Designing for Awareness: Towards the Design of an Intelligible Proactive Agent to Support Message-based Communication

by

Pranut Jain

B.Tech Computer Science and Engineering, Guru Gobind Singh Indraprastha University, India, 2014

M.S. Computer Science, The University of Texas at Dallas, USA, 2016

Submitted to the Graduate Faculty of

the Dietrich School of Arts and Sciences in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2023

UNIVERSITY OF PITTSBURGH

DIETRICH SCHOOL OF ARTS AND SCIENCES

This dissertation was presented

by

Pranut Jain

It was defended on

July 5, 2023

and approved by

Dr. Adam J. Lee, Department of Computer Science

Dr. Rosta Farzan, Department of Informatics and Networked Systems

Dr. Jacob Biehl, Department of Computer Science

Dr. Stephen Lee, Department of Computer Science

Designing for Awareness: Towards the Design of an Intelligible Proactive Agent to Support Message-based Communication

Pranut Jain, PhD

University of Pittsburgh, 2023

In mobile messaging, there is a lack of situational awareness about the state of message recipients. With the ubiquity of mobile devices, there is an expectation of fast responses, which can lead to message recipients feeling pressure to respond quickly to incoming messages. This can lead to distractions from ongoing tasks. At the same time, delayed responses to messages have also been shown to affect social relations negatively. To compensate, message recipients often need to apologize and explain these delays to message senders. Messaging applications share cues such as Online/Offline status, read receipts, and last-seen time to improve availability awareness. However, these cues have been shown to be poor availability indicators and can raise privacy concerns.

This dissertation contributes to the design, implementation, and evaluation of a proactive messaging agent, which improves situational awareness in messaging by detecting and sharing unavailability and related context. There are multiple stages involved in the design of this agent corresponding to its goals of (1) improving situational awareness in messaging to reduce the perceived obligation to respond immediately; (2) being fully automated to reduce distractions and effort on the part of the agent owner to share their availability; and (3) preserving user privacy through mutual awareness and understanding of context-sharing preferences.

In the first stage, we demonstrate that we can accurately detect user unavailability by leveraging data such as sensor values from a user's smartphone. At this stage, we also identify and understand user preferences related to the utility and comfort of the information the agent could share to inform unavailability. In the second stage, we present the results of evaluating the agent in the real world. In the third stage, we co-design explanations with this agent's users to make the agent function more intelligible for its users and allow for its more appropriate use. Through this work, we contribute to an improved understanding

of the crucial factors in designing a virtual assistant to improve situational awareness in mobile messaging and inform the design of future virtual assistants to support asynchronous communication.

Table of Contents

List of Tables

List of Figures

Preface

There are many people I am thankful to for helping me complete this milestone, but none more than my advisors, Dr. Adam J. Lee and Dr. Rosta Farzan. Coming into the program with minimal research experience, they taught me how to do research. They were always available to talk, advise, and help with any academic, professional, or even personal matters. I appreciate their patience and support in helping me navigate the challenges I faced these last few years. Thank you!

I want to thank Dr. Jacob Biehl and Dr. Stephen Lee for serving on my committee and for all the feedback and advice they have given me throughout my dissertation research. I also want to express my gratitude to Dr. Daniel Mosse for his insights and valuable feedback on my research and for working with me on one of my most memorable projects (MAFIA).

My family has always been very supportive of anything I took up. I want to thank my parents, Rajesh and Seema Jain, for their unconditional guidance, support, and encouragement. I want to thank my grandfather, Sumat Jain, and my aunts, Sangeeta Jain, Mamta Jain, and Dr. Indu Jain, for their kind words and encouragement throughout this journey.

Finally, I would like to thank all the friends I made at Pitt. First of all, thank you, Injung Kim! I always looked forward to our coffee breaks and winning at racquetball. I will always be grateful for your assistance with my research and outside research as I transitioned to a new phase. I would also like to thank Henrique, Andrew, Gavin, Ekaterina, Briand, Talha, and Pratik. I will forever cherish the time we spent together. Thank you!

1.0 Introduction

1.1 Motivation and Problem Statement

Communication is generally regarded as the exchange of thoughts, messages, or information by speech, signals, writing, or behavior[1]. This exchange between two people is known as *interpersonal communication* [28]. People communicate to share ideas and information and build/maintain relationships [28]. The need for social connectedness drives us to initiate and engage in conversations, i.e., back-and-forth exchanges of messages. With the significant advancements in wireless technologies and increased adoption of wireless portable devices (mobile phones, tablets, laptops), we are now part of an increasingly interconnected world.

With this evolution of technology, in addition to in-person face-to-face conversations, emails, voice or video calls, and instant messaging are all now possible on the go. Particularly, instant messaging has now emerged as the preferred method of communication when compared to voice or video calls[2] and email[3] [4]. Instant messaging is gaining popularity, mainly due to its informal nature [32]. Multiple prior works have described messaging as a valuable tool for quick scheduling, coordination of activities, and question answering [142, 82, 90]. While 'online chat' was designed to be a synchronous form of communication, i.e., exchange of messages in real-time, instant messaging systems later allowed messages to be exchanged 'offline' without the user needing to be logged in. This enabled messaging to be used more asynchronously, similar to emails while being less formal and enabling real-time communication when possible.

Mobile devices such as smartphones are generally tied to an individual, and people are generally expected to carry their phones with them to most places [69]. Thus, there is this general expectation of constant connectivity. Internet-based messaging applications such as

[1] `https://www.ahdictionary.com/word/search.html?q=communication`

[2] Gallup report, `https://news.gallup.com/poll/179288/new-era-communication-americans.aspx`

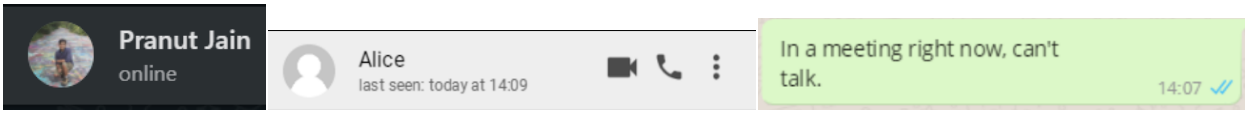
[3] GfK MRI Study, https://www.gfk.com/en-us/insights/press-release/smartphone-users-spend-as-much-time-on-entertainment-as-texting-gfk-mri-study/

[4] Flowroute Survey, https://www.flowroute.com/press-type/flowroute-survey-finds-consumers-overwhelmingly-prefer-sms-to-email-and-voice-for-business-interactions/

WhatsApp and Facebook Messenger on smartphones can instantly notify people (visually and audibly) of new incoming messages. Indeed, prior work has shown that people expect fast responses to their messages [129, 157]. To fulfill these expectations, message recipients feel pressure to respond to incoming messages immediately, even during inopportune moments such as being engaged in another task [9]. Attending to messages during inopportune moments has been shown to cause task and social disruptions [154, 183, 9] and can also affect task performance [12]. Further, being distracted during certain tasks can be even life-threatening. For example, texting while driving has been attributed as one of the leading causes of automobile-related injuries or fatalities [38].

Delays in responding to incoming messages have been shown to affect social relationships negatively. Past studies have shown that people begin to speculate when they do not receive a response within their expected time [192]. A survey of mobile users to understand message senders' interpretation of no response to their messages shows that a large percentage of senders interpret it negatively. It was found that 24% of the senders deem a recipient as 'is busy' whereas 15.4% respondents speculated that the recipient 'is pointedly ignoring me' or the recipient 'maybe in trouble' (5.7%), among other reasons [88]. Further, message senders may also be inclined to negatively adjust their responsiveness towards a contact if they feel their messages are not being responded to within their expectations [192, 157]. It has also been observed that message recipients often feel the need to justify and apologize for delays in responding [193]. Thus, the lack of awareness about the availability and activities of message recipients can cause negative emotions due to the expectation of fast responses.

This lack of situational awareness is more prominent in mobile messaging. In face-to-face communication, we can generally observe other's environments and infer their availability before initiating conversation. Even in a phone conversation, listening for background noise can provide hints about the callee's environment (e.g., commuting noise, background conversations). There is a lack of observable phenomena in messaging to gain context of other's situations. Mobile messaging applications such as WhatsApp and Facebook Messenger share cues like Online/Offline status, read receipts, and last-seen time to compensate for the lack of situational awareness. Figure 1 visualizes the presentation of these indicators on the popular WhatsApp messaging application.



a. Online status b. Last-seen time c. Read receipts

Figure 1: Typical indicators of availability used by messaging applications.

Research has shown that not only are Online/Offline status and last-seen time inaccurate predictors of someone's availability [157], but they can also raise social pressure to respond and could have privacy implications [157, 88, 50, 34]. For instance, Hoyle et al. [88] reported, based on the result of an online survey, the perceptions of message senders when their message is *seen* but not responded to, with almost 70% respondents reported feeling negative emotions (*upset/angry* or *slighted/ignored*) and 39% **speculated** that they are being ignored or may have been misinterpreted. The authors also reported how recipients were affected by the seen-time, as 68% of survey respondents reported deliberately avoiding viewing a message to pretend not seeing it. Similarly, Mai et al. [129] through an online survey, observed that *intensive* negative emotions are linked to delays in responses, especially when senders' are aware that their message has been *seen (or read)* but not responded to. Further, they also observed a higher perceived obligation to respond in message recipients due to the signaling of their message's *seen* status.

Thus, these observations point to the need to design better mechanisms to improve situational awareness in mobile messaging, focusing on reducing distractions while considering the user's privacy. This involves better understanding their needs to regain control of their attention.

1.1.1 Thesis Statement

In this dissertation, we explore the design of a virtual assistant that can improve situational awareness in messaging. In addition to supporting end-users in managing mobile messaging expectations, we also aim to support designers of intelligent agent systems in

developing adaptable and intelligible approaches tackling issues related to user modeling, context sharing, and the intelligibility of these systems.

It is possible to design an intelligible virtual assistant through user-centered design that can leverage mobile usage and sensor data to improve situational awareness in mobile messaging by predicting user unavailability and sharing relevant unavailability context.

1.2 Overview of Dissertation Work

This research focuses on designing technology to combat distractions from one of the most common sources of interruptions, mobile messaging [159, 120, 105]. Notably, we are looking into designing a virtual assistant to assist users with messaging interruptions. Virtual assistants have previously been shown to have the potential to reduce distraction in the workplace [107, 78] and have also been utilized as persuasive systems to foster positive behavior change related to physical and mental health [53, 205]. We aim to design an agent to support users in mobile messaging by (1) reducing distractions, (2) enabling situational awareness, (3) being considerate of user privacy, and (4) centering users' needs and preferences.

Target User: Based on these objectives, our target users for the messaging agent will be those who care about their responsiveness but can simultaneously make the most out of the agent mediation in their conversations without overly tweaking the agent model as this conflicts with the objective of reducing distractions and user engagement with their mobile devices.

We start by discussing some background and related work in Chapter 2. Particularly, availability management systems and setting up key concepts such as Situational Awareness, Attentiveness, and Responsiveness in messaging. We also discuss prior work in reducing distractions through agents, user modeling, and context-sharing. We end this chapter by discussing intelligent agents to motivate behavior change and prior attempts at improving the intelligibility of these agents.

To reduce distractions, automation of all aspects of the agent will be central. If the user needs to *ask* the agent to inform unavailability on their behalf, it leads to distrac-

tions and reduces the agent's utility. Thus, we wish to design this agent with a high level of proactivity [111]. The foundation work of this dissertation includes two parts. First, in Chapter 3, we focused on building an accurate prediction model. Further, in this chapter, we explore important considerations for utilizing different modeling approaches toward unavailability detection to understand their trade-offs and help designers pick the most appropriate approach based on the availability of user data. Next, once we establish automation by *accurately* detecting the user's unavailability state, in Chapter 4, we ask the question, *how do we improve situational awareness in mobile messaging?* We then explore **auto-responses** to incoming messages as a method to communicate unavailability and share unavailability-related context. As part of the design foundation of the agent, we explore *what can be shared as part of auto-responses to improve situational awareness?* Further, in the chapter, we explore *how do users perceive the utility and comfort of different information types shared through auto-responses?*

Following the foundation work, in Chapter 5, we explore *how does this agent work in practice?* In particular, *how do users perceive the usefulness of this agent to improve situational awareness?* We also explore *how users interpret the context shared by this agent generated from smartphone sensor data?* Finally, in this chapter, we also explore *in what ways the presence of this messaging agent affects user behavior?*

In the final part of this dissertation, we look into *how can we learn from user experiences to improve the agent design?* As users interact with the agent, they will develop mental models of how it works [58]. It becomes essential to accurately understand how the agent works to use it appropriately [162]. Thus, designing explanations for agent outcomes is central to making the agent more intelligible for its users. Through a co-design study, in Chapter 6, we first look into *how do users reason about the design and actions of the auto-response messaging agent?* and *what are their motivations for desiring explanations from this agent?* Answering these questions would allow us to identify gaps in user understanding and ways to augment agent design to fill these gaps to improve user experience.

Figure 2 shows the high-level research overview. There are three main stages to this work in reaching the thesis statement. In the first stage, we explore the design foundations, i.e., user modeling and context-sharing considerations. In stage two, we explore the important

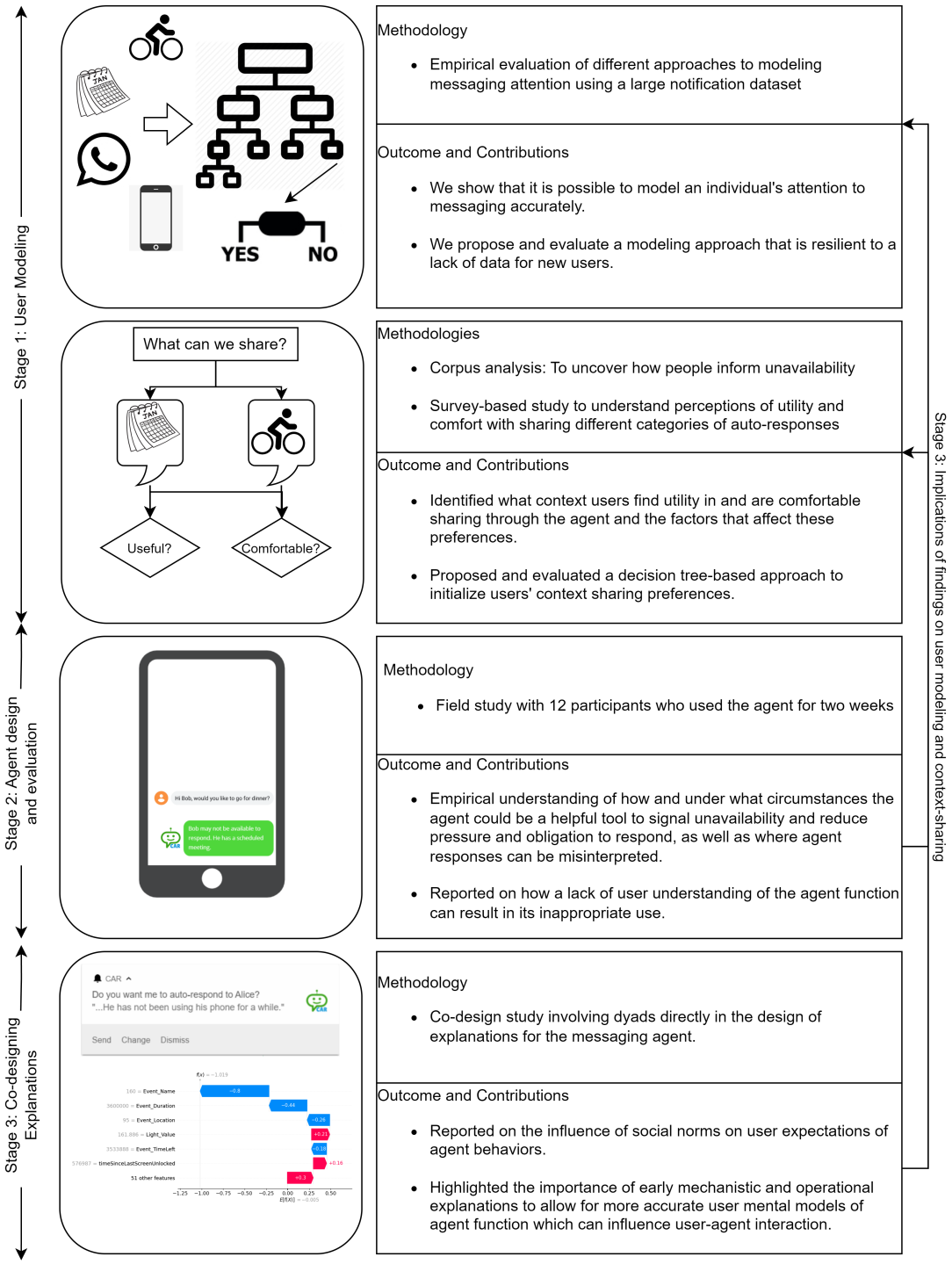


Figure 2: High-level research overview

design considerations for this agent and discuss the results from a field study. In stage three, we co-design agent explanations with end-users aimed to improve the agents' intelligibility and, subsequently, user agency in the outcomes of the agent. Finally, through the findings of stages two and three, we revisit stage 1 by exploring opportunities to improve user modeling and context sharing.

1. **Stage 1: Modeling unavailability and understanding context sharing through virtual assistants**

   In this stage, we explore the design foundations for an automated messaging agent. This involves understanding (1) what metrics and information we can use to determine user availability; (2) different approaches towards modeling availability and their trade-offs; (3) what information the agent can leverage to improve situational awareness; (4) user perceptions regarding the utility and comfort of the information the agent can share to inform unavailability.
2. **Stage 2: Designing and evaluating the auto-response messaging agent**

   In this stage, we describe the important design considerations for an auto-response messaging agent, followed by the results from a field study to evaluate its perception of utility in a real-world setting.
3. **Stage 3: Designing explanations for highly automated messaging agent**

   In the final stage, we discuss the challenges associated with improving the intelligibility of the agent due to its proactive nature. We describe and discuss the results of the co-design study with agent users and the implications on the design of future proactive messaging agents.

1.3 Broader Impact of this research

As more devices and services compete for our attention, it becomes essential to consider the economy of user attention in the design of technology and better support users in focusing on their tasks. We are starting to see research and industry trends in this direction. Our research explores unobtrusive solutions to minimize distractions caused by ubiquitous

connectivity. Through this research, we take the first steps towards finding ways to reinforce the asynchronous nature of messaging applications by looking into the design of an agent people can rely on while engaged in other tasks. Further, with this research, we aim to improve consumer awareness and understanding of intelligent agent functions and present directions towards getting back the control of technology around them.

2.0 Background and Related Work

In this chapter, we (1) provide background on and discuss important terms and concepts which are used frequently in this dissertation (Section 2.1); (2) discuss prior work in improving awareness in communication and their shortcomings (Section 2.2); (3) (Stage 1) discuss advances in user modeling for prediction of communication availability (Section 2.3); (4) (Stage 1) prior work in context-sharing approaches and their limitations (Section 2.4); (5) (Stage 2) how virtual assistants have been utilized for behavior change (Section 2.5); and (6) (Stage 3) ways in which we can improve the intelligibility of AI and ML systems (Section 2.6).

2.1 Concepts and definitions

2.1.1 Availability Management and Situational Awareness

Availability Management encompasses the activities and social processes involved in initiating, coordinating, and concluding social interactions [200]. One crucial aspect of availability management is **Situational Awareness**, which is the perception of the elements in the environment, comprehension of the situation, and projection of future status [70]. In face-to-face communication, observing the environment and state of others, such as their activities, can provide essential cues in determining appropriate avenues to initiate conversations. Even in voice communication, such as phone calls, environmental cues (i.e., background noise) can be utilized to get a sense of the activity of the callee. But what about Situational Awareness in mobile messaging? These cues are no longer present when initiating conversations through mobile messaging. People generally rely on their prior experiences and indicators provided by messaging applications when determining an opportune moment to initiate communication. However, contact initiation at inappropriate times can potentially disrupt task and social dynamics [154, 183, 9] and negatively impact task performance for the message recip-

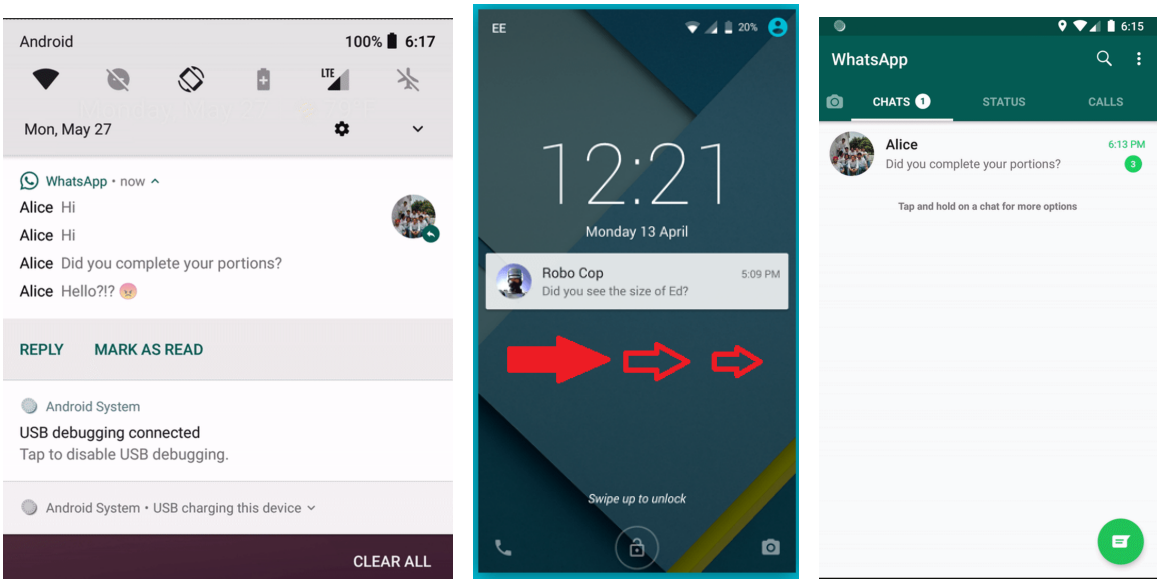
ient [12]. Further, disruptive or ill-timed communications are also likely to be ignored [39] or missed, for instance, due to the ringer profile set on the device [40, 166].

Social pressure and perceived obligation felt by message recipients can be inferred by their observed need to apologize and explain delays in responding. For instance, Voida et al. through interviews, observations, and text analysis, identified latent issues attributed to instant messaging [193]. One of the observed behaviors was the need felt by the message recipients to justify delays in responding by providing some situational context, possibly as a repair tactic to avoid coming off as rude (from [193]: "talking with Karen...sorry for delay in not talking") [166]. Further, it has been observed that individuals also feel the need to provide context when they need to steer away from a conversation (from [193]: "...I think I'm going to head home right now...can we talk later?") and may even use deceptive or dishonest explanations [81, 163, 166, 193].

