focused explanations of the agent's behavior during early interactions to develop accurate user mental models of the agent's behavior. Our study on the messaging agent further necessitates the importance of understanding and accounting for social norms in the agent's behavior due to its role as a mediator in human-human communication.

ACKNOWLEDGMENTS

This work was supported in part by the National Science Foundation under award CNS–1814866. We would also like to thank our participants for their contribution to this study.

REFERENCES

[1] Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y. Lim, and Mohan Kankanhalli. 2018. Trends and Trajectories for Explainable, Accountable and Intelligible Systems: An HCI Research Agenda. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (Montreal QC, Canada) (CHI '18). Association for Computing Machinery, New York, NY, USA, 1–18. https://doi.org/10.1145/3173574.3174156

[2] Ankit Agrawal and Jane Cleland-Huang. 2021. Explaining Autonomous Decisions in Swarms of Human-on-the-Loop Small Unmanned Aerial Systems. Proceedings of the AAAI Conference on Human Computation and Crowdsourcing 9, 1 (Oct. 2021), 15–26. https://ojs.aaai.org/index.php/HCOMP/article/view/18936

[3] Imtiaz Ahmad, Rosta Farzan, Apu Kapadia, and Adam J Lee. 2020. Tangible privacy: Towards user-centric sensor designs for bystander privacy. Proceedings of the ACM on Human-Computer Interaction 4, CSCW2 (2020), 1–28.

[4] Julio Angulo, Simone Fischer-Hübner, Erik Wästlund, and Tobias Pulls. 2012. Towards usable privacy policy display and management. Information Management & Computer Security 20 (2012), 4–17.

[5] Marilyn P Arnone, Ruth V Small, Sarah A Chauncey, and H Patricia McKenna. 2011. Curiosity, interest and engagement in technology-pervasive learning environments: a new research agenda. Educational Technology Research and Development 59, 2 (2011), 181–198.

[6] Daniel Avrahami and Scott E. Hudson. 2004. QnA: Augmenting an Instant Messaging Client to Balance User Responsiveness and Performance. In Proceedings of the 2004 ACM Conference on Computer Supported Cooperative Work (Chicago, Illinois, USA) (CSCW '04). Association for Computing Machinery, New York, NY, USA, 515–518. https://doi.org/10.1145/1031607.1031692

[7] Julia Bernd, Ruba Abu-Salma, and Alisa Frik. 2020. Bystanders' Privacy: The Perspectives of Nannies on Smart Home Surveillance. In 10th USENIX Workshop on Free and Open Communications on the Internet (FOCI 20). USENIX Association. https://www.usenix.org/conference/foci20/presentation/bernd

[8] Reuben Binns, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. 2018. 'It's Reducing a Human Being to a Percentage': Perceptions of Justice in Algorithmic Decisions. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (Montreal QC, Canada) (CHI '18). Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3173574.3173951

[9] Alexandros Bousdekis, Babis Magoutas, Dimitris Apostolou, and Gregoris Mentzas. 2015. A proactive decision making framework for condition-based maintenance. Industrial Management & Data Systems 115, 7 (2015), 1125–1250.

[10] Clara Bove, Jonathan Aigrain, Marie-Jeanne Lesot, Charles Tijus, and Marcin Detyniecki. 2022. Contextualization and Exploration of Local Feature Importance Explanations to Improve Understanding and Satisfaction of Non-Expert Users. In 27th International Conference on Intelligent User Interfaces (Helsinki, Finland) (IUI '22). Association for Computing Machinery, New York, NY, USA, 807–819. https://doi.org/10.1145/3490099.3511139

[11] Alexandra Burton, Claudia Cooper, Ayesha Dar, Lucy Mathews, and Kartikeya Tripathi. 2022. Exploring how, why and in what contexts older adults are at risk of financial cybercrime victimisation: A realist review. Experimental Gerontology 159 (2022), 111678. https://doi.org/10.1016/j.exger.2021.111678

[12] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (San Francisco, California, USA) (KDD '16). Association for Computing Machinery, New York, NY, USA, 785–794. https://doi.org/10.1145/2939672.2939785

[13] Zhenyu Chen, Mu Lin, Fanglin Chen, Nicholas D. Lane, Giuseppe Cardone, Rui Wang, Tianxing Li, Yiqiang Chen, Tanzeem Choudhury, and Andrew T. Campbell. 2013. Unobtrusive Sleep Monitoring Using Smartphones. In Proceedings of the 7th International Conference on Pervasive Computing Technologies for Healthcare (Venice, Italy) (PervasiveHealth '13). ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), Brussels, BEL, 145–152. https://doi.org/10.4108/icst.pervasivehealth.2013.252148

[14] Hyunsung Cho, Jinyoung Oh, Juho Kim, and Sung-Ju Lee. 2020. I Share, You Care: Private Status Sharing and Sender-Controlled Notifications in Mobile Instant Messaging. Proceedings of the ACM on Human-Computer Interaction 4, CSCW1 (2020), 1–25.

[15] Michael Chromik, Malin Eiband, Felicitas Buchner, Adrian Krüger, and Andreas Butz. 2021. I Think I Get Your Point, AI! The Illusion of Explanatory Depth in Explainable AI. In 26th International Conference on Intelligent User Interfaces (College Station, TX, USA) (IUI '21). Association for Computing Machinery, New York, NY, USA, 307–317. https://doi.org/10.1145/3397481.3450644

[16] Victoria Clarke and Virginia Braun. 2013. Successful qualitative research: A practical guide for beginners. Sage publications ltd. 1–400 pages.

[17] Vladan Devedzic and Danijela Radovic. 1999. A framework for building intelligent manufacturing systems. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) 29, 3 (1999), 422–439.

[18] Yagil Engel and Opher Etzion. 2011. Towards Proactive Event-Driven Computing. In Proceedings of the 5th ACM International Conference on Distributed Event-Based System (New York, New York, USA) (DEBS '11). Association for Computing Machinery, New York, NY, USA, 125–136. https://doi.org/10.1145/2002259.2002279

[19] Ethan Fast and Eric Horvitz. 2017. Long-Term Trends in the Public Perception of Artificial Intelligence. In Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (San Francisco, California, USA) (AAAI'17). AAAI Press, 963–969.

[20] Eun Go and S Shyam Sundar. 2019. Humanizing chatbots: The effects of visual, identity and conversational cues on humanness perceptions. Computers in Human Behavior 97 (2019), 304–316.

[21] Herbert P Grice. 1975. Logic and conversation. In Speech acts. Brill, 41–58.

[22] 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.

[23] Steven R Haynes, Mark A Cohen, and Frank E Ritter. 2009. Designs for explaining intelligent agents. International Journal of Human-Computer Studies 67, 1 (2009), 90–110.

[24] Jonathan L. Herlocker, Joseph A. Konstan, and John Riedl. 2000. Explaining Collaborative Filtering Recommendations. In Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work (Philadelphia, Pennsylvania, USA) (CSCW '00). Association for Computing Machinery, New York, NY, USA, 241–250. https://doi.org/10.1145/358916.358995

[25] Pranut Jain, Rosta Farzan, and Adam J. Lee. 2019. Adaptive Modelling of Attentiveness to Messaging: A Hybrid Approach. In Proceedings of the 27th ACM Conference on User Modeling, Adaptation and Personalization (Larnaca, Cyprus) (UMAP '19). Association for Computing Machinery, New York, NY, USA, 261–270. https://doi.org/10.1145/3320435.3320461

[26] Pranut Jain, Rosta Farzan, and Adam J. Lee. 2019. Are You There? Identifying Unavailability in Mobile Messaging. In Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems (Glasgow, Scotland Uk) (CHI EA '19). Association for Computing Machinery, New York, NY, USA, 1–6. https://doi.org/10.1145/3290607.3312893