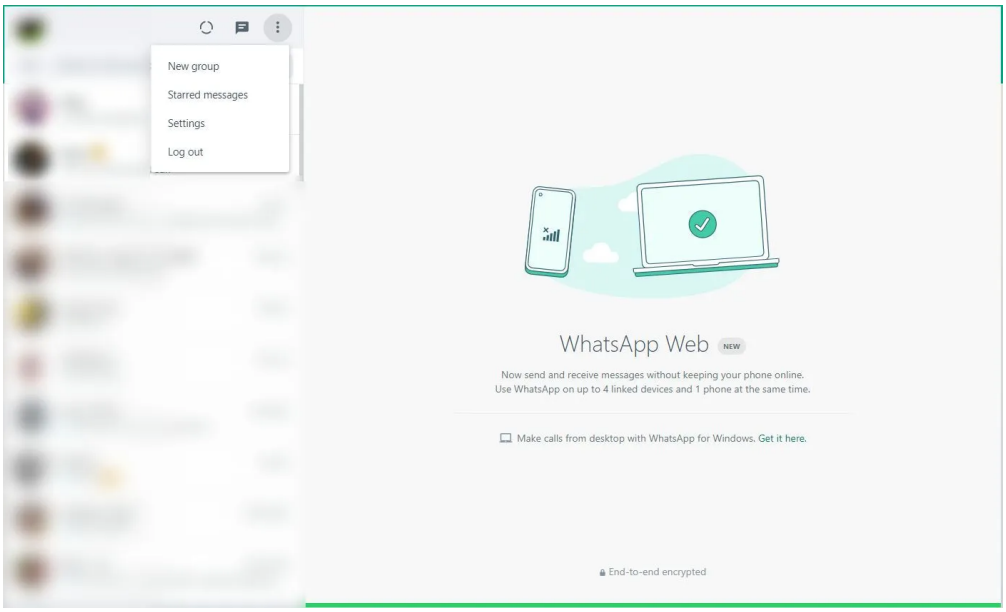
While in synchronous communication methods like phone calls, availability management is implicitly important; the above-mentioned observations point to issues surrounding communication and the need for better awareness mechanisms in messaging. In this dissertation, we build upon these observations regarding the importance of timely responses. Specifically, we explore the possibility of generating context-relevant automatic responses for incoming messages when the recipient is unavailable. We show that not only can onboard sensor data be used for classifying unavailability, but it can also be used to explain unavailability.

2.1.2 Attentiveness vs. Responsiveness in Mobile Messaging

Two other important concepts that we refer to frequently in this dissertation are **attentiveness** and **responsiveness** in mobile messaging. Figure 3 and 4 demonstrate the different ways to attend to or respond to an incoming message. A user is *attentive* to messaging if they are aware of an incoming message and any details about it [157]. Modeling attentiveness to messaging deals with predicting whether or not the user will attend to an incoming message within a few minutes. A user can attend to an incoming message by accessing the notifications drawer, swiping away the notification at the lock screen, opening the application which generated the notification, or accessing the message on another device [64].



a. Notification drawer b. Lock screen notification c. Messaging App



c. Whatsapp Web

Figure 3: Different ways to attend to an incoming message (a) Access the notification drawer; (b) swiping away notification in the lock screen; (c) opening the messaging app; and (d) accessing the message on another device.

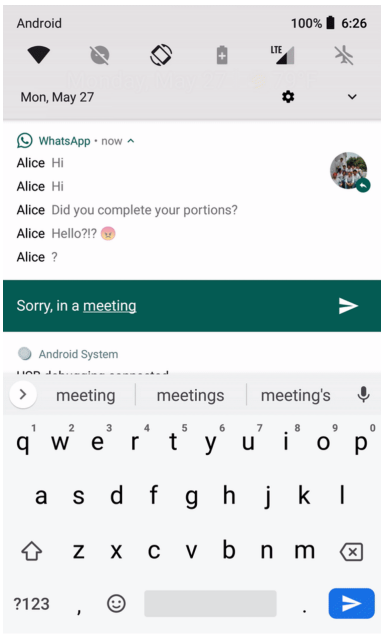


Figure 4: Typing and sending a reply within a certain time threshold constitutes responsiveness.

A user is *responsive* to a message if they respond to an incoming message within a certain threshold of time [10].

In this dissertation, we build on prior work on attentiveness modeling [157]. We focus on attentiveness rather than responsiveness as modeling responsiveness typically requires deep consideration of message content and relationship context, both of which could influence the user's decision to respond [135].

2.2 Prior attempts at reducing interruptions and improving situational awareness in communication

There have been several studies that looked into identifying inconvenient moments to engage in communication to defer notifications to minimize interruptions and their related expenses [134, 210, 152, 156]. For instance, the method proposed by Rosenthal et al. used

predictive models to automatically silence the user's phone to prevent disruptions from incoming notifications [165]. While these studies tackle the issue of interruptions due to ill-timed notifications, they do not address the lack of situational awareness in mobile messaging, which, as mentioned in the last section, can have social implications.

Researchers in computer-mediated communication have explored various methods and techniques to facilitate awareness in communication. Early attempts have utilized media spaces with video or audio streams to connect collaborators and improve presence and activity awareness [85]. Although, due to the use of visual and auditory data, the use of media spaces has raised concerns about their privacy implications [27, 66]. Media spaces are also constrained by locality, as their coverage is limited to the local space the sensors are installed. Thus, they can become inadequate in improving awareness as collaborators move away from these spaces [20].

People sometimes resort to unconventional methods without direct mechanisms to infer others' availability. For instance, Nardi et al. reported that people utilized IM applications' online/offline status to deduce someone's presence in their office space [142]. Early use of IM was limited to office workers' desktop systems, and when actively using their system, the IM software reports them to be online. People on the IM contact list could then use this information to infer if someone they were looking to contact was at their desk. As the technology evolved to be more portable such as the widespread use of laptops and mobile devices, this approach for presence detection became no longer viable. Later attempts utilized extending awareness information through status indicators for mobile devices. For instance, Tang et al. leveraged device usage indicators such as device-idle and last-used times for calls and IMs to augment status information and improve communication awareness [184]. In comparison, Handel et al. leveraged sources such as calendars and shared databases with team information to create a web application that individual users can leverage to infer the availability of other team members before setting up communication with them [83]. Since these proposed systems provide cues that users can then use to infer information on the state of others, the inference may not always be accurate. Instead, Wiberg et al. developed *The Negotiator* system, which relied on manually set status messages or preset text messages with availability information (e.g., "I will call you back in 0h25m") [200].

Later works proposed sharing richer contexts to tackle incorrect inferences due to limited cues to improve situational awareness. In a hospital workplace environment, Bardram et al. explored the use of sharing manual status, calendar information, and location among clinicians to improve situational awareness to coordinate shared tasks in hospital activities [14]. Buschek et al. developed a mobile application called *ContextChat* [37], which shared multiple contextual cues such as local weather, activities, the approximate distance between sender and recipient, and whether media is playing in the background. Their application augmented text messages with these contextual cues, limiting the amount of information shared compared to continuous data streams of prior works, particularly media spaces. Sharing a static set of contextual information has certain limitations. For instance, not all shared contexts might be relevant. Sharing indoor location and activity is helpful in hospital settings where this information can provide valuable insight into clinicians' availability [14]. In contrast, this shared context holds less utility in environments such as office spaces where employees may generally spend most time sitting at their desks. Trying to make sense of multiple streams of contextual values may increase the user's effort to make sense of multiple streams of sensor values. It may also result in the inaccurate inference of availability [19]. Further, privacy concerns were unaddressed with sharing multiple sensor data streams, mainly when not all data is relevant to improving situational awareness. As discussed in the next section, user modeling could help utilize multiple streams of sensor data to predict user state and use these user models to identify which context is relevant in a given situation.

2.3 Leveraging user modeling for unavailability detection

As mentioned before, for an agent to be useful in reducing distractions, it needs to be able to automatically detect the user's unavailability state and take proactive action on their behalf. This would enable users to focus on their ongoing tasks rather than trying to reply to every incoming message at inopportune moments.

In this dissertation, we will focus on modeling user behavior modeling in the context of communication. For instance, there is significant work in predicting opportune moments

for allowing notifications to minimize interruptions [152, 134, 210]. In terms of predicting communication availability, prior work has looked into availability prediction for both messaging [157, 64, 10] and phone calls [155, 168]. Particularly in mobile messaging, Pielot et al. used contextual data such as ringer mode, screen status, and proximity status to model users' attentiveness level to incoming messaging notifications [157]. They used aggregate notification data from *24* users collected over *two* weeks to train a general model and achieve a prediction accuracy of 70%. Their model predicted whether the user will *attend* incoming notifications within *6.15* minutes. While not focused on mobile messaging but rather instant messaging, Avrahami et al. instead modeled responsiveness to messaging for a desktop-based messaging client [10]. They used features such as the status of the message window (*open, closed*), buddy (or friend) status, and desktop environment features such as the *last accessed app* in the decision tree model. Their model could predict responsiveness (within *5* minutes) to incoming messages with accuracy as high as 90%.

In this dissertation, we extend prior work in user behavior modeling by evaluating a personalized modeling approach toward attentiveness prediction. It has been shown that users' device usage and messaging behavior can vary [4, 195, 194]. We hypothesize that by modeling users *individually* rather than using an aggregate of messaging data from multiple users, we can achieve better performance predicting user unavailability. In this dissertation, we also discuss the main limitation of a personalized modeling approach, i.e., *cold-start* problem, and how group-based and, subsequently, an adaptive modeling approach can overcome this limitation.

2.4 Improving communication awareness through Context-sharing

Context can be represented through various information types such as Location, Time, Activity, and Identity [101]. Mobile devices have enabled representations of more detailed and richer contexts which also incorporate aspects of usage and interactions in a more dynamic way [157, 134, 152]. In the last section, we questioned the need for multiple streams of information needed to improve awareness. The DASS framework proposed by Niemantsver-

driet et al. establishes three main themes concerning developing awareness systems by sharing information [144]. The themes are (1) What information is needed for awareness?; (2) How can awareness information be embodied?; and (3) How can the awareness be used effectively in interaction? The theme (1) on information for awareness further embodies subthemes - type, detail, inference, and privacy. Niemantsverdriet et al. emphasize the importance of weighing the trade-off between usefulness and privacy of the shared information for improving awareness [144], which was missing from some prior works listed in the previous section.

In this section, we explore prior work evaluating the perception of utility and privacy of sharing contextual information. Khalil et al. assessed the perception of comfort in disclosing four types of contextual information with different social relationship types to improve phone call awareness [103]. They recruited *20* participants for *10* days to understand their context-sharing preferences. Participants indicated feeling more comfortable sharing some context, such as company and in-conversation than their location of activity information. Their results showed that the social relationship was a significant factor in participants' disclosure rate of different contextual information. While this work considered the callee's perspective in their willingness to disclose different types of contextual information, it did not evaluate the utility of shared information for the caller. The work by Avrahami et al. looked into the effectiveness of different contextual information in allowing callers to make better decisions on when to initiate a phone call or leave a voice message [8]. Their work evaluated urgency as a factor in callers' decisions but did not include social relationships and privacy considerations from the callee's perspective. Further, both these works evaluated a limited set of contextual information. As mentioned earlier, data captured by smartphones can enable a much richer collection of contextual information [157, 93] and thus should be evaluated to understand its utility in improving communication awareness.

The work by Guzman et al. tackles this limitation [57]. They conducted a diary study with *13* users for *four* weeks to understand the perception of a more comprehensive set of contextual categories such as location, time, physical availability, social availability, task status, and emotional availability. Another study that evaluated a more extensive set of contextual information was done by Knittel et al. through a survey with contextual cate-

gories such as location, appointments, activity, phone usage, ringer profile, calling state, app usage, number of people in the vicinity, and mood [109]. Although, some of these categories of information cannot be automatically acquired. They require user input which can be distracting or annoying for the user if asked for frequently [17]. The authors suggested that body-worn sensors can be used to automate the inference of these categories. Although, this not only limits the practicality of the approach but also requires making inferences about the users' state, which can be ambiguous to some degree [212].

The works described so far have been evaluating sharing contextual information for improving callers' awareness about the callee's state before initiating communication. Although, it is unclear how the perceptions related to the utility and comfort of sharing various contextual information also apply to mobile messaging. The perception of what constitutes availability might differ between mobile messaging and phone calls. For instance, is the expectation of no response similar to phone calls and text messages if the communication recipient is in a library? In that environment, they cannot take a phone call due to the set rules of the location, but that doesn't prevent them from being able to respond to text messages. Thus, contextual information might differ in perception of utility and comfort in these cases when considering them with mobile messaging. Our work bridges this gap and evaluates the utility and comfort of a comprehensive set of contextual information for improving mobile messaging awareness.

2.5 Virtual Assistants and Behavior Change

There has been a stream of recent work toward developing virtual assistant systems for several applications. Some tasks where virtual assistants have been utilized include smart-home automation for people with special needs [149], academic advising and guidance for students [136], route navigation [151], and assistance with cooking [150]. For the listed applications, either the focus is on developing a virtual assistant or augmenting the capabilities of existing assistants for new tasks (e.g., adding new skills to Amazon Alexa).

Virtual assistants have also been utilized to reduce distractions in workplaces. Kimani et

al. designed and evaluated a conversational agent called Amber that users can interact with to schedule tasks and breaks [107]. The assistant can also detect distractions if it detects the user is going over a specific set time on social media. Work on persuasive systems [148], have also looked into making users more aware of their distractions through time spent on various activities to allow users to reflect on these activities and induce behavior change. For instance, persuasive systems have been utilized for positive behavior change in health and physical activity [53] and improving productivity through Digital Productivity Systems [205].

While researchers have evaluated virtual assistants for their role in behavior changes related to fields such as health and physical activity [53] and medicine adherence [13], we are mainly focusing on behavior changes related to the use of technology such as social media and any appropriation of technology to better suit individual needs [162]. The findings of Kimani et al. [107] reported that participants found agent suggestions around breaks and reflection useful and reported behavior changes in their routine with the use of the agent. Further, Grover et al. extended the work of Kimani et al. by introducing anthropomorphic features through a voice assistant [78]. They observed that this improved agent perception and its use for some participants. Similarly, in persuasive agent designs, agent nudges have been observed to help reduce time spent on social media [199]. In addition to self-behavior changes related to the use of technology, people have also been observed to appropriate technology to suit their needs better. For instance, in terms of communication, Retore et al.'s findings suggested that people tailor the way they use different controls on messaging applications (such as Slack and WhatsApp) depending on the *context-of-use* i.e., based on their situations and types of controls offered. These findings suggest that virtual assistants continue to show potential for improving the general well-being of their users. While virtual assistants can block notifications or silence smartphones to reduce disruptions, there is a stronger sense of obligation to respond to incoming messages. Even if ignored, the lack of awareness of the recipient state can negatively affect social relations and often requires effort to repair these social situations (e.g., by apologizing and explaining delays [193, 95]).

Our work augments this body of knowledge on virtual assistants by presenting and evaluating a novel design of an agent to manage user communication. By being cognizant of its user's state, the virtual assistant described in this work can act as an intermediary between

message senders and the owner of the assistant. This is important as it can potentially disrupt the flow of human-human conversation. Further, through the results from an on-field study, we report on additional dimensions associated with agent interaction that could not be identified in survey-based or lab studies.

2.6 Explanations to improve AI understanding

Multiple prior works have explored how explanations can help improve the understanding of intelligent agent systems. The work by Haynes et al. focused on understanding what explanations users desire as they interact with an intelligent agent [86]. The study involved familiarizing participants with the agent controls and involved domain experts and developers. Their findings indicated that users frequently expressed a desire to get operational explanations, i.e., '*how do I use it?*'. mechanistic explanations, i.e., '*how does it work?*', ontological explanations, i.e., identify, definitions, and relations for different components, and finally, the design rationale for the agent constructs. Generally, in explainable AI research, the focus is on generating explanations to explain either the model (**global** explanations) or the predictions made by the model (**local** explanations). In particular, for classification tasks, the focus is on features that have the most impact on model predictions [126, 79]. The design process of explanations for ML systems tend to rely on researchers or developers' institution and thus usually follows a more algorithmic view of explanations [137, 122], which studies have shown may not be the most appropriate for novice or non-technical users of the ML system [26, 181, 122]. These users have previously been shown to prefer local explanations, i.e., explanations for individual model prediction, rather than get a bigger picture of the model reasoning process [122]. At the same time, local explanations focused on specific predictions have also been shown to often be misleading for novice users and may result in an inaccurate understanding of the system [49, 23]. Explanations can be textual [110], visual [191], interactive [114, 176], or a mix of different types [181]. Novice users have also been shown to prefer visual explanations, although they tend to draw inaccurate conclusions from them [181].

Our work builds upon these prior works on explanations for AI systems by studying what explanations users desire in the context of a messaging agent. This agent has several unique attributes compared to other agent-based systems, such as recommender or conversational agents. The messaging agent is proactive by design to reduce distractions in mobile messaging. It can take multiple actions, mainly when the user is inattentive to their device before they get a chance to view these explanations. Thus, it becomes crucial to understand which explanation users desire, in what situations, and when is the ideal time to show these explanations. This can help reduce information overload from too many explanations presented at inopportune moments [108, 2], which can cause them to be ignored [145, 5, 178] or misunderstood [181]. The messaging agent also has social aspects associated with its use, i.e., it acts as an **intermediary** in human-human communication. Thus, it becomes also important to understand how these social aspects affect the use and desire for agent explanations.

3.0 Adaptive modeling and evaluation of approaches towards Messaging Attention Modeling

3.1 Introduction

Users are generally inconsistent in updating their status [18, 34], so a manual approach towards setting unavailability status may not be appropriate. As a result, the first step towards the design of a messaging agent is to automate unavailability detection[1]. It is essential for the messaging agent to automatically detect when their users are not available to respond and to act on their behalf. This can help reduce distractions from incoming messaging notifications.

We can automate unavailability detection by trying to understand and approximate a user's messaging behavior. By recognizing patterns in their engagement with incoming messaging notifications and modeling user behavior, we can identify instances when users cannot attend to their incoming messages.

Users generally carry their phones with them most of the time [60]. Smartphones have several sensors that can capture a user's environmental data, such as the light levels around the room (ambient light sensor), motion (accelerometer), and noise (microphone sensor). Pielot et al. proposed using data captured from a smartphone to build a messaging attentiveness model [157]. They used seven features, such as ringer mode, screen status, and proximity status, from 24 users collected over two weeks to build a generic model. Their modeling approach achieved 71% accuracy in predicting whether a user will attend to an incoming message notification within 6.15 minutes. In this chapter, we explore whether we can do better in terms of predicting unavailability.

We start by asking whether using aggregate data from multiple users is the most appropriate approach to predicting unavailability. The assumption with a generic (or generalized) approach is that the patterns in data that we identify generalize to a broader population [157]. It has previously been reported that smartphone usage varies in users [4, 195, 194]. For ex-

[1]The material presented in this chapter was originally published as [93] and [94].

ample, some people may have a calendar linked to their phone while others may not [96]. In particular, people may use mobile messaging for different reasons, e.g., as part of their work or for personal conversations only [96]. Building and utilizing individual models may result in better prediction performance in these cases. Indeed, prior studies have shown that personalization can improve prediction performance in prediction tasks like interruptibility prediction [156, 165, 152], call availability detection [74, 155], and recommendations [216]. Although, there are situations where a personalized modeling approach may not always be the most appropriate, particularly when there is a lack of initial training data, also known as the *cold-start* problem [172]. In these cases, the personal model may perform worse than a general model [89]. One approach to tackling the cold-start problem is leveraging group-based modeling [132]. In this approach, we identify a cluster of users similar to the target user by leveraging a limited amount of data on the target user and use the attributes of this group as the basis for modeling as a middle ground between personalized and generalized models.

As previously explored in user modeling, group-based modeling approaches help support users of adaptive systems when information about individuals is unavailable or collecting such information is undesirable, e.g., collecting privacy-sensitive information [175]. In such approaches, users are often clustered based on all available information, including demographics and user interaction with the system. Consequently, the same recommendations are provided for all members of the group. However, group-based personalization models can face three challenges: (1) including information beyond the implicit users' interaction with the system, such as demographic information, can introduce additional barriers, such as privacy concerns associated with collecting demographic information or requiring the users to provide additional information explicitly; (2) the performance of the model can depend highly on the accuracy of the clustering methods and set of features used in the clustering approach; and (3) using a group based model after enough personal information is available can lead to unnecessary sub-optimal performance.

In this chapter, we present our approach for building an adaptive hybrid weighted model that addresses these challenges in predicting users' inattentiveness to mobile messages. We first present that in contexts such as mobile messaging where rich user-interaction data is

available, a user clustering approach based on interaction and usage data can outperform clustering approaches based on users' characteristics such as age and gender. Showing that there is no need to collect such additional information in such a context. We then describe our hybrid model of users' inattentiveness, a weighted aggregate of general, group-based, and personalized models. We present our results of an evaluation analysis of this hybrid model and compare it to each separate modeling approach. Our results highlight the ensemble model's importance in better predicting the inattentive state and tackling the cold-start problem.

Our work extends prior research in user modeling by presenting a hybrid modeling approach for highly context-dependent and unstable tasks over time. This work provides a detailed description of the modeling approach, supporting future researchers in replicating and extending our work.

3.2 Methods

This section describes the data we used in this study, the types of features we extracted, our target variable, and evaluation metrics.

3.2.1 Dataset description

We used large-scale smartphone sensor logs collected as part of another study [156] for the performance analysis of our proposed approach. The data contains sensor logs from *342* participants collected for an average of *four* weeks. The events logged in the dataset fall into one of the following categories: (1) change-based events, such as a change in screen status from *on* to *off* or *unlocked*; (2) usage-based events such as the number of incoming messages, notifications, and phone calls; and (3) state-based events captured every 10 minutes such as battery state and connectivity (e.g., cellular, WiFi) state.

From this data, we extracted logs of messaging notifications by filtering the notification logs based on the package names of messaging applications. We focused only on notifications

generated by WhatsApp messenger[2] since they comprised 92% of all notifications in communication category applications in the data. After extracting the messaging notifications, the final dataset contained 1,375,359 notification instances from 274 participants spanning an average of 3 weeks.

3.2.2 Feature Extraction

We extracted a total of 72 features from the sensor logs belonging to the following four categories:

- **Current state of Sensor and Device data**, e.g., device orientation (portrait/landscape) and semantic location of the user (home, work or passing), current activity (on foot, cycling)
- **Time elapsed since last event**, e.g., time since an application was last opened or an outgoing call was made
- **Device usage in the last hour**, e.g., number of notifications received and network data transmitted.
- **Device usage in the current day (last 24 hours)**, e.g., percentage of time spent at home or work and total battery time.

3.2.3 Class variable

Our class variable is the user's attentiveness to messaging at the time of the incoming messaging notification. Attentiveness to messaging has been described in Section 2.1.2. If the participant in the dataset attended a messaging notification within **5.2 minutes**, then they were marked as attentive in that context. This threshold of 5.2 minutes is the median attend time in the dataset averaged across all users in the dataset [157]. To consider a notification as attended, a user either (1) accesses the notification tray on their device, which shows the notification details, including the message (or part of it), (2) opens the messaging application associated with the notification, or (3) access the notification on another device (e.g., through WhatsApp Web).

[2]WhatsApp, `https://www.whatsapp.com/`

Feature name	Description	General Model feature-score	Personalized Models: fraction of users	Group Models
timeSinceLastOpenApp	# ms since any app opened	7578	40.31%	1,2,3
Screen_Value	current screen status	2448	41.47%	1,2,3
timeSinceWhatsAppOpened	# ms since any whatsapp opened	1178	17.44%	1,2,3
timeSinceScreenChanged	# ms since screen changed	843	22.48%	1,2,3
Charging_Value	whether the device is charging	540	2.32%	1,2,3
HourOfDay	current hour of the day	466	1.55%	2
App_Value	current foreground app	427	4.26%	1,2
CellTower_GSMErr	amount of signal error	402	0%	-
perc_noloc	% time device unable to get loc	394	10.65%	-
timeSinceNotifCenter	# ms since notif center accessed	391	10.46%	2

Table 1: Top features identified by the general model along with their score (gain), the fraction of users who have that feature in the top 5 features of their personalized model, and the group models with that feature in their top 10 features.

3.3 Comparing modeling approaches

In this section, we describe three modeling approaches towards modeling messaging attentiveness, i.e., generalized and personalized, and group-based modeling approaches. We also compare their performance and discuss their shortcomings.

3.3.1 Generalized Modeling Approach

In a generalized modeling approach, data from multiple participants are aggregated to form a single general model [157]. We constructed this general model from the dataset described in Section 3.2.1, using a scalable gradient boosting decision tree approach called XGBoost [43]. The parameters for the XGBoost algorithm were set as follows after following the parameter tuning process: 'max_depth', i.e., the maximum depth of the tree to '5' and 'min_child_weight', i.e., the minimum weight to further partition the tree to '20'. Other parameters were set to their defaults as they did not significantly impact the model performance when testing different parameters.

We evaluated this general model through a 10-fold grouped cross-validation approach. We used *UUID* (Universal Unique Identifier) to group messaging instances by individual users to ensure that these messaging instances are not split between the training and testing folds during cross-validation. This approach to grouped cross-validation helps estimate how the model would perform for new users for whom we do not yet have any training data.

With our generalized modeling approach and grouped cross-validation, we achieved an accuracy score of 72.28% and an f-measure score for the inattentive class of 0.651. Our accuracy is similar to the 71% reported by Pielot et al. in their study, which also used a generalized modeling approach [157]. Further insight into the general model can be observed from Table 1, which shows the top features of the model ordered by the '*gain*' metric of that feature towards the model.

3.3.2 Personalized modeling approach

We created the personal models by using each participant's data individually. We again used the XGBoost algorithm to train these individual models. We used default parameters (with boosting iterations set to '20') as we did not notice a significant variation in model performance when testing different parameters.

Identifying messaging sessions is essential when building personal models on a messaging notification dataset. Since notification logs in our dataset are time-ordered and may contain sessions of fast message exchange [10], this can create a dependency structure between instances. Thus, when using randomized cross-validation, the model performance would be overestimated (notification instances within these sessions could be split into training and testing folds), while sequential cross-validation would underestimate the model performance [62, 164]. To tackle this issue, we grouped notifications into sessions by identifying clusters of notifications that arrived close to each other (15 seconds). Thus, when using cross-validation, we ensured that notifications within a session were not split across training and testing folds.

Evaluating the personalized modeling approach, we achieved an accuracy score of 84.21% and an average f-measure score for the inattentive class of 0.744. Both these metrics show

substantial improvement over the generalized modeling approach. Table 1 lists the fraction of personalized models with the same top feature as the general model in their top-5 features. We observed that only 40% of the personal models had the same top feature as the general model in their top-5 features. This suggests that with the personalized modeling approach, the models learn or assign higher weights to features depending on individual users.

3.3.2.1 Cold-start problem with personalized modeling

A significant concern with personalized models can be the lack of initial training data for a new user, which can lead to sub-optimal performance, even in comparison with a general model [89]. To investigate how much data will be sufficient for a personalized model to outperform the generic model, we assessed the individual models with a gradual increase of the training data in increments of days. For each user, we split the available data in the proportion of p/d, where d is the number of days represented in that user's data and p is the number of days to be used for training, which was varied from 1 to $(d-1)$. The rest of the data was used as testing data. We followed the session-based evaluation approach, as mentioned earlier. The process was repeated ten times for each user, and the results were averaged.

Figure 5 presents the change in F-measure for the inattentive class as more days of training data are added. The average performance increases as the number of training data days increases. After using *seven* days of training data, the personalized modeling approach outperforms the generalized approach, and with 16 days of training data, the model performance stabilizes.

3.3.3 Group-based modeling

Next, we discuss two methods of clustering users into groups, (1) demographics-based and (2) usage-based. The process of creating group models is similar to the generalized modeling approach, i.e., aggregate data from group members is used to train the model. We similarly evaluated these group models, following a 10-fold grouped cross-validation, where data from each user is not split in the training and testing folds to estimate how the model

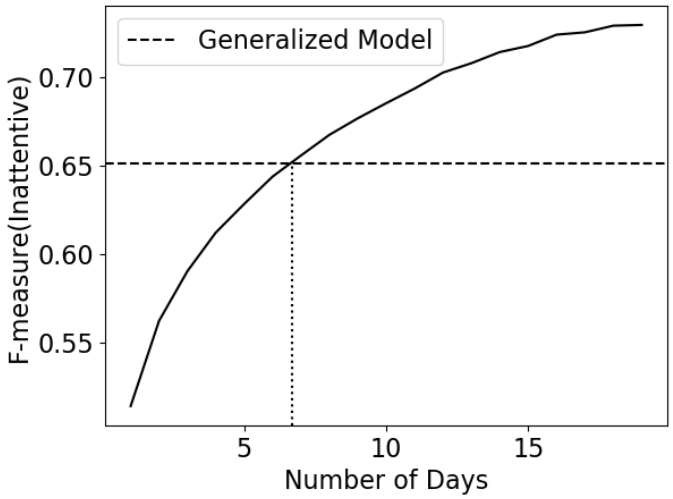


Figure 5: Number of days of training data and F-measure (inattentive)

	Group	Users	Accuracy	F-measure
Age	18-26	50	75.40	**0.660**
	27-35	78	69.84	0.574
	36-43	57	69.00	0.588
	44-50	50	70.46	**0.697**
	51-66	39	63.45	0.611
Gender	0	128	72.25	**0.664**
	1	146	72.47	0.634
Daily Behavioral	Cluster 1	137	72.60	**0.679**
	Cluster 2	87	70.59	0.589
	Cluster 3	50	71.14	**0.704**

Table 2: Grouping Summary. The accuracy and F-measure (inattentive) are computed by evaluating the model formed from the aggregate data of the group members.

will perform for a new user.

3.3.3.1 Demographics-based clustering

User demographics, i.e., *age* and *gender* have previously been shown to influence how people use their smartphones [4]. This is particularly relevant for mobile messaging where users have been observed to have high variation in their use based on their demographics [139]. Thus, demographics-based user grouping may help identify patterns in users with similar behavior that could help with predicting inattentiveness for a new user that belongs to that group.

(1) Clustering users by age. We start by grouping users based on their age group. The user age in the dataset ranged from *18 to 66* years. We used Jenks Natural Breaks optimization to find appropriate thresholds for the age distribution in the dataset. Setting the number of breaks to *five* resulted in the highest GVF (Goodness of Variance Fit) value of 0.92. Thus, we clustered users into five groups and evaluated the resulting attentiveness model. The resultant groups and their model performance are listed in Table 2.

Only the attentiveness model for the age group *44–50* outperformed the general model in detecting the inattentive state. This can be attributed to the fact that members of this group were less attentive to messaging than other groups (52% inattentive vs. 48% attentive instances).

(2) Clustering users by gender. The dataset had two genders, and we grouped users based on their reported gender[3]. The attentiveness model performance for gender-based groups is summarized in Table 2. *Gender 0* comprised 47% of all messaging instances in the dataset. Its attentiveness model showed only a minor improvement over the general model. Whereas, *Gender 1*, which makes up 53% of the messaging instances in the dataset, showed even lower performance than the general model.

Based on our evaluation results, the demographics-based clustering approach does not significantly improve inattentive state detection over the general model.

[3]The dataset represents gender only as 0 and 1 without association to any specific gender

3.3.3.2 Usage-based clustering

Feature Extraction. For usage-based clustering, we utilized the daily smartphone usage profile of the participants in the dataset. The first step to identify clusters based on usage is determining which smartphone usage vectors users have the most variation. Prior research indicates that location [119], application use [207, 215], movement patterns, and connectivity [194] are the dimensions where users have the most variations. Without making any assumptions of user behavioral attributes, We extracted an exhaustive feature set from all sensor events for the following categories: (1) environmental context-based features, e.g., *time spent at home, at work, and commuting*; (2) device-based features, e.g., *the number of times device was plugged in, screen state changed events, and device orientation changed events*; and (3) communication-based features, e.g., *the number of phone calls received, duration of incoming calls, and the number of messages received*.

User demographics were not included in the feature set for clustering. The final behavioral matrix X_i is of the shape $N \times K$ where N (=274) is the number of users and K (=52) is the number of feature dimensions. Each row of matrix X_i represents a user's daily behavior on average.

Clustering approach. We used a Bayesian Gaussian Mixture Model (BGMM) utilizing variational inference [7, 24] to estimate the membership of data points to a cluster. BGMM can be used as an unsupervised clustering approach. It does not require a pre-defined number of clusters as it chooses the optimal number of components to best fit the data. In our approach, each component was set to have its general covariance matrix allowing them to adapt to any shape and position. We set the number of expectation maximization iterations to 200 with ten initialization. We got *three* components (or clusters) upon fitting the model to the behavioral matrix X_i.

Interpreting the user clusters. Table 2 provides the details of usage-based clusters. Cluster 1 comprised 137 users, Cluster 2 comprised 87, and Cluster 3 comprised 50. We conducted PCA (Principal Component Analysis) to visualize the identified clusters along the dimensions of high variability and find correlated features [100] in the behavioral matrix X_i. Each feature f_i of X_i was standardized before computing the principal components. The

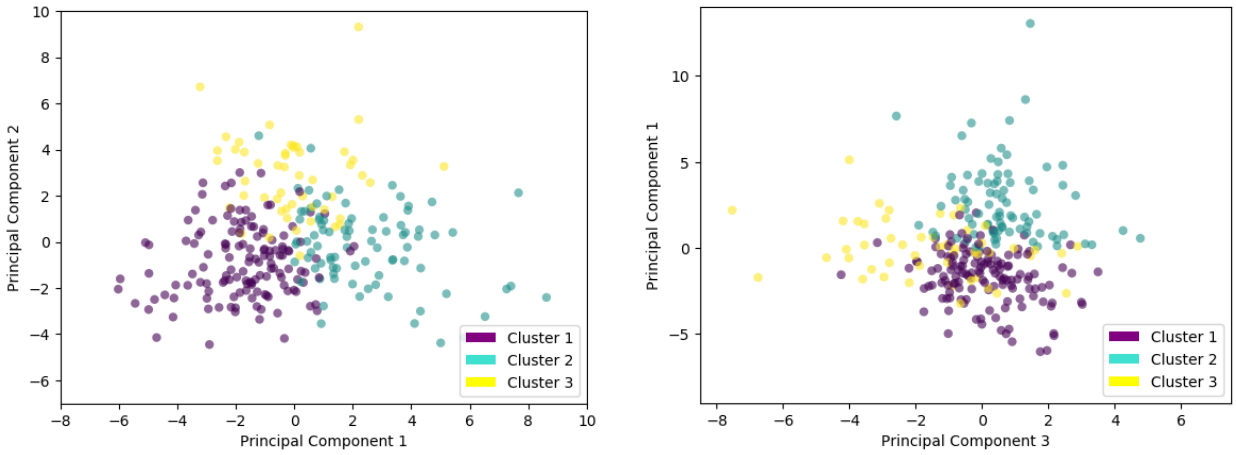
Principal Component	Feature	Score
PC-1 (variance_ratio = 0.155)	num_comm_dismissed	+0.286
	num_app	+0.284
	num_notifcenter	+0.277
PC-2 (variance_ratio = 0.091)	num_incomingcall	+0.351
	time_incall	+0.337
	num_missedcall	+0.291
PC-3 (variance_ratio = 0.063)	time_data_conn	-0.348
	time_wifi_noconn	-0.307
	num_outgoingcall	-0.302

Table 3: Top three Principal Components for the daily usage behavioral matrix X_i

top 3 principal components, along with their associated features, are summarized in Table 3.

Principal component 1 (PC-1) accounts for 15% of the variance in the data. The three main features included in PC-1 are the *number of communication notifications dismissed*, *number of applications opened*, and *number of times notification center was accessed*. The second principal component makes up 9% of variability in the data. It is comprised of features such as the *number of incoming calls*, *time spent on incoming calls*, and *number of missed calls*. The third principal component captures 6% variability in the data. It comprises features such as the *time connected to mobile data*, *amount of time not connected to a WiFi network*, and *number of outgoing calls*. Based on the comprised feature weights for PC-1, it signifies variability between users in terms of how actively they check and interact with their phones. PC-2 signifies user variability based on how actively a user engages in phone calls, and PC-3 signifies user variability based on users' network connection status.

Cluster assignments based on the top three principal components are visualized in Figure 6. A distinction between the three identified clusters can be observed for both PC-1 vs. PC-2 (Figure 6a) and PC-3 vs. PC-1 (Figure 6b) plots. On further analysis of the



(a) PC-1 vs PC-2

(b) PC-3 vs PC-1

Figure 6: Plot comparing cluster assignments against Top 3 principal components

three clusters, it can be observed that cluster 2 users show comparatively **more active** use of their device (they frequently check their phones and open a greater number of applications throughout the day). In comparison, cluster 1 users are **less active** users who receive fewer notifications per day and generally have comparatively less interaction with their devices. Cluster 3 users are moderately active regarding interaction with their phone but are **active callers** as they receive and make relatively more phone calls than the other two groups. They are also, on average, connected longer to a cellular data connection than a WiFi connection. Further, they spend more time traveling as they have higher daily on-foot, cycling, and in-vehicle average times, which also explains the extended periods of cellular data connection.

Model Evaluation. The evaluation results of usage-based group models are shown in Table 2. We observed significant improvement in the mean f-measure score for the inattentive class for clusters 1 and 2. On the other hand, the cluster 2 model performed worse than the general model. As noted earlier, cluster 2 users are generally more active in the use of

their devices. This explains the lower model performance since it becomes harder to detect the inattentive state due to the class imbalance in cluster 2 users' data (39% inattentive vs. 61% attentive instances).

Further, it has previously been reported that recent communication, such as making or receiving phone calls, is associated negatively with a user's availability for further communication [156, 158]. This would explain the better performance compared to the general model for detecting the inattentive state for the cluster 3 model, as the users in this cluster communicated more frequently via phone calls. Table 1 shows which group models share the same top features as the general model in their top 10 features.

3.4 Adaptive weighted modeling

While personalized models provide more accurate modeling of an individual's messaging behavior as the basis for prediction, they require sufficient user data to do so. In the face of insufficient personal data, a general model can outperform a personal model. Further, a group model will outperform a general model, given the correct association for a new user to a behavioral cluster. Our results of usage-based clustering analysis show that the group-based attentiveness model outperforms a general model for predicting a user's inattentive state for two of the three identified user groups. Thus, if a new user demonstrates daily behavior similar to users in these two groups, their group model should be utilized rather than the general one. However, relying on a single type of model may not be the most appropriate approach as (1) depending upon usage behavior and lack of initial data, the general model may perform the best for some users; (2) a behavior-based clustering approach requires at least a day of usage data to detect the behavioral group for a user, which may not be representative of the user's behavior as group membership could change as more data becomes available; (3) even with the adequate amount of data, a personalized model would require time to adapt to sudden changes in user's behavior and environment.

Thus, it may be beneficial to consider a more dynamic modeling approach rather than relying on a single approach. One method could be to select the model type based on the

Define:

clu, gen, per = group, general, personal models

f^* = set of f-measures of each model

day_usage = aggregate user behavior for the current day

day_instances = message instances current day

Input : x: a new instance of incoming message

Output: *state*: attentiveness state

begin

`/* check if a new day has begun */`

if $getday() \neq current\_day$ **then**

f^* = compute_models_performance(y_preds, y_true)

clu = get_cluster(day_usage)

w^* = update_weights(f^*) using eq. 3

per = update_personalized_model(day_instances)

reset f^*, day_usage and day_instances

current_day = getday()

$P_{gen}(y_i = 0) = gen(x_i)$

$P_{clu}(y_i = 0) = clu(x_i)$

$P_{per}(y_i = 0) = per(x_i)$

$P(y_i = 0)$ = combine predictions using equation 2

if $P(y_i = 0) > 0.5$ **then**

state = *inattentive*

else

state = *attentive*

y_preds.add(model, state)

return state

end

Algorithm 1: Adaptive Modeling Approach

current user situation, i.e., the amount of data available and their group association. Instead, we propose a hybrid approach that integrates predictions from multiple models to adapt to the situations mentioned above without relying on the amount of available data. This also covers situations where the users' behavior or environment changes, e.g., when they go on a vacation.

Algorithm 1 describes our adaptive modeling approach. Given a data point x_i, its class y_i can be determined by

$$y_i = \sum_{c \in C} w_c * y_c(x_i) \tag{1}$$

where $C = \{cluster, general, personalized\}$ is the set of models in use, w_c is the weight associated with model c and ranges between $\{0, 1\}$ and $y_c(x_i)$ is the class predicted by model c for the data point x_i. For modeling approaches that return the probability of each class for a given data point rather than the class value, we can rewrite equation 1 with the probability value returned by the model for the inattentive class, $P_c(y_i = 0)$:

$$P(y_i = 0) = \sum_{c \in C} w_c * P_c(y_i = 0) \tag{2}$$

We can then consider that if $P(y_i = 0) > 0.5$, set the class as *inattentive* or adjust that threshold to different values for more relaxed or more conservative models.

To set the weights w_c assigned to each model, the simple approach can be to set them to a pre-computed static value or as a function of the amount of data available for a user since heuristically, as more data becomes available, the weights for the personalized model should be increased while reducing the weights for the group and general models. However, statically set weights would not consider sudden changes in user behavior, which can affect the model performance.

Therefore, to address this limitation, we propose a dynamic approach to update the model weights based on how well a model performed recently for a given user. Previously, prediction accuracy through RMSE (Root Mean Squared Error) has been used to derive weights for classifiers in the ensemble model [33, 197]. This method of accuracy-weighted voting does not work well for unbalanced datasets [42]. Hence, instead, we use F-measure (for inattentive class) to determine the 'fitness' of a model in the ensemble [42].

Let w_c^{t+1} be the weight of the model to be used at the next time step $t+1$, then

$$w_c^{t+1} = \frac{f_c^t + \alpha(\Delta f_c^t)^3}{\sum_{m \in C} f_m^t + \alpha(\Delta f_m^t)^3} \tag{3}$$

where f_c^t is the performance of model c in terms of f-measure for the inattentive state at the current time-step t, α is a constant and Δf_c^t is the change in the performance of model c from previous time-step i.e.

$$\Delta f_c^t = f_c^t - f_c^{t-1} \tag{4}$$

The denominator normalizes the weight to be between $\{0, 1\}$. We take the cube of Δf_c^t to emphasize more considerable gains while keeping the *sign* of the change in performance. As observed from equation 3, the model's weight for the next time-step only depends upon the model's performance in the current time-step and the change in performance from the previous time-step. The term $\alpha(\Delta f_c^t)^3$ will either reward or penalize the model based on the change in its performance. The parameter α can be tuned based on the granularity of the time-step t. If the weights are updated per instance basis, then α should have a lower value while it should be set to a higher value with day-to-day weight update.

This type of weight assignment scheme allows the adaptive model to adjust to the amount of user data available and adapt to users' most recent behavior. For instance, a user's messaging patterns might change while on vacation. The personalized model might not have observed the user's behavior in this new environment in the past, and thus its performance would likely suffer. Detecting this drop in performance, the adaptive model would penalize its weight for the next timestep until the personalized model adapts to this new environment.

Identifying the most important features in a model is often essential to improve the model or, in the case of a messaging agent, form explanations for users' inattentive state. To compute the relative importance of features, we multiply the individual feature scores of each model with the model weight and then pick the top k scoring features. The feature scores can be the 'gain' provided to the model by the feature or other metrics, such as the information gain ratio.

foreach *user* $u \in U$ **do**

`/* train general model without user u */`

$gen_u = train_{gen}(data - data_u)$

`/* perform clustering without user u */`

clusters = user_clustering(U-u)

for $d \in range(1, k)$ **do**

`/* get cluster membership based on average cumulative daily data for day d */`

user_cluster = get_cluster(clusters, $daily_u^d$)

`/* train group model with similar users data */`

$clu_u = train_{clu}(user\_cluster)$

train_size = d/k

`/* get user data split by day, 10 folds */`

train_data, test_data = groupCV(train_size)

$per_u = train_{per}(train\_data)$

if *d = 1* **then**

`/* Initialize model weights using training data for day 1 */`

w^* = initialize_weights(train_data)

$pred_{gen} = gen_u(test\_data)$

$pred_{clu} = clu_u(test\_data)$

$pred_{per} = per_u(test\_data)$

$pred_{adapt}$ = combine predictions using equation 2

f^* = compute_models_performance(y_preds, y_true)

w^* = update_weights(f^*) using eq. 3

end

end

Algorithm 2: Evaluation process

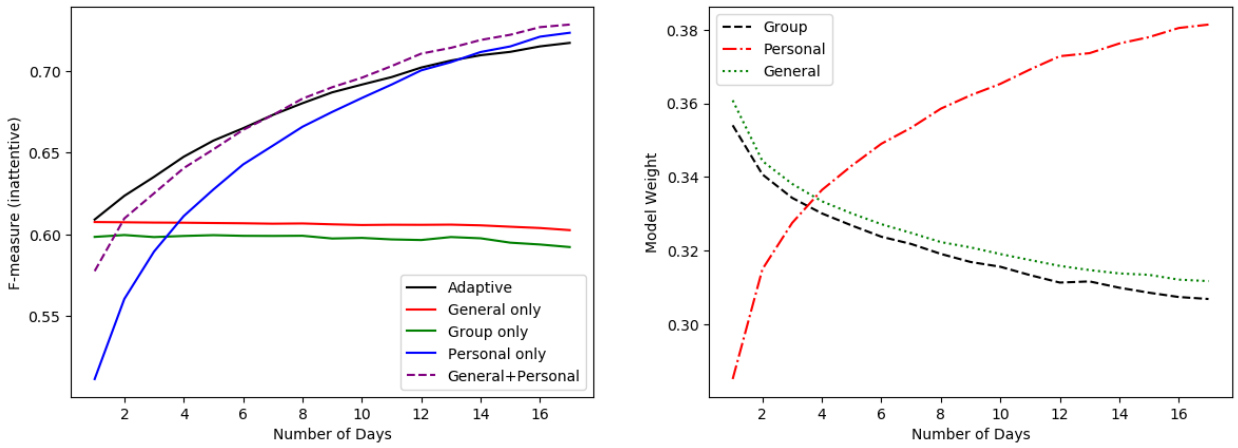
3.4.1 Evaluation

Our evaluation process has been summarized in Algorithm 2. The objective of the evaluation was to simulate multiple modeling approaches for a new user and get an estimate of how each performs as more data becomes available over time.

For each user, the amount of data available was gradually increased in one-day increments. The available data for the user was split in proportion of d/k where d is the number of days of data to use for training, and k is the total number of days of data available for that user. This forms the training set for the personalized model, and the remaining $(1 - \frac{d}{k})$ data becomes the testing set. For consistency in the number of users during the evaluation process, we only considered users with at least 18 days of messaging data available in our dataset, making up 79% (216) of all users. The general model was trained as discussed in Section 3.3.1 while not including the target user's data.

Similarly, cluster detection, as discussed in Section 3.3.3.2, was performed to find and model user groups without including the target user. To determine initial cluster membership, only one day of usage data of the target user was utilized, and as more data became available, cluster membership was re-evaluated. The general and group models were also evaluated on the same test data as the personalized model. The predictions of all three models were then combined as discussed in Section 3.4 to get the predictions for the hybrid weighted model. We repeated this process for each user in the dataset and averaged the performance of each model over all users for each day. The plot comparing the average model performance with the increasing amount of available data in terms of the number of days is shown in Figure 7a. It can be observed that the personalized model performance is considerably low during the first few days due to the lack of training data. The general and group models show consistent average performance throughout the testing period. Group models, on average, slightly underperformed when compared to the general model since the general model performed significantly better for one of the discovered clusters in detecting inattentiveness, bringing the average down.

The adaptive model performs better than all other models during the starting few days and eventually settles at personalized model performance. To assess what impact the group



(a) Change in performance

(b) Change in weights

Figure 7: Comparing model performances and change in model weights based on days of data available

model has on the adaptive model, we included a plot of the adaptive model performance without including the predictions of the group model. It can be observed that there is a performance drop until day 6, confirming that the group models provide a significant gain to the adaptive model for the initial few days. While a few days might not seem significant, it should be considered that most users decide to utilize a new application based on their initial experiences. A disappointed new user would likely not return to the application [106].