[27] Pranut Jain, Rosta Farzan, and Adam J. Lee. 2022. Laila is in a Meeting: Design and Evaluation of a Contextual Auto-Response Messaging Agent. In Designing Interactive Systems Conference (Virtual Event, Australia) (DIS '22). Association for Computing Machinery, New York, NY, USA, 1457–1471. https://doi.org/10.1145/3532106.3533493

[28] Clara Jimenez, Edgar Saavedra, Guillermo del Campo, and Asuncion Santamaria. 2021. Alexa-based voice assistant for smart home applications. IEEE Potentials 40, 4 (2021), 31–38.

[29] Patrick Gage Kelley, Joanna Bresee, Lorrie Faith Cranor, and Robert W. Reeder. 2009. A "Nutrition Label" for Privacy. In Proceedings of the 5th Symposium on Usable Privacy and Security (Mountain View, California, USA) (SOUPS '09). Association for Computing Machinery, New York, NY, USA, Article 4, 12 pages. https://doi.org/10.1145/1572532.1572538

[30] Patrick Gage Kelley, Lucian Cesca, Joanna Bresee, and Lorrie Faith Cranor. 2010. Standardizing Privacy Notices: An Online Study of the Nutrition Label Approach. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (Atlanta, Georgia, USA) (CHI '10). Association for Computing Machinery, New York, NY, USA, 1573–1582. https://doi.org/10.1145/1753326.1753561

[31] Torkel Klingberg. 2009. The overflowing brain: Information overload and the limits of working memory. Oxford University Press.

[32] Pigi Kouki, James Schaffer, Jay Pujara, John O'Donovan, and Lise Getoor. 2019. Personalized Explanations for Hybrid Recommender Systems. In Proceedings of the 24th International Conference on Intelligent User Interfaces (Marina del Ray, California) (IUI '19). Association for Computing Machinery, New York, NY, USA, 379–390. https://doi.org/10.1145/3301275.3302306

[33] Matthias Kraus, Nicolas Wagner, Zoraida Callejas, and Wolfgang Minker. 2021. The Role of Trust in Proactive Conversational Assistants. IEEE Access 9 (2021), 112821–112836.

[34] Fifi Kvalsvik and Torvald Øgaard. 2021. Dyadic interviews versus in-depth individual interviews in exploring food choices of Norwegian older adults: A comparison of two qualitative methods. Foods 10, 6 (2021), 1199.

[35] Béatrice Lamche, Ugur Adıgüzel, and Wolfgang Wörndl. 2014. Interactive explanations in mobile shopping recommender systems. In Joint Workshop on Interfaces and Human Decision Making in Recommender Systems, Vol. 14.

[36] Nicholas D Lane, Mu Lin, Mashfiqui Mohammod, Xiaochao Yang, Hong Lu, Giuseppe Cardone, Shahid Ali, Afsaneh Doryab, Ethan Berke, Andrew T Campbell, et al. 2014. Bewell: Sensing sleep, physical activities and social interactions to promote wellbeing. Mobile Networks and Applications 19 (2014), 345–359.

[37] Gierad Laput, Yang Zhang, and Chris Harrison. 2017. Synthetic Sensors: Towards General-Purpose Sensing. In Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (Denver, Colorado, USA) (CHI '17). Association for Computing Machinery, New York, NY, USA, 3986–3999. https://doi.org/10.1145/3025453.3025773

[38] Q. Vera Liao, Matthew Davis, Werner Geyer, Michael Muller, and N. Sadat Shami. 2016. What Can You Do? Studying Social-Agent Orientation and Agent Proactive Interactions with an Agent for Employees. In Proceedings of the 2016 ACM Conference on Designing Interactive Systems (Brisbane, QLD, Australia) (DIS '16). Association for Computing Machinery, New York, NY, USA, 264–275. https://doi.org/10.1145/2901790.2901842

[39] Q. Vera Liao, Daniel Gruen, and Sarah Miller. 2020. Questioning the AI: Informing Design Practices for Explainable AI User Experiences. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (Honolulu, HI, USA) (CHI '20). Association for Computing Machinery, New York, NY, USA, 1–15. https://doi.org/10.1145/3313831.3376590

[40] Scott M. Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In Proceedings of the 31st International Conference on Neural Information Processing Systems (Long Beach, California, USA) (NIPS'17). Curran Associates Inc., Red Hook, NY, USA, 4768–4777.