Figure 7b, shows how the dynamically assigned model weights change over time as more data becomes available. This plot can also be interpreted as the relative model importance with respect to time. The weight for the personalized model increases sharply as more data becomes available, and after day four, it has more weight than the group and general models. The weight change subsides around the 16-day mark with general at 0.312, cluster at 0.307, and personalized model at 0.381.

3.5 Discussion and Summary

In this chapter, we presented an approach for building an ensemble model to accurately predict instances when users are inattentive to messaging. We present how this hybrid approach can overcome challenges faced by different modeling approaches alone. Our approach allows the model to adapt to user behavior as more data is collected by (1) considering a dynamic, usage-based clustering approach and (2) creating a hybrid weighted model that optimally combines information about the user being profiled with models of more general user classes.

Computationally, our approach involves three modeling stages. First, we must train the general model, which needs to be done infrequently unless the user population changes significantly. Second, we must maintain up-to-date group models, which require identifying group memberships for individual users and training group models. While the group membership for a user can change over time, the group model does not need to be retrained frequently. Third, we must regularly update personalized models to adapt to user behavior and environmental changes. In this work, we used a batch training approach, which required retraining the model again as more data became available. This frequent retraining not only takes up computation resources but also requires storing batches of user data which can subject the users to privacy compromises of their data. Another approach would be to use an online or incremental classifier [68, 152, 210]. Incremental approaches update the model per instance or in mini-batches and often do not require previously used training data while reducing the training time significantly [35]. However, they do not perform as well as batch-trained models in many cases [174, 35, 47].

Detecting instances of inattentiveness accurately is the first step towards designing an intelligent messaging assistant to support users during moments of unavailability. In this chapter, we tackled part of the first challenge, i.e., the automation of the agent. The agent can use the proposed modeling approach to accurately detect when the user is unavailable and take some action on their behalf. As discussed earlier, that action is to share context to improve situational awareness. The next challenge is what context we can share as part of the agent action to improve situational awareness.

Our next steps include generating textual auto-responses to explain a recipient's unavailability to the message sender. Constructing such responses requires understanding what contextual factors are affecting a user's availability at the time of an incoming message. This information can be extracted from the user's attentiveness model, which captures the user's messaging behavior. However, several challenges are involved in this endeavor, mainly to identify accurate and *effective* [57, 109] information regarding the user's state and ensuring the protection of the user's privacy.

4.0 Improving Situational Awareness through Auto-responses

4.1 Introduction

In the last chapter, we established that hybrid modeling enabled us to detect unavailability with high accuracy even with a lack of initial data for a new user. The next step in the design of the messaging agent is to understand what actions we can take upon detecting that a user will not be able to attend to incoming messages. Recall that with the design of this agent, our goals are to (1) reduce distractions related to frequently checking messaging notifications through automation; (2) improve situational awareness regarding the state of the message recipient, and (3) account for their privacy preferences.

Toward these goals, we propose the use of auto-responses in messaging to improve situational awareness (Figure 8)[1]. The agent can send auto-responses automatically upon detecting the unavailability of the message recipient without requiring user intervention and thus potentially reducing distractions (goal 1). Through these auto-responses, the agent can share context related to the user's unavailability improving situational awareness (goal 2). Since these responses are shared in the same thread of conversation as the incoming message, the message recipient is aware of what context is shared and with whom enabling mutual awareness [46] (goal 3 partially). Another benefit of the automated response approach is that these auto-responses can be sent after the sender initiates communication and the recipient is predicted to be unavailable. This way, the sender is not discouraged from starting a conversation by observing a busy flag before initiating a conversation. The recipient will not miss any messages for that reason. However, generating accurate, useful, and trustworthy auto-responses that share information that the user is comfortable sharing (goal 3) remains an open area of research.

Availability models built by Pielot et al. [157] and in our work [93, 94] used information available directly from a user's smartphone to establish *context* that characterizes the situation of an individual or their device [59]. For instance, Pielot et al. used 17 features, such

[1]The material presented in this chapter was originally published as [95].

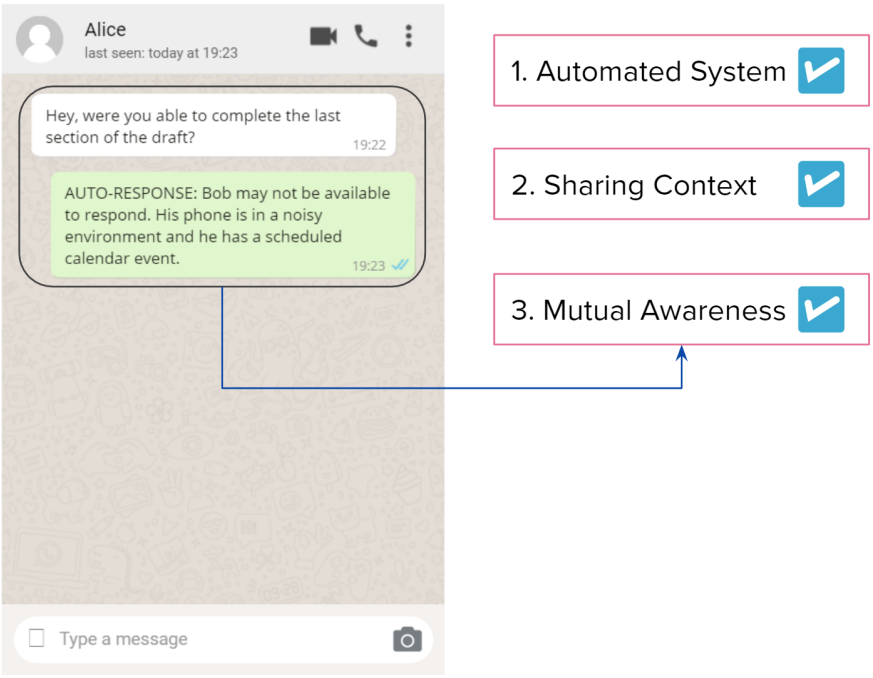


Figure 8: Auto-responses as a way to improve situational awareness.

as the state of the proximity sensor, ringer mode, and screen setting in their model. In our work, we utilized a more comprehensive set of 72 features to represent context at the time of an incoming message (see Section 3.2.2 for more details). Both these representations of context characterize the environment of the user and their device in a relatively static form. Dourish argues that context is rather an emergent property of the ongoing interaction [65]. That is, (1) not all features are always relevant when accounting for the availability of an individual; i.e., sharing an irrelevant feature as an explanation may not provide any benefit to the message sender and worse, may even further confuse the issue, (2) the characteristics of an interaction, such as the purpose of the communication or the relationship between the communicating parties, can influence the shared context utility by affecting how that context is interpreted. Further, these features may contain information that a user might consider sensitive, such as their location. Even if the user is comfortable sharing certain information, that does not imply that the message sender (with whom the information would be shared) will find that information useful or adequate for explaining unavailability. Thus,

it is important to consider context more as a dynamic property of the interaction.

In this chapter, we analyze users' perception of several automated contextual responses from the perspective of both message senders and recipients. We specifically target one-to-one communication since the expectation of fast responses is more apparent in those conversations than in group conversations, where a message is usually directed toward multiple conversational participants. A *message recipient* in a one-to-one conversation is the person who receives a message from one of their contacts but cannot respond at that moment. In contrast, a *message sender* is the communication initiator and the contact who gets the automated response back. We analyze the usefulness of automated responses from the perspective of message senders and individuals' comfort level in sharing contextual information from the perspective of the message recipients. Based on our analysis, we then provide design guidelines for generating automated responses to manage users' unavailability in responding to mobile messages. Moreover, we provide insight into how different people (both message senders and recipients) perceive such messages differently and what characteristics contribute to that difference.

More specifically, in this chapter, we explore the following research questions:

- **RQ1: What types of automated responses can be generated using contextual information collected from an individual's smartphone?** Auto-responses can be generated in different ways, including simple standard messages, pre-defined 'canned' messages written by users, or messages considering users' current status. In this work, we are particularly interested in messages generated based on the context that can be automatically inferred from sensors on people's mobile devices with little or no extra work for the individuals. However, as mentioned earlier, this context can involve many different features. Prior research has identified a significant number of features (as high as 72) used to represent user context [158, 93]. Thus, the first step in generating context-based auto-responses involves identifying which types of responses can be put together to create meaningful responses. Therefore, our first research question focuses on identifying the appropriate types of context-based auto-responses.
- **RQ2: What is the perceived usefulness of different types of automated responses? And How comfortable do people feel with sharing each type of**

automated response? For each type of automated response that can be generated, it is critical to understand how message senders (i.e., communication initiators) perceive the usefulness of that response. Additionally, it is crucial to understand the message recipient's perception of how comfortable they are with an auto-response sent on their behalf. Therefore, our second research question focuses on assessing senders' and recipients' perceptions of the usefulness and comfort of context-based auto-responses.

- **RQ3: How do users differ in their perception of the usefulness of and comfort associated with automated responses?** It has been observed that people have varying privacy concerns [130, 30, 131]. Further, people tend to differ on how they utilize technology [139, 4]. Personalization is now becoming an integral part of multiple application areas such as web-based systems [21] [182], learning and education [16], banking [198] and even availability management [94, 93]. Thus, it becomes essential to consider not only the utility of automated responses but also individual differences which can affect the perception of these responses. Identifying characteristics associated with individual preferences can help the autonomous agent adapt to different user groups to address their needs most effectively.
- **RQ4: What is the role of message urgency and social relationship in the perceived usefulness and comfort level associated with automated responses?** We hypothesize that communication context in the form of the urgency of the message and sender-recipient relationship can play a role in how automated responses are perceived by the users. Multiple previous works have reported the role of relationships in messaging [135, 73], self-disclosure [104, 214, 124], location-sharing [54] and context-sharing [109]. Previous works have also pointed out the role of urgency concerning the reception of communication [50, 185]. Thus, our fourth research question focuses on understanding how these factors impact users' perception of comfort and usefulness for context-based auto-responses.

To address our research questions, we first analyzed a text messaging corpus to understand what context people generally provide when communicating or explaining unavailability and conducted a survey through Amazon Mechanical Turk to understand perception of utility and comfort with sharing different categories of auto-responses . Our findings in-

dicate varying perceptions about an automated response depending on the context of the information shared through the messages, the relationship with the sender, and the message's urgency. Our contribution in this work is two-fold: (1) We present the findings of corpus analysis and how it informed the design and implementation of an online survey about user perception of automated responses; (2) We discuss the implications based on our findings from our survey to design an assistive agent which can support individuals' interpersonal communications through messaging.

4.2 Methods

In this section, we describe the creation of our survey instrument, methodology, and the analytical approaches used to understand peoples' utility and comfort assessments of contextually generated auto-responses in instant messaging platforms.

4.2.1 Analyzing ways people communicate unavailability

To develop an agent to construct contextual auto-responses, the first step is understanding whether and how people typically communicate unavailability. For this purpose, we analyzed an existing text message corpus [147]. The corpus contains a relatively small number of drug-related criminal messages (labeled), while the rest are regular text messages. This corpus is one of the few publicly available messaging corpora with metadata information such as *message time*, *contact id*, and *message type (incoming/outgoing)*. The availability of metadata information makes it easier to identify instances of delayed responses along with their explanations. The corpus contains $4,934$ messages, including 289 drug-related messages.

From this corpus, we identified categories of explanations people provide when communicating unavailability. Table 4 lists the identified categories and examples taken from the corpus. On linking these categories to sensors or features previously used in modeling messaging attentiveness, we established 13 categories of automated responses based on the

contextual information they contain. These are listed in Table 5 and represent the categories we evaluated in our survey. We discuss the corpus analysis in more detail in Section 4.3.1.

4.2.2 Survey Design

To understand people's perceptions of usefulness and comfort with sharing context-based auto-responses in one-to-one conversations, we designed and conducted a web-based survey [2]. It was distributed using Amazon Mechanical Turk. The study was reviewed and approved by our University's Institutional Review Board. Respondents were paid 3.50 USD for completing the survey. It included two major sections. One of the sections assessed respondents' perception as message senders, while the other assessed their perceptions as message recipients about the different types of automated responses listed in Table 5. These were guided by the corpus analysis discussed in section 4.3.1. Additionally, our survey included questions regarding users' demographic information and privacy concerns.

After introducing the survey, participants were randomly presented with questions corresponding to either message senders' or recipients' perspectives first, followed by the other perspective to balance out potential carry-over effects. Further, the respondents were not informed that they would be asked about the other role and were not allowed to go back and change their responses after completing a section.

4.2.2.1 External factors: message urgency and social relationship

Previous work has shown that a message can be received differently depending on the content of the message and sender-recipient relationship [73, 135]. Two important factors identified by prior work to impact communication and information sharing are the message's urgency and the social relationship between the sender and recipient. Social relationships have been found to affect the willingness to share information [124, 201, 214, 54, 109, 103]. Church and Oliveira, in their user study, pointed out that expectations vary based on the nature of the communication ("If I started a conversation and it's something urgent, then I expect them to respond immediately [50]. If the message isn't important, I personally don't

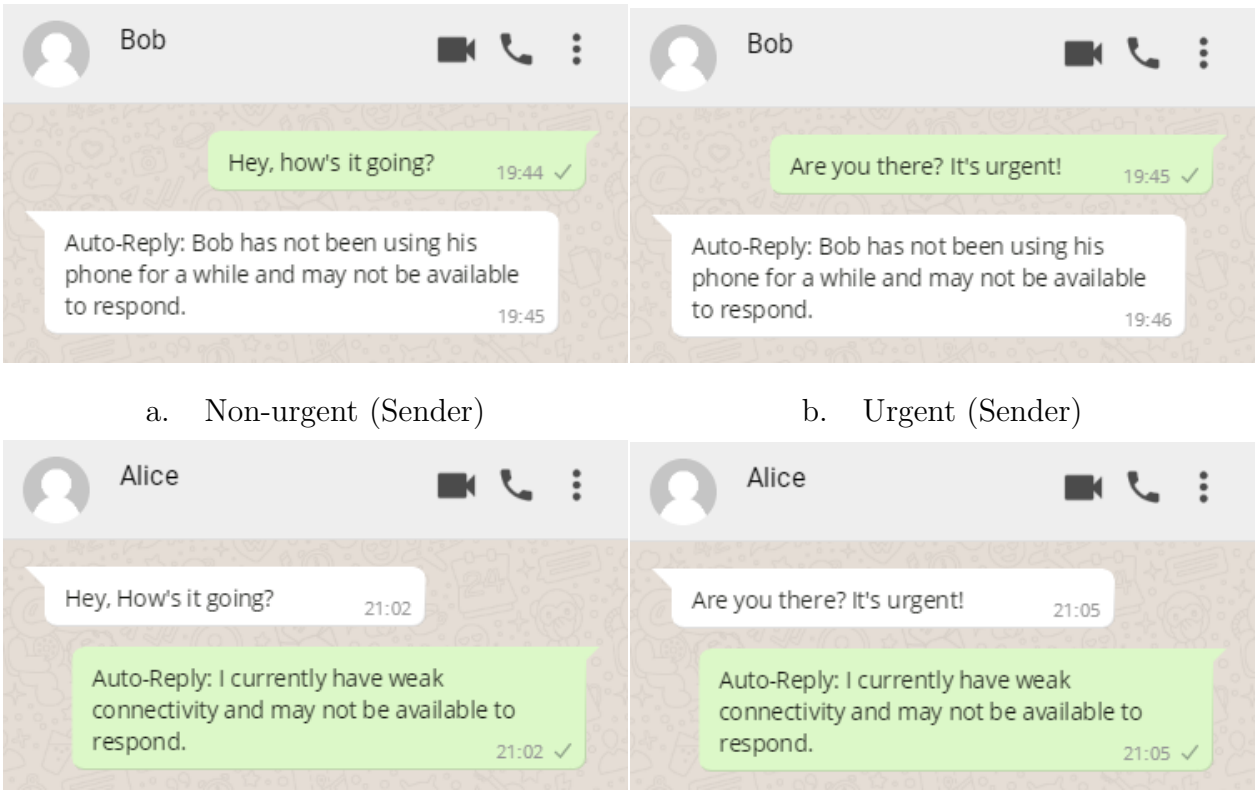
[2]Survey link: `https://people.cs.pitt.edu/~pranut/messaging_study/mstudy_survey.pdf`

care. I think people respond whenever they find time or whenever they feel like it"). Teevan and Hehmeyer also observed that communication is affected by whether users perceive a communication attempt to be urgent or important [185]. Thus, in the design of our survey, we account for the context of communication in terms of the strength of the social relations and the urgency of the message when evaluating different auto-response types.

4.2.2.2 Message Senders Perspective

Respondents in this section were asked how *useful* they find each category of contextual auto-response on a 3-point scale [92] (3-Useful; 2-Somewhat Useful; and 1-Not Useful). They were presented with four scenarios corresponding to the urgency of their message (i.e., Urgent vs. Not Urgent) and their relationship to the recipient of their message (i.e., Close or frequent contact vs. Distant or infrequent contact). Rather than having fixed relationship groups (such as *friends, families* and *coworkers*) as part of our evaluation, we chose to evaluate the effect of social relations based on closeness and frequency of communication since within social groups, the degree of closeness may vary and closeness has been observed to have a more profound effect than the social group on sharing behavior [201]. Figure 9a and 9b shows the sample screens presented to survey respondents. Here, the top bubble corresponds to their message to the recipient, and the bottom bubble corresponds to an auto-response. The respondents were asked to rate the usefulness of the auto-response in the four scenarios mentioned above.

After evaluating the *13* auto-response categories, the respondents were asked an open-ended question to provide any additional information from the message recipients that they would find useful. We also assessed how the granularity of information could influence their judgment of the usefulness of particular messages. For instance, an auto-response message related to a calendar event can include general information about the recipient being in a commitment or more detailed information about being busy with a *meeting* or a personal event such as *attending a game*. Similarly, auto-responses including location information can include only the general information such as *'not at home'* or *'at work'* or include more detailed information about the exact location.



a. Non-urgent (Sender)

b. Urgent (Sender)

c. Non-urgent (Recipient)

d. Urgent (Recipient)

Figure 9: Screen captures distinguished by urgency. (a) and (b) were shown during the Message Sender's block, and (c) and (d) were shown during the Message Recipient's block

4.2.2.3 Message Recipients Perspective

In this section, the respondents were asked to assume the role of recipient who received a message from one of their contacts and were asked to rate how comfortable (3-Comfortable; 2-Neutral; and 1-Not Comfortable) they were with the agent automatically sharing different types of contextual information about their state when they were deemed unavailable. Symmetric to the message senders block structure, these questions assessing comfort levels were asked by including communication context (urgency and relation). Figure 9c and Figure 9d show the respondents' screen samples when assessing their comfort levels. The top bubble corresponds to the incoming message from a contact, and the bottom bubble corresponds to

an automated response shared by the agent on the respondent's behalf.

Similar to the sender's block, the questionnaire in this section also included questions related to comfort levels associated with different categories and the granularity of shared information within an auto-response category.

4.2.2.4 Measuring Privacy Concern

One assumes that privacy concerns can be an important factor in the design of an auto-response agent, particularly concerning how comfortable the individuals are with sharing information about their situational context. Therefore, in the third section of the survey, we asked questions relating to the respondent's privacy views to measure their level of privacy concern. Prior work has shown that directly prompting respondents about privacy topics can lead to inflated levels of privacy concern or otherwise biased results [131, 30]. To avoid priming respondents in this manner, we purposefully asked users about their comfort with sharing and the utility of auto responses *before* collecting information on general privacy concerns.

We used the well-established second-order IUIPC (Internet Users' Information Privacy Concerns) scale to measure a respondent's level of privacy concern. This scale includes ten items based on three dimensions - *control* (over information), *awareness* (of privacy practices), and *collection* (of information) [130]. The ten items are measured on a seven-point scale ranging from *'strongly disagree' (1)* to *'strongly agree' (7)*.

Since this set of questions is directly asking about privacy, we expect the responses to be somewhat inflated [30], but that should not affect our analysis since we are only interested in measuring *relative* privacy concerns among respondents and relate that to their responses for usefulness and comfort levels in sharing different information through auto-responses.

4.2.3 Response Analysis

We utilized several statistical analyses in analyzing our survey responses. Here, we describe each analysis approach.

4.2.3.1 Factor Analysis

We evaluated respondents' perceptions about 13 categories of automated responses in four different contexts for a total of 52 items for usefulness and comfort. We performed factor analysis to determine if there is a latent structure as to how respondents rated these different categories and if some categories or subsets of categories measure the same aspect of perception for an automated response.

Usefulness. Our usefulness response dataset includes responses from 99 respondents for 52 items, each row representing a respondent's ratings for each auto-response category. To conduct the factor analysis, we restructured the data into 396 rows, where each row represents the response for a specific category of auto-response under a unique combination of urgency and social relation values (i.e., *frequent and non-urgent, frequent, and urgent, infrequent and non-urgent, infrequent and non-urgent.*

We then performed PCA (Principal Component Analysis) followed by varimax rotation on the transformed data to find components or factors which represent maximum variation in usefulness ratings and to identify any latent structures in how respondents rate different auto-response categories. To find the right number of components, we created the scree plot, which compared the eigenvalue with the different numbers of components and observed an 'elbow' with two components. The second factor having an eigenvalue of 1.090 also satisfies the Kaiser rule of selecting factors > 1.0 eigenvalue. The resulting two components explained 59.996% of the overall variance. Bartlett's test of sphericity was significant ($\chi^2(78) = 2627.052, p < 0.001$). The Kaiser-Meyer-Olkin Measure of Sampling Adequacy (KMO value) was 0.941, indicating that the strength of relationships between variables is high.

Comfort. We similarly structured the comfort data and conducted a similar factor analysis. Similarly, we observed two components to explain 59.104% of the variation. Bartlett's test of sphericity was significant ($\chi^2(78) = 2652.834, p < 0.001$). The Kaiser-Meyer-Olkin Measure of Sampling Adequacy (KMO value) was 0.936, again indicating the high strength of the relationship between variables.

The factor loadings for both usefulness and comfort are listed in Table 6. The variation

captured by component 2 was more significant than component 1 in the unrotated components returned by PCA. We interchanged them to better visualize the comparison with the usefulness factor loadings.

4.2.3.2 Cluster Analysis

To assess whether there are groups of respondents who are similar in their responses in terms of their perception of usefulness and comfort, we conducted cluster analysis on the restructured usefulness and comfort datasets with regression scores obtained from factor analysis. We used the *k-means* clustering approach with *k-means++* algorithm for selecting initial cluster centers. In this approach, after randomly selecting the first cluster center from all data points, the subsequent centers are chosen based on probability proportional to the squared distance from existing cluster centers.

To determine the optimal number of clusters k, we used the *elbow* approach by plotting the distortion score associated with different numbers of clusters. The distortion score computes the sum of squared distances from each point to its assigned center. The test range of k varied from 1 to 6. For both usefulness and comfort datasets, the *elbow* was observed for $k = 3$, which also had the highest silhouette average of 0.461 for usefulness and 0.464 for comfort.

4.2.3.3 Regression Analysis

We performed regression analysis to estimate the relationship of respondent attributes (age, gender, etc.) and message context (relation, urgency) with respondent preferences which could be linked to the group they belong to identified from cluster analysis.

Since a respondent rated the usefulness and comfort of sharing a response category multiple times for each communication context, our dataset includes repeated measures of an auto-response category for each respondent. Therefore, we used *GEE (Generalized Estimating Equations)*, which is a method used for parameter estimation for correlated data [121]. In addition to the consideration for dependencies between cases, GEE also does not have distributional assumptions [153].

We used a logistic response model with GEE, with the usefulness or comfort group association as the dependent variable. Each model included respondents' demographics (i.e., age, gender, employment, and education) along with the self-reported preferred method of communication, frequency of checking for unread messages, IUIPC metrics (i.e., control, awareness, and collection), and the message context (i.e., relation and urgency) as the independent variables.

4.3 Findings

In total, we received 101 responses to our online survey. We removed two responses for failing the *attention check* questions, general low-quality responses (copying question text in open-ended questions), or completing the survey in significantly lower time than the median time of all participants. Our final response set consisted of 99 responses, of which 70 respondents reported their gender as *male* (70.71%) and 29 reported as *female* (29.29%). In terms of the age distribution, 46 respondents reported their age between *18-34* (46.46%), 31 between *35-44* (31.31%), and 22 reported greater than *44* (22.22%). Respondents reported their education level as 16 *high-school or lower* (16.16%), 28 *college or 2-year degree* (28.28%) and 55 *4-year degree or higher* (55.55%). Employment was reported as 84 *employed full-time* (84.85%), and 15 reported *part-time or unemployed* (15.15%).

In terms of the preferred method of communication, among our respondents, 58 prefer *messaging* (58.59%), 24 *email* (24.24%), and 17 prefer *calling* (17.17%). We further asked respondents about how frequently they checked their phones for unread messages, with nine reporting *every 5 minutes or less* (9.09%), 51 reporting *couple or more times an hour* (51.52%) and 39 reporting *not more than once an hour* (39.39%).

In terms of *privacy concerns*, the measured concerns among the respondents along all three constructs, i.e., *control* ($\mu = 6.077, \sigma = .953$), *awareness* ($\mu = 6.350, \sigma = .840$) and *collection* ($\mu = 5.942, \sigma = 1.007$) were high.

Even though the Mechanical Turk worker population in the US has recently been reported to be predominantly male [63], our response set has a more considerable gender bias towards

Category	Example	Count
Location	"I am at work"	13
Physical/Motion Activity	"I'm on the way back to campus now"	8
Specific/Other Activity	"I will check later today. I am in a meeting."	10
Sleeping	"Hey sorry took nap"	6
Busy (no context)	"Sorry for not responding, got sidetracked"	7
In conversation	"Still at dinner, in a good conversation. Didn't forget about you."	3
Did not see/notice	"Sorry for getting back to you do late, left my phone on table in other part of house."	7
Weak Connectivity	"I am on the store, getting toilet paper. No reception. What's up?"	2
Low/Dead Battery	"Yeah, my phone died earlier"	3

Table 4: Categories of explanations identified from the forensics corpus with example and frequencies.

the male population. At the same time, other demographic measures align with the general Mechanical Turk population[3].

We also checked for order effects in respondent ratings which were insignificant for both, i.e., comfort (p =.861) and usefulness (p=.667).

4.3.1 RQ1: What types of automated responses can be generated using contextual information collected from an individual's smartphone?

Analyzing the messaging corpus described in Section 4.2.1, we identified 59 explanations that provide situational context for a recipient's unavailability. This includes explaining delays in responding to incoming messages (e.g., "Sorry, just got your text. My phone

[3]http://crowdsourcing-class.org/readings/downloads/platform/demographics-of-mturk.pdf

Category	Example
Busy (no context)	Bob is currently busy and may not be available to respond.
Activity	Bob is currently biking and may not be available to respond.
Connectivity	Bob currently has weak connectivity and may not be available to respond.
Battery Status	Bob's phone is currently low on battery and he may not be available to respond.
Location	Bob is currently at work and may not be available to respond.
Noise Level	Bob is currently in a noisy environment and may not be available to respond.
Charging	Bob's phone is currently charging and he may not be available to respond.
Proximity	Bob's phone is currently covered (in a bag or pocket) and he may not be available to respond.
App Status	Bob is currently playing a game on his phone and may not be available to respond.
Calendar	Bob is currently in a meeting with Joe and may not be available to respond.
Ringer Mode	Bob's phone is currently on silent mode and he may not be available to respond.
Phone Unused	Bob has not been using his phone for a while and may not be available to respond.
Call Status	Bob is currently on a call and may not be available to respond.

Table 5: Auto-response categories along with examples.

locked up and i had to do a hard reboot."); missing an incoming phone call (e.g., "What's up? Was in church when u called"); or being unable to communicate at the moment (e.g., "I will check later today. I am in a meeting."). We categorized each message based on the context provided in the explanation. Table 4 lists the identified categories, with examples of explanations in each category and the associated count in the corpus. The top recurring context provided in the explanation included location, activity, or physical motion: in 13 cases (22%), the explanations included location-based context, in 10 cases (17%) a specific activity, and in 8 cases (14%) some indication of physical motion. It should be noted that these explanations may not all be accurate or true and may be using deception to politely maintain the social connection [163, 81, 166].

In categorizing the situational context for communicating unavailability, we identified cases like *In conversation*, *Did not see/notice*, and *Specific/Other Activity* can have different interpretations depending on more detailed context. For example, the explanation *Did not see/notice* can be due to different reasons, such as the phone is on silent or *DND (Do Not Disturb)* mode or the phone is in a location not being noticed. Similarly, someone can be *In*

conversation either face-to-face or on the phone. In generating the messages, however, we posit that such interpretation can be left to the recipient of the auto-response. It is more appropriate for the auto-response only to include the relevant details [196, 67]. Given the classification and this assumption, our final auto-response categories are listed in Table 5. These categories can be directly linked to sensors or features used in previous works in modeling messaging attentiveness [157, 93]. Categories such as *Location*, *Physical Activity*, *Weak Connectivity*, and *Low Battery* can be inferred directly from an individual's phone sensors, whereas more complex categories such as a *Specific/Other Activity*, can either be inferred by the auto-response recipient from an individual's *Calendar* or the application they are using on their device (*App Status*). Similarly, *Did Not See/Notice* can be inferred from *Ringer Mode*, Last phone use (*Phone Unused*), whether the phone is in pocket/bag (*Proximity*), whether it is *Charging* and what is the surrounding *Noise Level*.

These categories can further be classified based on the information they represent. We define *user-state* categories as those which describe the state or environment of the user. In contrast, *device-state* categories indicate the characteristics or state of the user's device. From the categories described in Table 5, *Activity, Location, Noise Level, Calendar, App-Status, Busy (no context)* and *Call-status* would be classified as user-state categories while *Connectivity, Battery Status, Charging, Proximity, Ringer Mode* and *Phone Usage* form the device-state categories. For instance, a user's calendar describes their current schedule, and their activity describes their physical state (walking, running, etc.). Similarly, while the *Busy* category lacks additional context, it still describes the user rather than their device. In contrast, battery state and charging categories describe the current power state of the device.

Some categories of explanations have also been mentioned in work by Voida et al. Quotes from participants in the study included situational context such as *In-conversation* ("talking with Karen...sorry for delay in not talking"), engaged in *another activity* ("...was reading email on my laptop") and *location* (not at home) ("...I'm going to head home right now...can we talk later?). However, Voida et al. did not categorize the types of contextual indicators included in these explanatory messages nor explore the possibility of automatically constructing contextual replies. Cho et al. analyzed the types of manual statuses set by participants

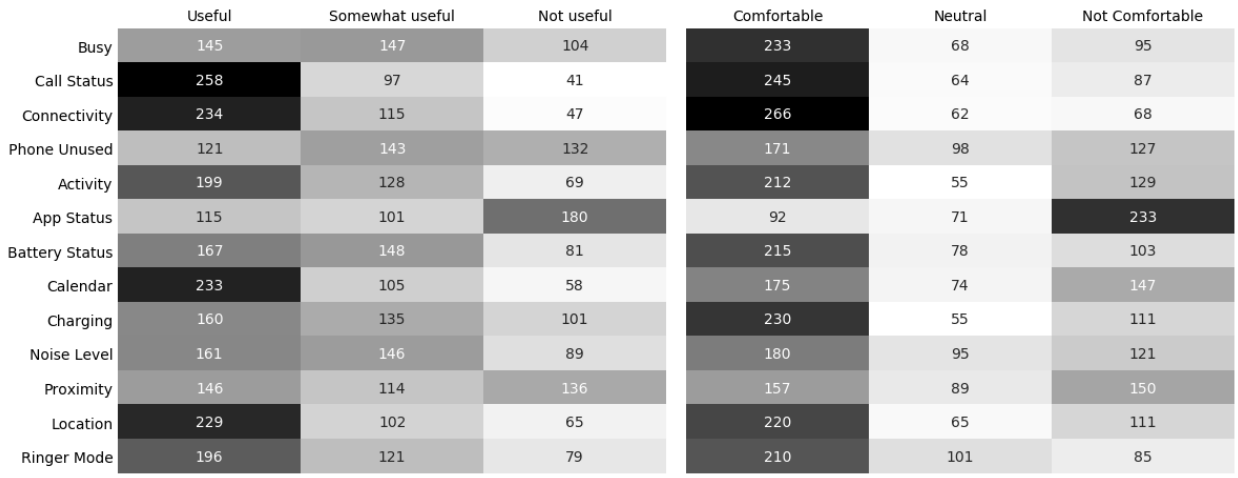


Figure 10: Plot showing overall ratings for different auto-response categories.

to be automatically shared for incoming messages. They observed six high-level categories of statuses set by participants, i.e., *Activity*, *Availability*, *Emotional/Physical*, *Location*, *Conversation* and *None*, which are similar to the categories of explanations we identified in our analysis. This further validates our finding into what people already (manually) share when communicating unavailability but also indicates what people might be comfortable sharing automatically on incoming messages. However, their work was limited to close contacts (friends or couples). It did not explore or evaluate the utility of these categories of status messages for message senders, nor how these messages can be generated and shared without manually setting preferences related to each category. We build upon this prior work by first trying to understand (i) the types of context that are useful in explaining unavailability and (ii) the availability of sensor data on the phone to facilitate the creation of auto-responses explaining recipient unavailability.

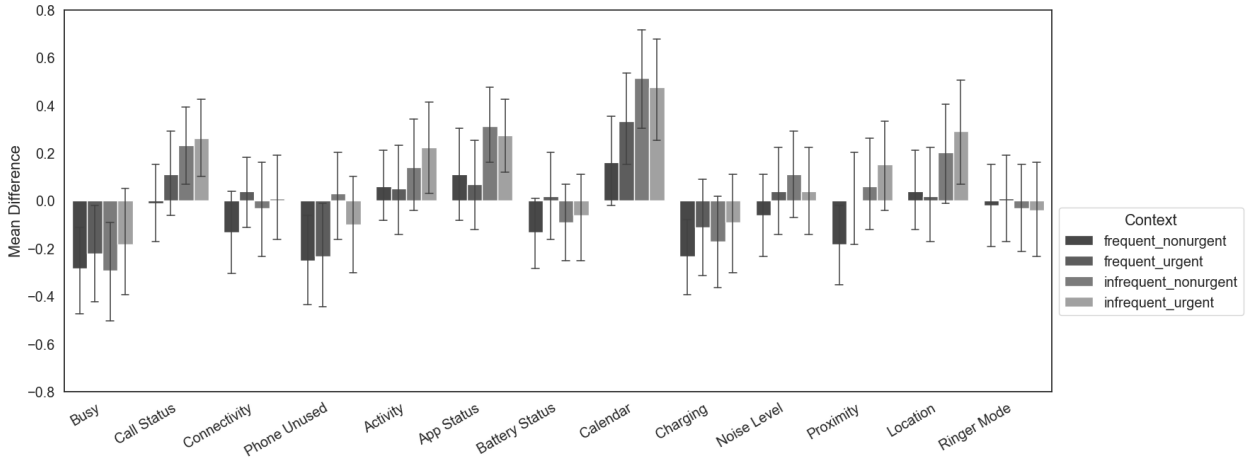


Figure 11: Plot showing the differences between usefulness and comfort ratings for all categories. The error bars represent a 95% confidence interval obtained using bootstrapping.

4.3.2 RQ2: What is the perceived usefulness and comfort in sharing different categories of automated responses?

Figure 10 shows the comparison of the usefulness and comfort ratings for all response types. It can be observed that the usefulness ratings of different categories are more spread out than comfort ratings which are more aligned toward high comfort for most of the categories. This is surprising given the high average privacy concern of our respondents.

Comparing the ratings of different categories, *connectivity* and *call-status* were comparatively perceived as more *useful*, as well rated more *comfortable* in sharing. The utility of the recipient's *connectivity* state, in particular, is an interesting result since previous works on evaluating the value of sharing different contextual information didn't consider the recipients' *connectivity* state since the focus was on reducing disruptions for the callee [109, 103]. *Read-receipts* in applications such as WhatsApp and Facebook also include a state that represents whether the message has been *delivered* to the recipient (e.g., double white ticks in WhatsApp), though not everyone might be aware or may interpret it as such [88]. This result points to the perceived usefulness of explicitly making the senders aware of the *connectivity* of the recipient. Further comparing our results to other works, in terms of usefulness, our

respondents rated *calendar* and *call-status* higher compared to *phone-usage*, whereas in the findings of Knittel et al., the utility of *phone-usage* was rated higher [109]. In terms of comfort, *location* and *activity* were rated lower compared to *call-status*, similar to Khalil et al. [103]. Respondents for the study by Knittel et al. also rated sharing *App-Status* lower in terms of comfort. In contrast, *ringer-mode* and *abstract location* had higher disclosure rates pointing to some similarities in terms of comfort of sharing context irrespective of the communication medium [109]. As we will see in Section 4.3.3, the perception towards different categories varied further based on communication context (social relation and urgency), which was only partially considered in the works by Knittel et al. [109] and Khalil et al. [103].

In terms of alignment between usefulness and comfort ratings, Figure 11 visualizes the differences in how respondents rated *usefulness* and *comfort* of different auto-response categories. Directly comparing usefulness and comfort ratings may not accurately represent differences given that Likert scale ratings may not be perceived equidistant from one another by respondents [180]. More so, in our survey, the middle point for the usefulness rating scale was *somewhat useful*, which might tend towards positive polarity compared to *neutral* in the comfort rating scale, though the effect this has would be less pronounced than the perceived polarity at the extremes of the Likert scale [117]. Nevertheless, analyzing the mean absolute difference would still give some indication as to where usefulness and comfort ratings differ the most, which we observed to be low (i.e., ranging from .50 for *call-status* to .78 for *calendar*). At the same time, the standard deviations were observed to be high (i.e., ranging from .670 for *battery-state* to .771 for *calendar*). On average, categories such as *Busy, Phone-unused, Battery-status* and *Charging* were rated higher on comfort than usefulness. In contrast, categories such as *Call-status, Activity, App-status* and *Calendar* were rated higher on usefulness than comfort in sharing, indicating the existence of varied opinions for some categories concerning utility and comfort in sharing. Further, the variation between usefulness and comfort ratings for different categories was affected by the communication context (social relation and urgency). For instance, usefulness and comfort associated with sharing *Calendar* are more equally aligned for *frequent* contacts than *infrequent* contacts, which relates to a previous finding that relationship category impacts in what way and with whom people share their calendars with [186]. These results indicate that

Usefulness			Comfort		
Category	Component 1	Component 2	Category	Component 1	Component 2
Calendar	.793	.103	Calendar	.847	.140
Call-status	.784	.233	Call-status	.556	.495
Location	.780	.269	Location	.626	.382
Activity	.621	.469	Activity	.711	.326
Busy	.596	.365	Busy	.380	.497
Noise-level	.447	.613	Noise-level	.471	.603
App-Status	.024	.783	App-status	.640	.278
Connectivity	.615	.367	Connectivity	.161	.788
Ringer Mode	.595	.465	Ringer Mode	.468	.538
Battery	.317	.671	Battery	.221	.804
Charging	.423	.664	Charging	.343	.786
Proximity	.328	.763	Proximity	.359	.671
Phone-state	.456	.614	Phone-state	.456	.636

Table 6: Factor Loadings for Usefulness and Comfort ratings.

our respondents had varying preferences regarding the perceived utility of different response types, and communication context affected their perceptions.

4.3.2.1 Variation in preferences based on whether a category represents User-state or Device-state

The factor analysis results are presented in Table 6. It can be observed that categories *Calendar, Call-status, Location, Activity* and *Busy* show higher loadings on component 1 than component 2. Whereas categories *Battery, Charging, Proximity*, and *Phone-state* have higher loadings for component 2 than component 1 for both usefulness and comfort. Most categories with higher factor loadings in component 1 represent user-related information, i.e.,

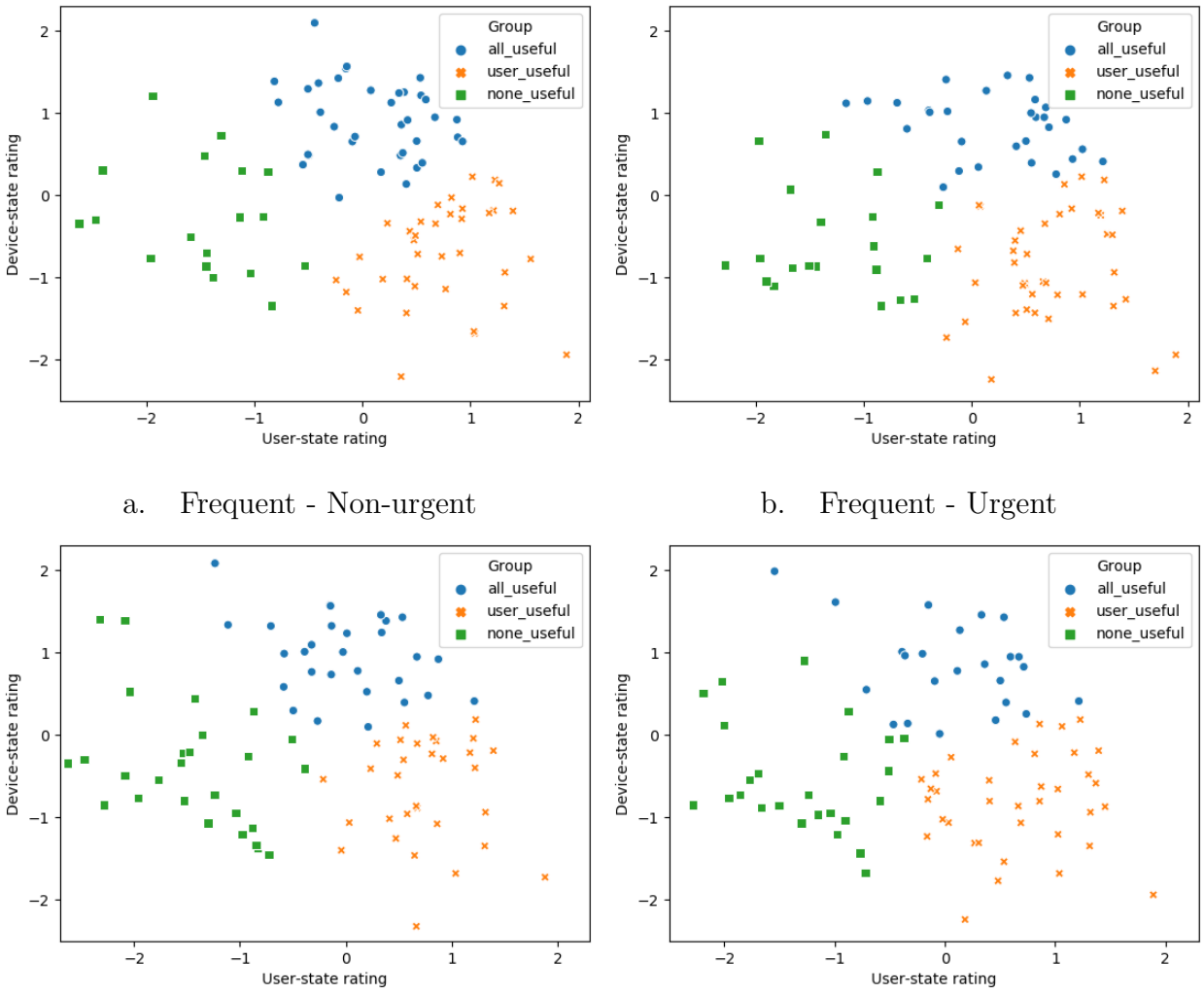
user-state categories. In contrast, most categories with higher factor loadings for component 2 represent device-state categories as mentioned in Section 4.3.1. This result suggests how people's perception of potential auto-responses depends on the distinct context of user-related information versus device-related information. Some categories, though, such as *App-status*, did not correspond to a single component for both usefulness and comfort, i.e., the respondents' usefulness ratings of app-status were closer to device-state categories than user-state categories. In contrast, for comfort, *App-status* was rated similarly to user-state categories. This means that respondents found the usefulness of app status similar to device-state categories. In contrast, the comfort in sharing was similar to the comfort they felt sharing other user-state categories. Connectivity ratings showed a converse pattern, i.e., respondents rated the usefulness of connectivity similar to user-state categories. In contrast, the comfort in sharing was rated similarly to other device-state categories. For the rest of this dissertation, we will refer to component 1 as *user-state* categories and component 2 as *device-state* categories.

4.3.3 RQ3: Emergence of user-groups with varying preferences in relation to the communication context

The standard deviation from the mean for *usefulness* ratings of different categories varied from .675 for *call-status* to .848 for *app-status* and for *comfort* ratings, varied from .771 for *connectivity* to .905 for *activity*. This indicates that, with respect to the usefulness of messages and participants' comfort in sharing the information, there can be high variation among the respondents. Through cluster analysis described in Section 4.2.3.2, we identified user groups with varying preferences for different categories of auto-responses.

4.3.3.1 Usefulness

Figure 12 shows the plot of the identified clusters in different communication contexts, i.e., *Frequent and non-urgent* (Figure 12a), *Frequent and urgent* (Figure 12b), *Infrequent and non-urgent* (Figure 12c) and *Infrequent and urgent* (Figure 12d). Table 7 lists the number of respondents in each cluster for different contexts.



a. Frequent - Non-urgent

b. Frequent - Urgent

c. Infrequent - Non-urgent

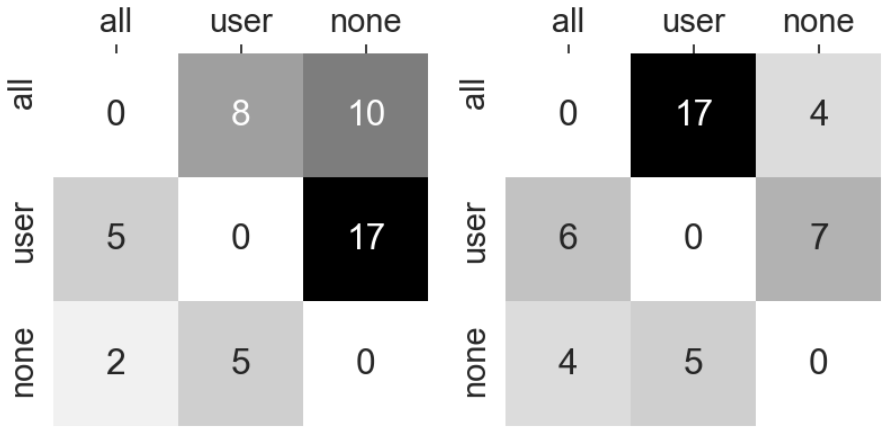
d. Infrequent - Urgent

Figure 12: Scatter plot visualizing user groups based on the usefulness ratings for different types of categories identified from Factor Analysis.