[41] Deborah Lupton, Sarah Pink, Christine Heyes LaBond, and Shanti Sumartojo. 2018. Digital Traces in Context: Personal Data Contexts, Data Sense, and Self-Tracking Cycling. International Journal of Communications Special Issue on Digital Traces in Context, Vol 12, 647-665 (2018).

[42] Prashan Madumal, Tim Miller, Liz Sonenberg, and Frank Vetere. 2019. A Grounded Interaction Protocol for Explainable Artificial Intelligence. In Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems (Montreal QC, Canada) (AAMAS '19). International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 1033–1041.

[43] Lisa M Mai, Rainer Freudenthaler, Frank M Schneider, and Peter Vorderer. 2015. "I know you've seen it!" Individual and social factors for users' chatting behavior on Facebook. Computers in Human Behavior 49 (2015), 296–302.

[44] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. Artificial intelligence 267 (2019), 1–38.

[45] Jun-Ki Min, Afsaneh Doryab, Jason Wiese, Shahriyar Amini, John Zimmerman, and Jason I. Hong. 2014. Toss 'n' Turn: Smartphone as Sleep and Sleep Quality Detector. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (Toronto, Ontario, Canada) (CHI '14). Association for Computing Machinery, New York, NY, USA, 477–486. https://doi.org/10.1145/2556288.2557220

[46] Helen Nissenbaum. 2011. A contextual approach to privacy online. Daedalus 140, 4 (2011), 32–48.

[47] Florian Nothdurft, Felix Richter, and Wolfgang Minker. 2014. Probabilistic human-computer trust handling. In Proceedings of the 15th annual meeting of the special interest group on discourse and dialogue (SIGDIAL). 51–59.

[48] Martin Pielot, Rodrigo de Oliveira, Haewoon Kwak, and Nuria Oliver. 2014. Didn't You See My Message? Predicting Attentiveness to Mobile Instant Messages. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (Toronto, Ontario, Canada) (CHI '14). Association for Computing Machinery, New York, NY, USA, 3319–3328. https://doi.org/10.1145/2556288.2556973

[49] Ana Paula Retore and Leonelo Dell Anhol Almeida. 2019. Understanding Appropriation Through End-User Tailoring in Communication Systems: A Case Study on Slack and WhatsApp. In Social Computing and Social Media. Design, Human Behavior and Analytics, Gabriele Meiselwitz (Ed.). Springer International Publishing, Cham, 245–264.