The *x-axis* represents the user-state categories rating, and *y-axis* represents the device-state categories rating. Higher values on the *x-axis* represent high rating for user-state categories (activity, location, etc.), and higher values for *y-axis* represents higher ratings for device-state categories (phone-status, battery-status, etc.). As presented in the plots, one of the emergent groups (depicted in blue dots) has comparatively higher ratings for

Context	all_useful	user_useful	none_useful
Frequent, Non-urgent	44	37	18
Frequent, Urgent	38	41	20
Infrequent, Non-urgent	38	32	29
Infrequent, Urgent	33	37	29

Table 7: Number of respondents in each group for different contexts.



a. Change in relation b. Change in urgency

Figure 13: Changes in respondents' *usefulness* group association (y-axis-*from*, x-axis-*to*) with change in communication context (a. frequent to infrequent and b. non-urgent to urgent). *all* represents *all_useful*, *none* represents *none_useful* and *user* represents *user_useful* groups.

both user-state and device-state categories which we will refer to as the *'all_useful'* group. Whereas another group (depicted in green squares) has lower ratings for both user-state and device-state categories which we will refer to as *'none_useful'* group. The third identified cluster (depicted in orange crosses), has a higher rating for the user-state category but a lower rating for the device-state category, which will be referred to as *'user_useful'* group.

Further, we observed that respondents' group association varied based on the commu-

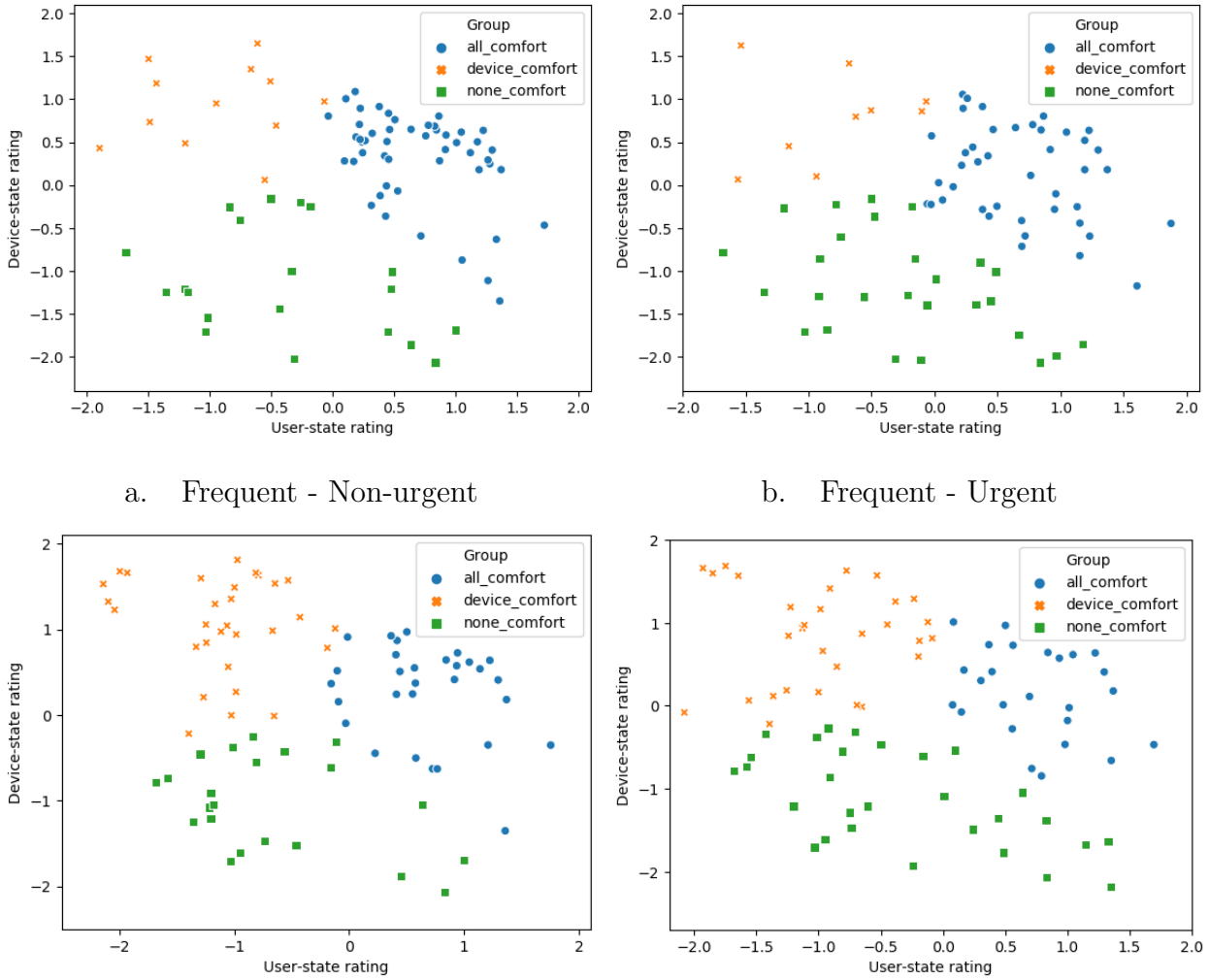
nication context, and varying social relations or message urgency resulted in respondents moving from one group to another. Figure 13 shows the change in respondent group association with the change in the communication context. For instance, the group association of 27 respondents switched to the 'none_useful' group when considering infrequent contacts indicating that for these respondents, both types of contextual information (user-state and device-state) were perceived as not useful when trying to communicate with distant contacts. Similarly, 22 respondents switched group association to the 'user_useful' group for urgent messages, indicating the importance of knowing the user state in urgent situations.

4.3.3.2 Comfort

Figure 14 shows the plots of the identified clusters in different communication contexts, and Table 8 lists the number of respondents in each cluster for different contexts.

Like the usefulness plots, the *x-axis* represents the user-state category rating while the *y-axis* represents the device-state category rating. The first identified cluster (depicted in blue dots) has comparatively higher comfort ratings for both user-state and device-state sharing, which we will refer to as the *'all_comfort'* group. In contrast, the second cluster (depicted in green squares) has a comparatively lower rating for both user and device state sharing categories which we will refer to as *'none_comfort'* group. The third cluster (depicted in orange crosses) has higher ratings for the device-state category and lower ratings for the user-state category, which we will refer to as *'device_comfort'* group.

We also observed that respondents' group association changed when the communication context was varied for comfort groupings. Figure 15 shows how the group associations change in different communication contexts. For instance, nearly half of all respondents' group associations changed from 'all_comfort' to 'device_comfort' group when considering infrequent contacts indicating that these respondents felt comfortable sharing only device-state categories when communicating with distant contacts. Similarly, 23 respondents switched group association to the 'none_comfort' group for urgent messages indicating that respondents were uncomfortable sharing both user-state and device-state categories in urgent situations. We elaborate further on this in the discussion section (Section 4.5).



a. Frequent - Non-urgent

b. Frequent - Urgent

c. Infrequent - Non-urgent

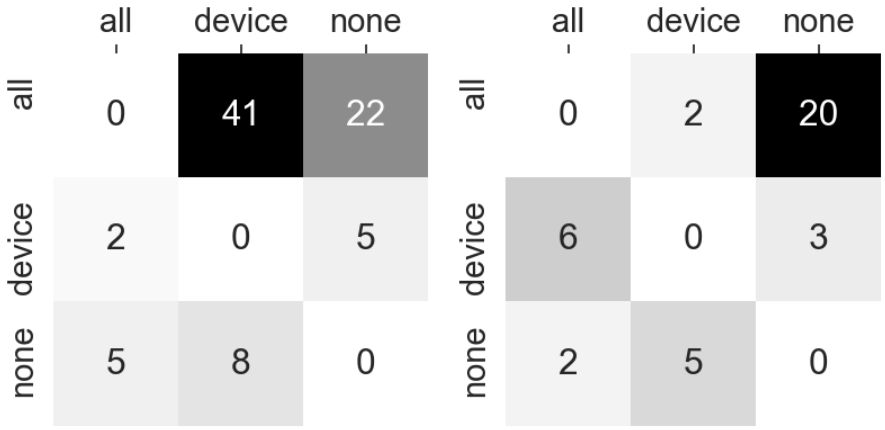
d. Infrequent - Urgent

Figure 14: Scatter plot visualizing user groups based on comfort ratings for different types of categories identified from Factor Analysis.

Overall, there is a more considerable shift in group association with context change in comfort ratings compared to usefulness ratings. In general, we observed that respondents' found both user-state and device-state more useful and were more comfortable sharing those with frequent contacts in non-urgent contexts. Whereas, in urgent contexts, some respondents perceived user-state information to be more useful. While for infrequent contacts,

Context	all_comfort	device_comfort	none_comfort
Frequent, Non-urgent	65	13	21
Frequent, Urgent	58	10	31
Infrequent, Non-urgent	37	32	30
Infrequent, Urgent	30	33	36

Table 8: Number of respondents in each group for different contexts.



a. Change in relation b. Change in urgency

Figure 15: Changes in respondents' *comfort* group association with change in communication context (y-axis-*from*, x-axis-*to*). *all* represents *all_comfort*, *none* represents *none_comfort* and *device* represents *device_comfort* groups.

some respondents rated device-state categories as more comfortable in sharing compared to user-state categories.

These observations indicate that respondents' had varying preferences for different auto-response categories, and preferences were affected by the communication context.

4.3.4 RQ4: Role of user-attributes and communication context on preferences

We observed that communication context affected respondent preferences for different auto-response categories (Section 4.3.3). The parameter estimates from regression analysis (Section 4.2.3.3) indicate how significant the effect of user attributes and communication context was in the perceived usefulness and comfort associated with different auto-response categories.

4.3.4.1 Usefulness

We observed that relation was a significant factor in usefulness group association ($\beta = .373, Exp(\beta) = 1.451, \chi^2(1) = 7.720, p = 0.005$). This suggests that both user and device-state automated responses are *1.5* times more likely to be found useful when coming from frequent contacts. Employment status was also marginally significant ($\beta = .913, Exp(\beta) = 2.492, \chi^2(1) = 3.770, p = 0.052$) with full-time employed being *2.5* times more likely to find both user and device-state based automated responses useful.

Neither message urgency nor the interaction effect between social relation and message urgency were significant factors in determining the usefulness group association. This implies that message urgency did not significantly affect the perception of usefulness for different auto-response categories. Further, other attributes such as gender, age, education, and IUIPC were also not significant factors in the usefulness group association.

4.3.4.2 Comfort

We observed that gender ($\beta = .846, Exp(\beta) = 2.330, \chi^2(1) = 5.658, p = 0.017$) was a significant factor, and male gender was 2.3 times more likely to be comfortable sharing both user and device-state based automated responses. A similar observation was made by Khalil et al. [103] and Knittel et al. [109], where men were more likely to share context compared to women. Relation ($\beta = .916, Exp(\beta) = 2.499, \chi^2(1) = 21.350, p < 0.001$) was also a significant factor, and respondents were 2.499 times more likely to be comfortable sharing both user and device state auto-responses with frequent contacts. This observa-

tion confirms social relations' importance in information disclosure [109, 103, 214, 124, 54] also holds when sharing contextual information through auto-responses. Further, urgency ($\beta = .330, Exp(\beta) = 1.391, \chi^2(1) = 5.030, p = 0.025$) was also a significant factor, with respondents being 1.391 times more likely to be comfortable with sharing both user and device-state auto-responses for non-urgent messages compared to urgent messages. We further discuss the implications of message urgency in Section 4.5.1.

Similar to usefulness, the test for model effects indicated insignificant interaction effect between social relation and message urgency. Other attributes such as age, education, employment, and IUIPC were also insignificant in the comfort group association.

4.4 Predicting Usefulness and Comfort preferences

Automated sharing may raise concerns about unintended or unwilling information disclosure [201]. Further, the auto-response sent should also be perceived as 'acceptable' by the sender to communicate unavailability [166] effectively. Thus, it is important to account for user preferences with regard to usefulness and willingness to share different auto-response categories. If the user is responsible for setting up their preferences, that would add an additional burden of creating and maintaining policies on the user [115, 201]. Rather, for the agent to be accepted, it should be able to learn from the users' context and adapt [133]. Initial preferences for the user can be set based on the group to which they belong in different communication contexts. For instance, given a communication context, if it can be determined that the user belongs to the group 'none_comfort', then it can be implied that they are not comfortable sharing either device or user-state categories in that context.

In Section 4.3.4, we presented that along with the communication context, the user group association was also affected by demographics such as gender for comfort groups and employment for usefulness groups. Relation strength between contacts can be inferred by looking at the frequency of message exchanges [201], and urgency can either be determined using NLP techniques on messages or utilizing an 'important' flag like used in emails [87]. However, demographic information may not always be available, and the collection and/or storage might

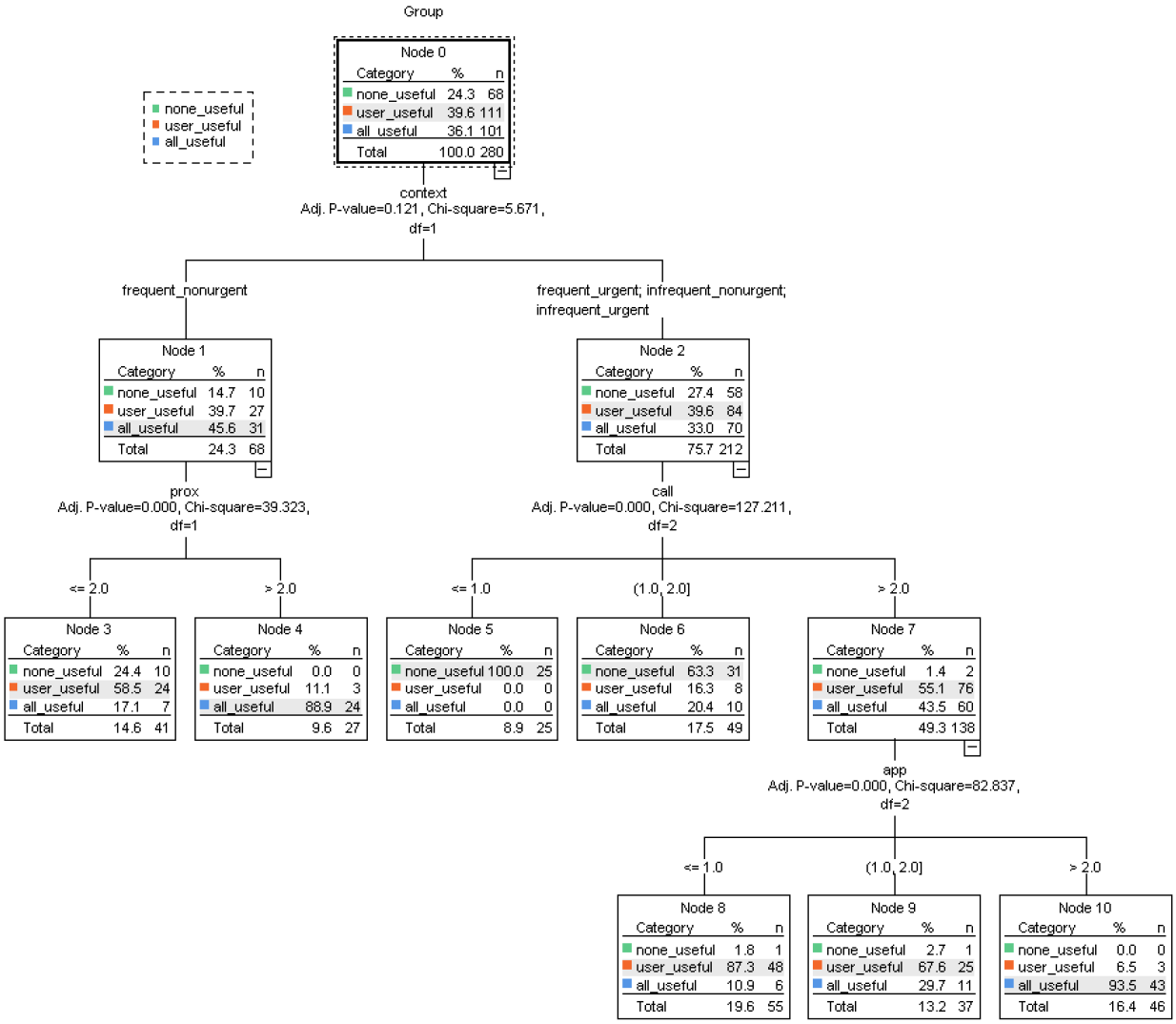


Figure 16: Decision tree visualization for predicting Usefulness group association

raise privacy concerns, as this information might be considered sensitive [11]. Another way to get initial user preferences is to ask the user to rate all auto-response categories for all communication contexts. This might not only be too cumbersome for the user, but user preferences might change over time. Thus, in this section, we evaluate how accurately the usefulness or comfort group association can be predicted and ratings for which categories or

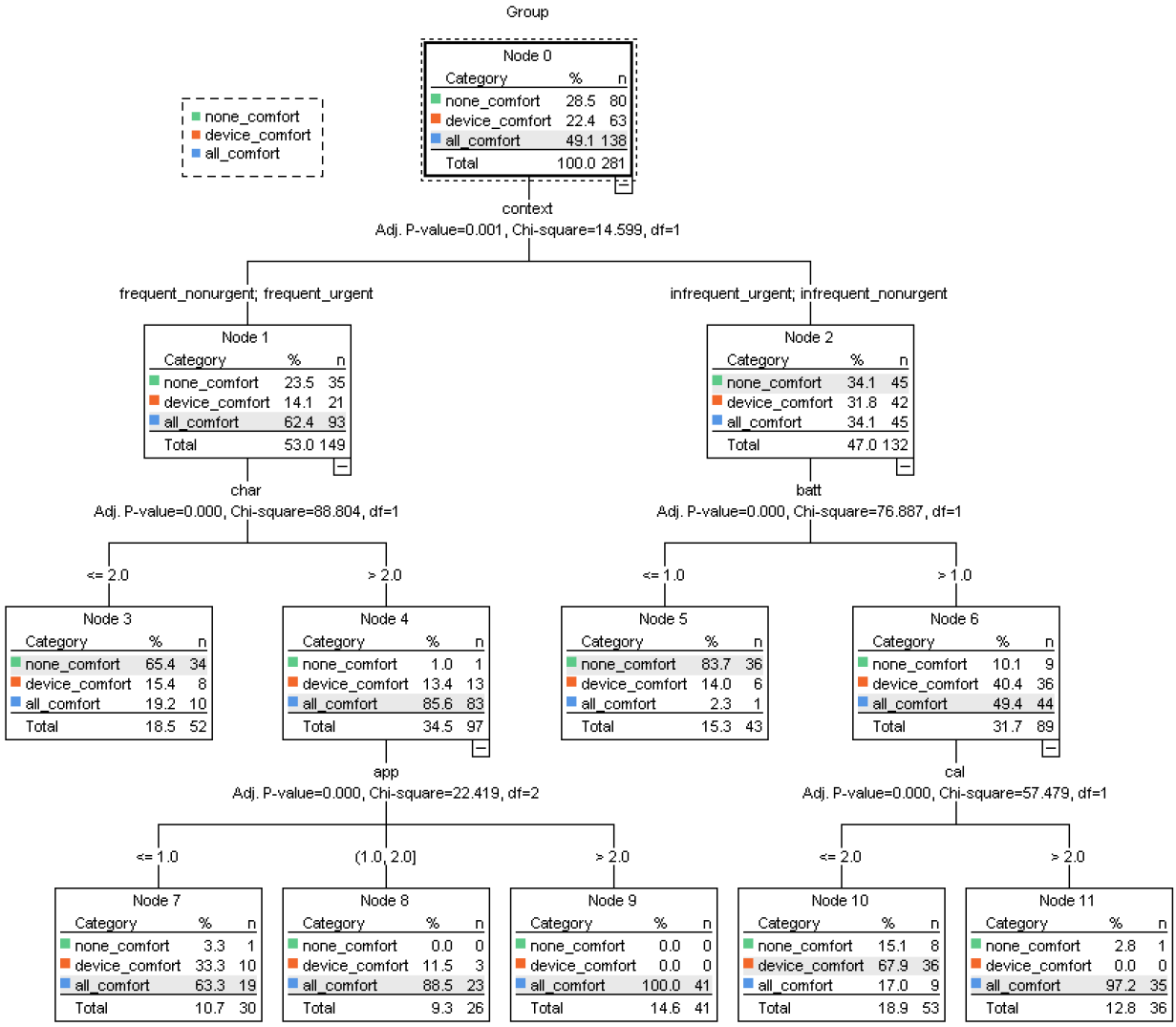


Figure 17: Decision tree visualization for predicting Comfort group association

subset of categories can best discriminate between different group associations for usefulness and comfort groups.

For this purpose, we built decision trees for the transformed usefulness and comfort ratings response sets with the group associations as the class or ground truth for each case. One advantage of using decision trees over other classification techniques is easier interpretation

and implicit feature selection [80]. Further, rules can easily be generated by traversing from the root node to the leaf nodes of a decision tree. Since the cases in the usefulness and comfort response sets are not independent, we would need to build 4 models to represent each communication context. Rather than doing that, we split the first or root node of both usefulness and comfort decision trees based on the communication context. We used *CHAID* growing method with 60 minimum cases in the parent, 25 in child nodes, and *max_depth* of 3 to prevent overfitting. Using a $70-30$ training testing split, we got (1) 78.6% training and 74.1% test accuracy for Usefulness group classification with the model asking 3 questions for usefulness rating of proximity, call-status, and app-status; (2) 79.7% training accuracy and 75.7% test accuracy for comfort group classification with the model asking four questions for comfort ratings of sharing battery-status, charging, calendar, and app-status. This indicates that by knowing the message context (urgency and social relation), the agent can make a fairly accurate prediction of an individual's usefulness and comfort group association.

The decision tree models for usefulness and comfort group classification are shown in Figure 16 and 17, respectively. The root node in both trees represents the set of all instances (ratings) used in the training phase (280/396). The *usefulness* decision tree splits on three nodes representing *proximity, call-status* and *app-status*. In comparison, the comfort decision tree splits on four nodes *battery-status, charging, calendar* and *app-status*. Knowing a user's usefulness and comfort in sharing preferences for these subsets of auto-response categories, along with the communication context, would allow the agent to predict their initial group associations. As an example, based on the usefulness model, for *frequent contacts in non-urgent* situations, if the message sender rates proximity (device-state category) as *not useful* or *somewhat useful* (≤ 2) they would be classified in the *user_useful* group indicating that they find only user-state categories useful in that communication context. Similarly, for comfort group classification, for *infrequent contacts in non-urgent* situations, if the user rates battery (device-state category) as *neutral* or *comfortable* (> 1) and calendar (user-state category) as *neutral* or *not comfortable* (≤ 2) then they would be classified into *device_comfort* group indicating they are comfortable sharing only device-state categories. As discussed in Section 4.3.2.1, app-status usefulness was rated similar to device-state categories of component 2, and comfort rating was similar to user-state categories of component 1. This can

also be observed with the usefulness group classification tree, where one branch represents the rating of app status. A high rating corresponds to classification to *all_useful* group, whereas a low rating corresponds to *user_useful* group, i.e., finding categories of user-state (component 1) useful vs. finding all categories useful. While the app-value rating does not change the classification of comfort group association, it does affect the confidence associated with the prediction as the number of instances for *all_comfort* group association reduces in the leaf with lower ratings of app-value category.

While this classification method can initialize a user's preference, further improvements can be made using methods like reinforcement learning, where the model can be updated by asking the user to rate the auto-responses as they are sent over time. This can create a more personalized model based on user preferences. Other dimensions of customization in terms of granularity of information in an auto-response and finer or customized contact groups can be considered to further improve the model and understand the associated preferences. These are beyond the scope of this paper and will be investigated in future work.

4.5 Discussion and Summary

The social information processing theory (SIPT) points to people using any available cues in CMC (Computed Mediated Communication) to make decisions about others and form relations [67, 196]. Limited or incomplete communication cues may lead to unwarranted speculations in message senders such as *'feeling ignored'* [88]. Providing more relevant context when the recipient may not be available to respond may allow senders' to make better inferences about the recipients' state and also allow for better management of expectations. For instance, when detecting an instance of unavailability for a message recipient who has not been checking their phone, an agent can respond by saying that the recipient *'has not been using their phone for a while'*. This may relieve the sender that they are not being ignored; the recipient has just not been looking at their phone. When constructing such replies, it is important to consider the perceptions surrounding different response types. In particular, message senders should find the information in an auto-response useful, and the

recipient should be comfortable sharing this information.

This work contributes to the growing body of CSCW and HCI research on awareness in remote communication. In particular, this work improves the understanding of automatically acquired context in informing availability. Our analysis identifies important factors for perception and how they can be used to set preferences for individual users. Combined, the results of this work augment the body of knowledge for designing awareness systems that are cognizant of the communication situation, the type of awareness information, and the preferences of a specific user.

4.5.1 Design Implications

In this section, we present some implications for the design of an agent-based availability manager based on our findings.

DI1: An agent-based availability manager should be cognizant of user and device state responses. Our findings indicate that the perceptions towards different categories of auto-responses varied based on whether a category represented the *user-state* or the *device-state* contextual information. While we evaluated a limited set of categories based on information that a user's smartphone can directly capture, as technology evolves, more information can be made available through additional sensors or in combination with other devices. For instance, multiple respondents noted in the open-ended question asking about other information that they would be comfortable sharing, that they would like the agent to share when they are *'sleeping'* (e.g., "I think I would be comfortable with an auto-response stating that I am sleeping.") which represents user-state and requires making inferences using information from multiple sensors [44]. The distinction between user and device-state categories would allow adding more response categories—beyond those studied in this paper—without needing to evaluate the utility of every new type of contextual auto-response.

DI2: An agent-based availability manager should account for communication context when determining the type of contextual cue to utilize for communicating unavailability. Another important finding was related to the communication context (i.e., social relationship and message urgency). We observed that social relationship has a

significant role in both the perceived usefulness and comfort of sharing an auto-response category. Our respondents were more likely to share both user-state and device-state based auto-responses with *close or frequent* contacts rather than with *distant or infrequent* contacts (e.g., "I'd be comfortable with just about anything except for people I don't know/talk to often knowing that I might be ghosting them while using my phone like gaming, YouTube, etc."). Message urgency, too, played a significant role in determining respondents' comfort level associated with both categories of auto-responses, with respondents being more likely to be comfortable sharing both user-state and device-state for non-urgent situations rather than for urgent situations. This observation can be attributed to the fact that people are likely to be more receptive to communication if they perceive it as urgent or important [8, 185, 39] and would probably like to be able to attend to urgent matters themselves (e.g., "I would be mostly comfortable with anything so long as it isn't urgent. If something were urgent, I would much prefer to be notified about it via some kind of emergency alert rather than an auto-response to an urgent message").

DI3: An agent-based availability manager should be aware of individual preferences for sharing different contextual cues. We also observed individual variations concerning the perception of auto-responses. As for a given communication context, some respondents were uncomfortable sharing any auto-responses category, whereas some were comfortable sharing only a device-based context. Similarly, some respondents found all types of contextual information useful for a given communication context, while others found only user-based contextual information useful. The respondents also differed in how much information they would be willing to share with their contacts as some were open to sharing finer details (e.g., "I would be comfortable sharing most any information with close contacts, like who I'm with, where I'm at or what I'm doing. I'd be comfortable with telling my close contacts what time I'll be available again for them to try me again at a more convenient time if I'm doing something I do on a schedule or calendar..."). In contrast, some preferred limiting the amount of details that would be shared (e.g., "The primary concern is that there would be an expectation to respond after I'm done with the activity. So, any activity that is timed and wouldn't take that long to do would be uncomfortable.").

4.5.2 Practical Considerations

Accuracy of auto-responses: While the focus of this work was to understand perceptions surrounding the utility of different contextual auto-response types, the correctness and accuracy of response are also important. An accurate availability model can extract relevant information about contextual features affecting a user's availability [93, 94]. These features can then be evaluated to identify which contextual information has the most influence on the recipient's unavailability and is also considered comfortable to share by the recipient and would be perceived as useful by the sender. The sender's preferences may differ from the recipient's regarding what they constitute as useful and need to be managed/tracked centrally by the agent designer. Another approach would be for the agent on the sender's phone to send their preferences when communication is initiated.

Burden on recipient: One of the goals of this research is trying to minimize the burden on recipients, whether it is due to manually setting an unavailability status or explaining delays in responding to messages. From the point of view of the recipient, they would not have to take any action in case of urgent messages since the urgent signal is directed towards the agent rather than the recipient, who then decides on what information to share from the recipient's context (as discussed in Section 4.4). This is in contrast to prior work by Teevan et al. [185] and Avrahami et al. [8], where the urgency context was directed towards the recipient for them to make an informed decision on whether to take the call or not.

Privacy and Mutual Awareness: From a privacy perspective, information should only be shared when communication is initiated, and the recipient is deemed unavailable. This way, even though more information is being shared about the recipients' context, it is limited in terms of accessibility compared to existing cues in messaging applications and awareness displays proposed in other works [57, 109]. Also, since information is only being shared when message senders initiate communication, the recipient is **aware** of the information that has been shared and with whom [46, 144].

Appropriate use: Finally, as mentioned in Section 4.3.1, people sometimes use deception in the form of butler lies to signal or explain unavailability [81, 163]. People also tend to appropriate technologies to better suit their needs which in terms of mobile messaging

might be by turning off 'last seen' [162] or by not viewing a message to avoid setting off 'read receipts' [88]. But this tailoring is often based on situational context, e.g., contact in question or contacts' setting (personal or professional) and the characteristics of messaging applications [162]. At the same time, manipulating a messaging agent would require understanding what the agent has already learned about their behavior and what can be done to manipulate the agent's behavior to the desired outcome. This might be too complex and could be a limitation in terms of the flexibility of the technology for users to appropriate. Further investigation is needed to assess how people would adapt to using a messaging agent for managing unavailability.

4.5.3 Limitations

Our evaluation of different contextual auto-responses only considered a single response category at a time. It is possible that the combination of multiple categories can have a higher utility in terms of explaining unavailability than individual categories. For instance, *physical activity* information together with *noise levels* may allow the message sender to infer more about the recipients' state. The number of possible combinations would make evaluation overly complex and time-consuming, especially for a survey-based study. Further, when asking the respondents to rate a category of contextual auto-reply, while we did mention the possible values that category could represent—e.g., *activity (driving, biking, walking, etc.)*—the rating of the respondent may have been biased towards the example which was presented through the sample screen (Figure 9). Further, the middle point for rating categories for usefulness and comfort scales was not the same, with the usefulness middle point being *somewhat useful* rather than *neutral*. While this might affect the direct comparison between these two dependent variables, it should not affect our analysis of usefulness and comfort individually, where the scale for all items was consistent, and each rating was considered a distinct class in our analysis.

As mentioned in Section 4.3, the measured privacy concerns of our respondents were high, which was expected since Mechanical Turk workers, on average, have greater privacy concerns than the general US population [102]. While the measured privacy concerns were

not a significant factor in both usefulness and comfort in sharing preferences, the perception towards various categories of contextual information may change with a population with more varied privacy concerns. On the positive side, even with more significant privacy concerns, the respondents were favorable in their perceived comfort associated with sharing different contextual categories indicating the approach's utility. In terms of generalization, our response set has a significant gender bias towards the male population (which is typical for Mechanical Turk-based studies[4]), and our findings may not be representative of the general US population [63]. Finally, our focus with this study was on one-to-one communication. With group messaging, the communication dynamics can be different. While it is possible to direct a message sent in a group conversation to a specific individual, group messages are usually intended for multiple participants. The expectation of fast responses or acknowledgments is generally relaxed. The utility of sharing individual context in these situations would require further investigation.

[4]https://www.cloudresearch.com/resources/blog/the-new-new-demographics-on-mechanical-turk-is-there-still-a-gender-gap/

5.0 Agent Design and Evaluation

5.1 Introduction

In the last two chapters, we explored foundations in the design of a messaging agent with the goal of improving situational awareness in messaging while reducing distractions and preserving user privacy. Through a personalized modeling approach, we can accurately detect unavailability and interpret the model to identify the top features corresponding to that unavailability state, forming the context to be shared to improve situational awareness. Although, as discussed in the last chapter, user preferences related to context sharing can vary depending on factors such as information type, social relations, and message urgency. We tackled this challenge through preference modeling to initialize user preferences based on communication context.

In this chapter[1], we build upon our prior work to explore the design and implementation of a fully automated approach for generating and sending auto-responses as a means to improve situational and unavailability awareness. As discussed earlier, the messaging agent evaluates each new messaging session to predict the availability of the message recipient. If deemed unavailable, an auto-response is generated, which shares the predicted relevant recipient's context, using a user attentiveness model. As we have shown that it is *possible* to design the agent to achieve the goals we have set for it, the next step involves designing and evaluating this agent to understand the perception surrounding its use and its potential to impact user behavior.

In particular, we explore the following research questions:

- **What are important considerations in designing an automated availability management agent to reduce device engagement while maintaining user privacy?**
- **What are users' perceptions and interpretations of the auto-responses generated from the information captured from sensor data?**

[1]The material presented in this chapter was originally published as [96].

- **In what ways does the presence of an availability managing agent affect user behavior?**

To answer these research questions, we developed and evaluated an availability management agent through an empirical two-week-long study with 12 participants who used our messaging agent on their smartphones for the study period. Our findings suggest that participants found the agent useful for communicating unavailability when they could not get to their phones. Participants also reported altering their behavior based on their understanding of the agent's design and function to appropriate it in their desired way. We also learned how inaccuracies in the agent's behavior lead to a sense of loss of interaction control. This occurred when the information shared to message senders by the agent was considered either irrelevant or inappropriate by the message recipients. This also resulted in an increased effort by message recipients to explain the agent's actions to their contacts.

Overall, our work contributes to the field of designing interactive systems by (1) presenting a novel design of a fully automated messaging agent that learns from users messaging behavior to identify and share relevant context related to their unavailability state (Section 5.2); (2) describing ways in which this agent can be useful (Section 5.5.1) and what factors affect its utility for its users (Section 5.5.2); (3) providing insights on how presenting mid-level sensed information (Section 5.2.2) rather than the inferred state could be perceived by users and under what circumstances such messages can be effective or misinterpreted (Section 5.5.3); and (4) empirically evaluating the role of the agent in both positive and negative users' behavior changes (Section 5.5.4).

5.2 Design of Automated Response Agent

The main design goals for the agent are to (1) reduce users' engagement with their devices when they are busy with other tasks; (2) improve situational awareness for users' social contacts; (3) maintain users' privacy through mutual awareness of user context. In this section, we present the design of the agent to achieve these goals.

5.2.1 Fully automated agent design

In order to reduce device engagement, the agent needs to act autonomously without requiring user intervention. An automated agent design would allow users to focus on their ongoing tasks, reducing distractions. Furthermore, as discussed previously, users are inconsistent in updating their status, so the agent should also ensure consistency in sharing users' status information and provide awareness to their social contacts. We designed a fully automated agent by modeling the users' messaging behavior and using this model to detect and share unavailability-related contexts.

5.2.1.1 Detecting and classifying messaging sessions

Similar to Avrahami et al. [10], to define a new messaging session, we used 5 minutes threshold since the last message from the same contact. This helped distinguish new messaging sessions from ongoing conversations and focus only on new sessions in modeling attentiveness rather than all incoming messages. In addition to tracking session initiation, the model also tracked when the user attended a message to generate class labels. We consider a session as *attended* if the user (1) *removes* the associated notification, (2) *opens* messaging application associated with the session, or (3) *accesses* the message on another device[2] (e.g., WhatsApp Web). For a session to be classified as "attended to", one of these events had to happen within *7.2 minutes* from when the message was received. The 7.2 minutes threshold comes from prior literature on attentiveness modeling, representing the average median attend time as the threshold for classifying attentiveness [157, 93].

[2]While we tried to detect web-interfaces of messaging application (e.g., WhatsApp Web), due to the nature of notifications on Android, this detection was not always reliable. This led to some messages being falsely flagged as a new session when the participant used the web interface on another device. Two participants reported being affected by this. One participant reported annoyance and described the agent as an "intruder in the conversation". The other participant reported the event as rare and were not significantly affected by it.

5.2.1.2 Sensors and features used to define context

We used 58 features[3] to create the user model. We derived the feature set from phone sensor data and phone usage data based on previous works on using smartphone data to create user models [157, 156, 93]. We logged two main types of information, (1) *time since an event* features - where events were cases such as change of screen's state (e.g., time since screen unlocked) or communication (e.g., time since phone was last answered); (2) *current status* features such as screen state (locked, unlocked, or covered), connectivity state (e.g., WiFi signal strength), or ringer mode (normal, silent, or vibrate). In addition to these, we also logged additional information such as (1) location, which the users semantically labeled as work, home, or other frequented locations; (2) level of background noise, using frequent processing of background sound through the phone mic [72]; and (3) Calendar information to represents events with which the users might be engaged [109, 95].

5.2.1.3 Modeling

We used personalized modeling of attentiveness [157, 158, 94] to predict when the user cannot attend to their messages. Prior work has shown personalized models (1) more accurately predict users' attentiveness to their mobile messages [94]; (2) are more flexible in terms of the modeling process, optimization, and retraining of the model [93]; and (3) can better support users' privacy by enabling modifications to individual models based on comfort with specific information used to model behavior without having to retrain the general model [93]. We created these individual models using a tree-boosting algorithm called XGBoost [43]. We used *binary logistic* as the objective method. We scaled the positive class weight to the ratio of the positive and negative class instances in our data to deal with potential class imbalance. The rest of the parameters were set to their default (*learning rate = 0.3*, *max depth = 6*, *minimum child weight = 1*) as they usually performed the best when testing on a dataset from another study [156]. Based on Jain et al. [94], we retrained these personal models once a day using cumulative data samples collected in the preceding days.

[3]List of all features used in modeling is available at `https://docs.google.com/spreadsheets/d/1S59ZCWfAmVDA1WucOKXCZu4FMb4QVhvJUyqasAuwC4o/`

5.2.1.4 Detecting and sharing relevant context

For the design of the agent, rather than sharing a status type, we chose to use auto-responses within the same thread of conversation to inform unavailability. This allows the users to keep track of the agent's behavior within the specific context of a communication thread. A sample auto-response is shown in Figure 18. We distinguish each auto-response from regular messages by using text *"AUTO-RESPONSE:"* to signal to message senders that the message came from the agent. The base auto-response is a simple phrase *"[NAME] may not be available to respond"*. We further augment this base auto-response to include specific contexts, which may help explain the unavailability prediction to the auto-response recipient. The motivation for this design comes from how a human assistant may communicate unavailability, for instance, by including information such as "they are in a meeting" or "they have left the office". We illustrate this in Figure 18. The context shared in this case is the noisy background and the calendar information.

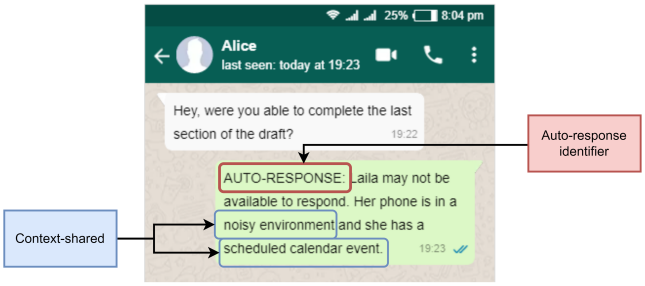


Figure 18: Sample Auto-response with two types of information being shared, device-state (noise level) and user-state (calendar event).

The next design consideration was identifying what information is relevant to share to form these augmented auto-responses. Since the availability models can achieve high levels of accuracy in predicting attentiveness [93, 10], understanding and interpreting the learned model can help identify relevant features associated with unavailability. We used the tree explainer component of SHAP [125] that utilizes Shapley values to produce local interpretations of each messaging session to identify these factors. Figure 19 visualizes an example of one such local interpretation for one of our participants. While these local

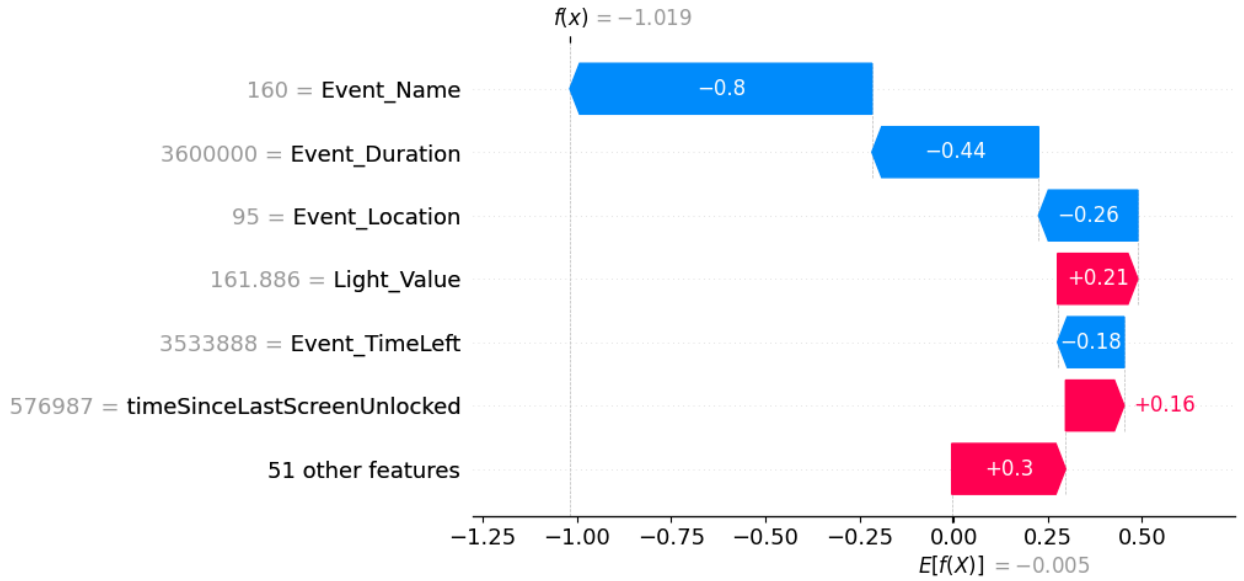


Figure 19: The figure shows a sample local interpretation for Participant P1 generated using SHAP waterfall visualization. Here, the y-axis represents features and their encoded values, while the x-axis bars represent the push of a specific feature toward a particular model output. The bars pointing towards the left or negative axis represent features pushing the model output towards unavailability. In contrast, the bars pointing to the right push the model output toward available prediction. Based on this interpretation, *'Event_Name'*, which signifies a calendar event, has the most significant push towards the unavailability state. At the same time, the high *luminance* and short *time since the phone was last unlocked* are pushing the model output towards the available state.

interpretations may not link to causality, they still help identify patterns for each local prediction. We limited the number of features included in the auto-response to at most three to limit the amount of shared information and reduce the cognitive load in understanding multiple items of information [190].

5.2.2 Privacy Considerations

Ensuring users' privacy is one of the key aspects of our design decisions. By design, the agent sends auto-responses only for (1) new incoming session initiations; (2) when the model predicts unavailability; (3) contacts saved in the address book. This limits access to status information compared to typical online/offline and attentiveness [157, 206] indicators which constantly broadcast application usage status. Further, this enables mutual awareness and transparency since the message recipient is aware that information was shared with their contact through an auto-response in the same thread of conversation [46]. This approach provides high transparency between the social contacts of who has what contextual information about their availability instead of social contacts passively checking a user's status on the application.

As a design decision, we ensured that auto-response messages were neither too low level (e.g., detailed sensor data such as actual decibel noise levels or proximity readings) nor too high level (i.e., the agent is not making any inferences of the *actual* activity of the user). We call this mid-level sensor data. For instance, the agent might report that the user is in a 'dark environment' and 'silent environment' rather than inferring an associated state (e.g., sleeping). Additionally, we aggregated some low-level context values into bundles or categories. For instance, the agent shares the application category instead of sharing the last application used (e.g., productivity and communication). Similarly, instead of sharing precise location coordinates, the agent shares only labeled semantic locations that follow a circular radius along a point of reference (GPS coordinate) the user is willing to share.

Further, the content of the message is not tracked or parsed by the agent. While the agent uses the contact name from the notification to identify new message sessions, it does not use this information to model availability. Finally, identifying new sessions is local to the device, ensuring the privacy of message content and contact information. While sensor data was sent to a remote server for modeling and prediction, as mobile devices become more capable of handling ML/AI tasks using neural co-processors, this processing can also be performed locally on the device, improving privacy even further.

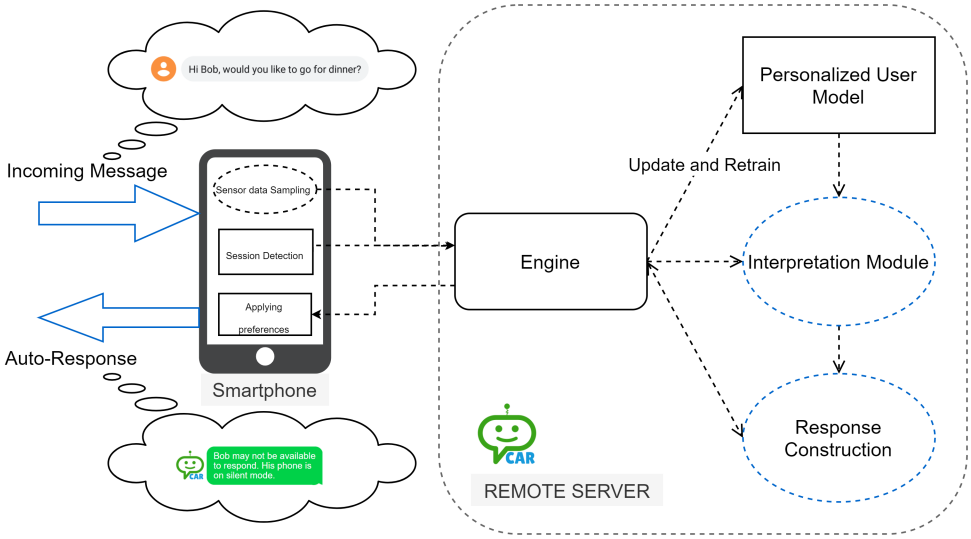


Figure 20: Agent System Design

5.3 Implementation of Auto-Response Agent and Messages

We illustrate the design components of the application in Figure 20. The individual User, Interpretation, and Response Construction modules were deployed on a remote server, while sensor data collection and session identification were performed locally on the user's smartphone. The AWARE Framework was used as a library in our Android application [72] for sampling data from some sensors, while for the rest, we added manual listeners using Android's SensorManager class[4].

5.3.1 Supporting multiple applications

One of our objectives for implementing the agent was to support multiple messaging applications. Approximately 36% of smartphone users have multiple messaging applications on their phones (not including the preinstalled SMS application). People might use these applications either for different purposes (e.g., Slack for work-related discussions and WhatsApp

[4] `https://developer.android.com/guide/topics/sensors/sensors_overview`

for personal conversations [162]) or for interacting with different types of contacts [188]. Thus, the agent's utility might be limited if it cannot support multiple applications. We used Android's Notification Listener Service to intercept all notifications on the phone. We leveraged Android's Quick Reply feature, which allows users to send responses within the notification without launching the messaging application. Using either Notification Actions (introduced in Android API level 19) or Wearable Actions (on older ¡19 API level versions), we were able to use the Quick Reply feature to send auto-responses programmatically. This approach allowed our application to support all applications that supported this feature. Messaging and Social media applications which supported this feature at the time of the study included WhatsApp, Facebook Messenger (and Messenger Lite), Telegram, Signal, Instagram, Google Messages (and other SMS applications), and Slack.