[50] Cristiele A Scariot, Adriano Heemann, and Stephania Padovani. 2012. Understanding the collaborative-participatory design. Work 41, Supplement 1 (2012), 2701–2705.
[51] Tim Schrills and Thomas Franke. 2020. How to Answer Why – Evaluating the Explanations of AI Through Mental Model Analysis. https://doi.org/10.48550/ARXIV.2002.02526
[52] Thomas B Sheridan and William L Verplank. 1978. Human and computer control of undersea teleoperators. Technical Report. Massachusetts Inst of Tech Cambridge Man-Machine Systems Lab.
[53] Kacper Sokol and Peter Flach. 2018. Glass-Box: Explaining AI Decisions with Counterfactual Statements through Conversation with a Voice-Enabled Virtual Assistant. In Proceedings of the 27th International Joint Conference on Artificial Intelligence (Stockholm, Sweden) (IJCAI'18). AAAI Press, 5868–5870.
[54] Clay Spinuzzi. 2005. The methodology of participatory design. Technical communication 52, 2 (2005), 163–174.
[55] Nili Steinfeld. 2016. "I agree to the terms and conditions":(How) do users read privacy policies online? An eye-tracking experiment. Computers in human behavior 55 (2016), 992–1000.
[56] Wolfgang Stroebe and Michael Diehl. 1994. Why groups are less effective than their members: On productivity losses in idea-generating groups. European review of social psychology 5, 1 (1994), 271–303.
[57] Maxwell Szymanski, Martijn Millecamp, and Katrien Verbert. 2021. Visual, Textual or Hybrid: The Effect of User Expertise on Different Explanations. In 26th International Conference on Intelligent User Interfaces (College Station, TX, USA) (IUI '21). Association for Computing Machinery, New York, NY, USA, 109–119. https://doi.org/10.1145/3397481.3450662
[58] John C Tang. 2007. Approaching and leave-taking: Negotiating contact in computer-mediated communication. ACM Transactions on Computer-Human Interaction (TOCHI) 14, 1 (2007), 5–es.
[59] Paul Thomas, Mary Czerwinksi, Daniel McDuff, and Nick Craswell. 2021. Theories of conversation for conversational IR. ACM Transactions on Information Systems (TOIS) 39, 4 (2021), 1–23.
[60] Nava Tintarev and Judith Masthoff. 2007. A Survey of Explanations in Recommender Systems. In 2007 IEEE 23rd International Conference on Data Engineering Workshop. IEEE Computer Society, 801–810. https://doi.org/10.1109/ICDEW.2007.4401070
[61] Nava Tintarev and Judith Masthoff. 2011. Designing and Evaluating Explanations for Recommender Systems. Springer US, Boston, MA, 479–510. https://doi.org/10.1007/978-0-387-85820-3_15
[62] Chun-Hua Tsai and Peter Brusilovsky. 2019. Evaluating Visual Explanations for Similarity-Based Recommendations: User Perception and Performance. In Proceedings of the 27th ACM Conference on User Modeling, Adaptation and Personalization (Larnaca, Cyprus) (UMAP '19). Association for Computing Machinery, New York, NY, USA, 22–30. https://doi.org/10.1145/3320435.3320465
[63] Jennifer Wiley and Jeannine Bailey. 2006. Effects of collaboration and argumentation on learning from web pages. Lawrence Earlbaum, Mahwah, New Jersey. 297–321 pages.
[64] Jennifer Wiley and Cara Jolly. 2003. When two heads are better than one expert. In Proceedings of the Annual Meeting of the Cognitive Science Society, Vol. 25.
[65] Michael Yeomans, Anuj Shah, Sendhil Mullainathan, and Jon Kleinberg. 2019. Making sense of recommendations. Journal of Behavioral Decision Making 32, 4 (2019), 403–414.
[66] Neil Yorke-Smith, Shahin Saadati, Karen L. Myers, and David N. Morley. 2009. Like an Intuitive and Courteous Butler: A Proactive Personal Agent for Task Management. In Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems - Volume 1 (Budapest, Hungary) (AAMAS '09). International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 337–344.
[67] Neil Yorke-Smith, Shahin Saadati, Karen L Myers, and David N Morley. 2012. The design of a proactive personal agent for task management. International Journal on Artificial Intelligence Tools 21, 01 (2012), 1250004.
[68] Nima Zargham, Leon Reicherts, Michael Bonfert, Sarah Theres Voelkel, Johannes Schoening, Rainer Malaka, and Yvonne Rogers. 2022. Understanding Circumstances for Desirable Proactive Behaviour of Voice Assistants: The Proactivity Dilemma. In Proceedings of the 4th Conference on Conversational User Interfaces (Glasgow, United Kingdom) (CUI '22). Association for Computing Machinery, New York, NY, USA, Article 3, 14 pages. https://doi.org/10.1145/3543829.3543834

Received January 2023; revised May 2023; accepted June 2023