5.3.2 Generating auto-response messages

We generate multiple auto-response patterns in the Response Construction module using the top features returned by the Interpretation module. These are patterns since we generate the actual response on the user's phone to include their name and gender preferences for an auto-response. We did not store any individual information on the application's server side. The response types, descriptions, and examples are listed in Table 9.

For each messaging session where the agent predicts the user as *'unavailable'*, it generates multiple auto-response types as listed in Table 9. The single-feature auto-responses include the top feature, second-best feature, and third-best feature auto-responses totaling seven auto-response types. We construct the *weighted top-features ensemble* auto-responses using a simple heuristic approach: From the list of features and weights returned by the Interpretation module, the Response construction module picks those features that push the model output toward an unavailability state. If the normalized sum of weights for the top 2 features makes up 80% of the overall weight (for the unavailability prediction), then those two features are used in the auto-response; otherwise, the top 3 features are used. Suppose some constituents of an auto-response type are missing (e.g., no device-state features returned by the interpretation module), then the agent skips that auto-response type. After generating

all auto-response types, the agent randomly picks one of them to send to the message sender.

The Response construction module also included a rule base. This rule-base defined rules for phrasing different feature-value pairs and combining multiple phrases to form coherent auto-responses. The rule base also defined a hierarchy of features based on the type of information at different levels, such as high level (e.g., user-state or device-state) or low level (e.g., connectivity or location). This prevented the creation of multi-feature auto-responses, which included highly correlated features such as those shown in Figure 19. For example, only one of *event_name* or *event_location* features will be picked for an auto-response since they both represent a calendar event. Although, we would still consider them independent if they had a unique feature encoding defined, i.e., if *event_name* feature has a different auto-response phrase than *event_location* feature.

5.3.3 Pilot run

The research team carried out a month-long pilot run to (1) determine what controls to add for the primary agent function; (2) fine-tune the phrasing of auto-responses, especially multi-feature ones; and (3) detect and iron out any bugs in the application. The controls added based on the pilot run results included adding a *delay* before sending an auto-response (default: 1 minute). The purpose of the delay value was to give the user a chance to respond when they were available contrary to the model prediction. This value was customizable and could range from 0 (instant auto-response) to 7 minutes (the threshold used for modeling). Another observation from the pilot run was the frequency of auto-responses for some contacts. These contacts with whom a user engages in conversation multiple times a day may get multiple auto-responses throughout the day. Multiple auto-responses in a short period can lead to annoyance and raise privacy concerns due to oversharing contextual data. Instead of limiting the number of auto-responses for a specific contact, we added a contact interval setting to the application, preventing another auto-response from being sent to the same contact for the set amount of time (default: 2 hours).

Regarding the phrasing of auto-responses, we observed that some features, when chained together, resulted in redundant information in an auto-response. For example, if the top 2

Sub-type	Description	Example
No-context/Simple	This response type did not share additional context as part of the auto-response.	Laila may not be available to respond at this time.
Single-feature	These response types use a single feature value from the Interpretation module.	Laila may not be available to respond. Her phone is covered (in a bag or pocket).
User & device ensemble	User-state information represents information about the activities or tasks of the user and their environment. In contrast, device-state information relegates information about the device (e.g., screen-state, ringer mode) [95].	Laila may not be available to respond. Her phone is covered (in a bag or pocket) and she has a scheduled calendar event.
Weighted top-features ensemble	This response type combined responses from multiple top features returned by the Interpretation module to form a single cumulative auto-response.	Laila may not be available to respond. She has not been using her phone for a while and has a scheduled calendar event and her phone is currently locked.
Clustered ensemble	The third type of multi-feature auto-response included features from at least two of three dimensions of locality, time, and task information, comprising top features as returned from the Interpretation module.	Laila may not be available to respond. She is at work and has a scheduled calendar meeting and has not unlocked her phone for a while.

Table 9: Auto-response types generated by the agent

features are the "high number of unattended notifications" and the "long time since screen unlocked", the resultant multi-feature auto-response would be *"Laila may not be available to respond. She has* ***not been checking her notifications*** *and has* ***not been using her phone for a while****"*. In this example, these statements sound redundant when taken

together as they directly relate to one another. To prevent this redundancy, we assigned a category to each auto-response feature based on the type of information it represented, i.e., whether it was related to a *device-state* information (charge level, ringer mode, etc.) or *user-state* information (current location, calendar information, etc.). We then augmented the agent rule base to prevent a multi-feature auto-response from including two features from the same category. The agent then picked the features with the higher weight from the Interpretation module to be included in the auto-response.

5.4 User Study

We evaluated our approach to modeling and implementing the auto-response agent in a two-week user study with 12 participants. We recruited our participants through advertisements on the university news web page, flyers around campus, and social media listings. They were briefed remotely about the description and requirements of the study. In the 15-minute session, we also described our Android application and answered any participants' questions. Following this, we sent the participant a link to install the application and a web-based guide describing the functions and controls of the application. Week one of participant recruitment was dedicated to data collection to build an initial model. The agent generated and sent auto-responses during week two using the participant's personalized attentiveness model trained using week 1 data. The application also sent daily questionnaires during week two. Participants were paid 30 USD for participating and completing the study. It is worth noting that the study took place between September to December 2020, when most organizations and universities were operating remotely due to the COVID-19 pandemic. These circumstances may have impacted our study results, as discussed later in this thesis.

Ethical Considerations. The university's Institutional review board (IRB) approved the study. During the briefing, we disclosed all the data collected by the application and the permission the application needs to function to the participants. This information was also available through the study web page[5] sent out in an email after the briefing. As mentioned

[5] `https://people.cs.pitt.edu/~pranut/messaging_study/index.html`

in section 5.2.2, the agent did not send any text messages or contact information to the remote server.

5.4.1 Application interface

On the first launch of the application, participants had to enter details such as name, gender, and age. Following this, the application presented the consent form describing the purpose of the study. The main screen had buttons to start and stop the background services and an options pane to customize aspects of the application. Upon hitting the 'start' button for the first time, the application prompted the participants to label some locations of interest. They were informed that these location labels would be used for prediction and could also be shared in auto-responses.

The server kept track of when the application was started and stopped and alerted the participant if the application was stopped or crashed for more than 6 hours. Upon stopping the application, all data collection was ceased, and the agent stopped sending auto-responses. At the end of the day, around 9:00 PM, the application generated a notification asking the participant to complete a daily questionnaire asking for their feedback on using the agent. This was also available within the application if the participant accidentally dismissed the notification or wanted to take the questionnaire earlier in the day. Participants could also take the questionnaire multiple times daily, and only unevaluated auto-responses were shown to them. All participants used the option to start the questionnaire from the application, sometimes even multiple times a day.

After two weeks, participants were sent an end-of-study survey within the application, which consisted of general questions about the overall perception of auto-responses. The survey and the daily questionnaire responses guided the semi-structured end-of-study interview, which lasted about 45 minutes on average.

5.4.2 Participants

We reached saturation in terms of new high-level findings after around 12 interviews, and at that point, we stopped recruiting. In total, we recruited 14 participants. One participant

could not run the application on their phone and had to withdraw after one week. After the briefing, another participant withdrew from the study because they were uncomfortable sending auto-responses to their contacts. We discarded any collected data for these two participants and only presented the analysis results of the remaining 12 participants. In terms of the demographics of our participants, six were in the age group 18-24, three in the group 24-34, two in the group 35-44, and one in the group 55-64—seven of our participants identified as female, and five identified as male.

5.4.3 Analysis

The primary researcher remotely conducted the interviews with all participants, which were audio/video recorded. The recorded interviews were transcribed with the built-in transcription of the recording software and further fixed by the primary researcher. We performed inductive thematic analysis on interview transcripts [29] and used Nvivo software for creating and categorizing codes. The primary researcher developed the initial set of codes from half (six) of the interview transcripts, which were then improved upon and categorized into themes and sub-themes during multiple discussion sessions among the research team. Another researcher, not part of this project's research team, coded one of the interview transcripts. We achieved a Kappa value of 0.813 after performing a reliability analysis. Given the high level of agreement, the primary researcher coded the remaining six interview transcripts.

There were 105 initial codes such as "customizing: contact-blocking", "perception of noise value", and "usefulness for family". Upon iterating and refining these nodes, some nodes were split and re-categorized. For instance, we split the "perception of noise value" code into two parts: "perceived utility of noise value" and "interpretation of noise value" categories. From this final set of nodes, we identified 16 first-level categories such as "interpretation", "customization", and "agent accuracy". Through rounds of discussion between the research team, we identified four major themes: "varying preferences related to agent function", "effect of misclassifications", "understanding of the agent and appropriation", and "utility of auto-response information type".

5.5 Results

During the two-week agent deployment, 310 auto-responses were sent ($\mu = 25.333, \sigma = 16.036$) with a minimum of 6 for P2 and a maximum of 61 for P10. The most common auto-response type was the phone-usage ("*Laila has not been using her phone for a while*"), which was sent 86 times (27.74% of all auto-responses). We expected this since phone usage was in the top features for multiple personalized models similar to prior works [94, 93]. The overall accuracy was 70%, the false-positive rate was 0.21, and the false-negative rate was 0.55. We discuss these metrics in more detail in Section 5.6.1.1.

Next, we discuss the major themes emerging from our interview data. The overall response from our participants about the agent and auto-responses was positive, with participants noting less obligation and pressure to respond and to explain their delayed responses. While half of our participants reported less engagement with their phones, which was the agent's goal, it was highly context-dependent. Factors such as the message's urgency, strength and nature of social relationships, and the format and content of auto-response messages all played a role in defining how beneficial the agent was for the users. The type of information that the agent shared, in particular, was an important consideration, as our participants noted that its misinterpretations could be consequential. Additionally, there were indications of behavior change related to device and agent usage arising from the understanding of the agent function and the effort to fix mistakes made by the agent. We expand on each of these in this section.

5.5.1 An auto-response agent can be a useful tool to communicate unavailability

Multiple participants reported various perceived benefits of using the agent and auto-responses, as detailed below.

5.5.1.1 Agent reduced pressure and obligation to respond

Overall, participants found the agent useful in reducing their attention to their phones. We observed an average of *5 minutes* increase in time to attend to new incoming messages

among our participants; i.e., in the first week of study, when there were no auto-responses, they took an average of *18 minutes* to attend to their messages. In contrast, in the second week, when the agent started sending auto-responses, they took an average of *23 minutes*. P4 mentioned that they felt less pressure to check their phone and focused better on their tasks. *P4: "It really put less pressure on me to have to check my phone and my messages all the time to just make sure that people knew I was okay and receiving them, it kind of took that off my plate, and I could be more focused on what I was doing in the moment and then at night or later in the day kind of check back to see if messages required more meaningful response from myself, but oftentimes I could just leave it at that, i.e., auto-response. So I really enjoyed it"*. Indeed, our data confirmed that P4 took significantly more time to attend to incoming messages with the presence of the auto-response, from 8.5 minutes in the first week (no auto-responses) to 19 minutes in the second week (with auto-responses). Although, it could be related to their unique situation described in the next section. Nevertheless, they ascribed lower engagement to the use of the agent.

5.5.1.2 Agent can help stay focused on important tasks

Beyond the general utility of the agent, we learned that there could be certain situations where auto-responses were particularly useful for some participants. For example, four of our participants felt that auto-responses would be helpful while driving, *P7: "A big bad habit I have is that when I'm at stoplights, I'll check my phone. With the auto-responses, I did not do that"*. Similarly, P4 mentioned being on a trip when auto-responses started (during the second week), *P4: "I was in a unique position because when it started up the auto-responses, I was on a long 10-hour road trip. So it's really helpful not to kind of have to worry about responding to people knowing that the app would respond for me, and it did"*.

Another participant (P11) brought up the usefulness of auto-responses while studying, as messages can be distracting during that time, *P11: "They were especially useful when studying because I like to put my phone away and Yeah, I guess, like the biggest drive to pick up that phone is to make sure no one has contacted me"*. Another unique situation brought up by P11 was when they were at a doctor's appointment, *P11: "One time it was useful was*

when I was in a doctor's appointment. First thing in the morning on my birthday, and there were a bunch of people texting me because it was my birthday. And I was like, well, I'm at the doctor for an hour".

5.5.1.3 Agent reduced the need to explain unavailability

Six participants indicated that they had to provide fewer explanations when the agent sent an auto-response. P2 attributed this to an accurate representation of their unavailability state in auto-responses, *P2: "Oh, because it was already laid out for me as to what was happening and why I wasn't available during that time"*. Similarly, P13 described how they felt that agent explanations were sufficient to justify delays, *P13: "Before I used that application and I was away from my phone. I would always get from the other party like where are you, what are you doing, how come you didn't message me back. And then I would have to sit there and just, you know, lengthily explain what I was doing. That's why I felt those auto-responses were helpful".*

5.5.2 An auto-response agent is more useful in some situations

We identified multiple factors influencing how participants felt about the agent and auto-responses in our analysis.

5.5.2.1 Urgent vs Non-urgent messages

Our participants reported variations in the usefulness of auto-responses based on the urgency of the messages. Out of the six participants who brought up urgent/time-sensitive incoming messages, three felt auto-responses were not helpful for urgent messages, while the other three felt they were. The participants who preferred auto-responses in urgent situations gave reasons such as stronger emotions linked to urgent messages and making the sender aware so that they can reach out to someone else, *P9: "When it's someone texting when it's urgent or important, then I'd really want them to get an auto-response, just so they know what's going on. I think that would be really useful because if they know that you're not*

available or something, then they could reach out to someone else". While participants who preferred not sending auto-responses in urgent situations felt that they needed to handle those situations themselves, *P4: "usually those (urgent) messages in the nature of my work on campus are more pointed towards me and are more time-sensitive. I guess that's the only reason"*. This finding falls in line with previous work that surveyed people on their perception of sharing contextual information, confirming that urgency matters and has varying implications on agent usefulness for different people [95].

5.5.2.2 Agent's personality and its content representation

While most participants felt that the tone and framing of the auto-responses were fine, i.e., not too formal or casual, there was a mixed response as to whether they would like auto-responses to sound like them or take on an independent agent personality. Four of our participants felt that auto-responses sounding like them would improve their acceptance for their contacts, *P10: "The person who gets those auto-responses will believe that these responses are from me"*. Further, seven of our participants also wanted to customize some aspect of the auto-responses by adding a personal touch, *P1: "Personalization messages are really big for me. I really like value using my own voice. And so I would definitely want to see that"*. Other participants preferred auto-responses not to sound like they would, to be distinctive from their own responses, and not confuse their contacts. P8 elaborated on this *P8: "I had a friend who used voicemail with, "Hello, are you there?". It sounded like she was actually picking up, and that always drove me nuts because I would try and actually talk to her. I feel like if it (the agent) sounded more like me, it might get more responses unnecessarily"*.

5.5.2.3 Usefulness for different contact groups

Another avenue of varied response was the utility of auto-responses for different contact groups. The qualitative design of the study allowed us to inquire about the perception of the agent for more distinctive contact groups, unlike previous survey-based studies, which were limited to two or three coarse groups [95, 109]. In addition, to close vs. distant groups, our

participants noted the relevance of more fine groups such as higher-authority figures (e.g., boss and advisor), family, friends, coworkers, and even personalized contact types (e.g., their doctors' offices, special-needs contact) as we discuss below.

Four participants mentioned agent usefulness towards an interesting contact group: a higher-authority group such as a boss, advisor, or professor. P9 emphasized the usefulness of auto-responses for their boss, *P9: "More important people like, say like a boss or someone that you always want to be more responsive to, you know, or keep them more in the loop"*.

In terms of close vs. distant contacts, our participants again had mixed perceptions of auto-response utility for these contact types. Some participants did not feel comfortable sending auto-responses to infrequent contacts, *P1: "Basically there are two kinds of people who contact me: people whom I think of as close friends and people who are acquaintances, or maybe who I don't know at all. And so for people whom I don't know at all or not very well or like acquaintances, I definitely don't want auto-responses to go to them because I don't feel the need to tell them anything about me until I've decided whether or not I want to engage"*. In contrast, some participants specifically found auto-responses useful for contacts they did not engage with frequently, such as distant family, *P2: "I have a cousin that's in [redacted] right now. It would have been really useful for her because there were times where I can't always get to her, and I hear sometimes I'm just entirely too busy to respond to her"*.

Similarly, some participants felt that close contacts already knew about their availability and schedules, making the auto-responses less helpful, *P8: "I think people whom I text very frequently, it was less useful. Like if people are already fairly aware of my schedule and (they) can kind of anticipate. It's not necessarily providing any new information"*. Two participants mentioned that while auto-responses were less useful for frequent contacts in general, they were helpful for their families, *P4: "All my family really liked it. I'd say my parents probably benefited the most from it while I was away on vacation. They enjoyed being able to "keep up with me" but know that I was safe and would respond at a later point. And then when we were driving it auto responded to my cousin whose house we were staying at, and she found that helpful as well"*. On the other hand, personal situations also made auto-responses to close family members such as parents not useful for some participants, *P2: "There might be people who just don't want the auto-responses to go to like my mom because she might*

actually need something at that point in time. She's more of a person that I need to get to right away because of health issues".

There were also instances brought up by participants discussing contact types that are more specific to them. For instance, P4 mentioned how auto-responses could be confusing to a special-needs person they interact with through messaging, *P4: "One of the individuals has special needs so, with her, I have to be very direct and blunt with the messages. So I just didn't want to confuse her".* Similarly, P9 mentioned wanting auto-responses to their doctor's office even if they are not on their contacts list to inform them of their unavailability.

5.5.3 Perception and interpretation of information shared by the agent

Our participants evaluated 263 auto-responses in the daily questionnaire. In terms of mean ratings for different categories listed in Table 9, we did not observe any significant difference concerning the usefulness and comfort of these auto-response types. Although, there were implications related to the content of the auto-responses, as we detail below.

5.5.3.1 Is the reason convincing?

Our participants discussed multiple factors as to what constituted a good auto-response. One of them is that the reason shared has to be convincing, *P10: "It's about what the information is, what the reason is, it could be very long, but [if] there is no specific reason, or there is no convincing reason, then I don't think the other person would be very friendly to you".* P8 had a similar opinion and elaborated using an example auto-response that the agent sent to their contact, *P8: "The ambient noise one, I'm like, just playing music in my own house. I don't think [it] makes me less likely to respond".* Similarly, P9 felt that silent environment (noise value) auto-response may not be indicative of unavailability in most cases, *P9: "I feel like there's a lot of cases where you're in a silent environment, but you're still available to respond. You're just like, say, in your room just like reading a book or whatever, like you're not necessarily you know focused on something very important or like, if you're in the library studying, Well, I guess, in that case, then it [would] be different but yeah I think there's just too many cases with that when that wouldn't be a good response".*

Most participants liked the auto-response sharing phone usage. P1 and P9 also noted the reduced privacy risk from sharing phone usage information compared to other user-state information such as location, calendar, and app usage. *P1: "Because it doesn't really tell you what I'm doing, it tells you what I'm not doing, and since what I'm not doing is relevant to their needs, then it makes a little more sense in terms of alignment to me".* P9 expressed a similar sentiment, *P9: "I think that might be one of the best ones just because like you know it's like general, It doesn't give too much information, but it gives enough to infer to the other person that he is not using his phone so he's probably just not available".* Similarly, ringer mode had a positive reception, *P9: "I thought that was really useful because I feel like when my phone is on silent mode, I probably won't want to respond, so I think that's always a good time to send an auto-response".*

5.5.3.2 Privacy implications of sharing app usage information

There was an overwhelmingly negative response to sharing app usage information in an auto-response even though the agent was sharing the category of application (e.g., productivity, communication, and entertainment app) rather than the exact name of the application last used. P1 and P12 mentioned that they were not comfortable sharing app usage due to the potential of sharing highly personal usage information. *P1: "I basically almost never want them to know which apps I'm using on my phone because if I want to look at [inappropriate content]. That's my own thing, not just good, but yeah, I definitely don't love that".* Sharing app usage was not always perceived negatively. P10 pointed out a stark variation in their perception of sharing app usage based on the type of app category shared in the auto-response. One of those auto-responses shared that they were last using an educational app, whereas the other said they were last using an entertainment app, *P10 (for education app): "I think this reason tells them that he's working on some project or something, educational and should not be disturbed."* Whereas, when the agent shares 'entertainment app', *P10 (entertainment app): "They might think like he's ignoring me but he's also using an entertainment app."*

5.5.3.3 Speculative and misinterpreted context

P7 felt that in addition to being convincing, the auto-responses should also not leave room for speculation, *P7: "I like the ones that are just a little bit shorter and clearish. I don't want [the sender] reading too much into it"*. As noted in Section 5.5.2.3, there was also some variation in preferences related to auto-response information concerning different contact groups. In general, while sharing that the user has a calendar event had a positive reception from most participants, P11 felt that sharing that they have a calendar event may lead to speculations and more questions, *P11: "I thought the calendar one was kind of unnecessary. It just kind of makes it begs like oh, what is the event or like begs more questions than a simple like not able to respond"*. P13 mentioned a similar issue with sharing 'not at usual location'. The agent picks this auto-response when the user is in an unlabeled location that affects their availability, *P13: "They want to know what's going on and where am I, that's what they'll be thinking"*.

P5 pointed out the ambiguity of sharing light value, *P5: "Oh, the low light one is kind of not useful. For me nor for them just because it could apply that my phone was just facing down"*. P9 recalled that their contacts found the dark and silent environment auto-responses 'creepy' and raised concern for them, *P9: "A couple of people thought that some of the responses were overly specific or like, you know, kind of creepy. I think they had mentioned the light level one and the silent environment one"*. Similarly, for most participants, noise value auto-responses raised concerns about being misinterpreted due to their potential locality inference. P2 pointed out an example of this. They had an auto-response sent saying that they were in a 'noisy environment', whereas they were in bed, sleeping. While discussing this auto-response, they recalled that it was probably due to their room's loud air conditioning, which their phone's mic might have picked up. So even though the information in the auto-response was technically correct, their contact misconstrued the auto-response itself, *P2: "Didn't think it was appropriate since it sounds like I was at a party and I wasn't, and that one was to my dad. So he's probably like, where is she?"*. P7 and P11 raised similar concerns regarding noise value: *P7: "When I think of a noisy environment. I think it's like crowds, and if it's going to coworkers and my parents, that's not really the*

image I want to put forth". P11: "I feel like it gives the illusion that I'm in like it begs like where are they that's noisy". Prior survey-based studies which evaluated the comfort of sharing noise data did not report on the potential locality inference arising from sharing noise value, making this finding interesting [95, 109].

Another observation that P9 noted was related to potential long-term effects or assumptions based on the information shared in the auto-response. For instance, the participant mentioned that an auto-response sharing 'not responsive at this time of the day' might prevent contacts from initiating the conversation at that time in the future, *P9: "They just assume like yeah this, he doesn't want to be disturbed this time of day and I'll just hold off for later".* Although, P10 felt that this auto-response was particularly useful for them since there were times in the week when they did not respond to messages, *P10: "I think this will clear up the fact that this is not a good time to text because anyway, he won't text you back".*

5.5.4 Behavior change related to the agent and device usage

We were interested in identifying how the agent as a whole and auto-responses influence a change in how our participants were using their devices. Our findings reveal both positive and negative aspects of using the agent to handle communication.

5.5.4.1 Reduction in device engagement when the agent works as expected

As described in Section 5.2, the main goal in the design of the agent was to reduce device engagement by enabling the agent to handle incoming communication. Thus, understanding the effect of the agent and auto-responses on device engagement was one of our focuses for the evaluation. Overall, half of our participants reported *reduced* device engagement with the use of the agent, while the other half reported an *increase*. Most participants initially reported *increased* engagement with their device due to curiosity regarding the tool's novel features. However, perceptions of engagement *decreased* in the latter part of the study, as indicated by the following quotes. *P7: "At first, whenever it first started sending the auto-responses, I checked like, "Oh did it send an auto-response cool!". After that initial checking of messages, I stopped checking them as much because I felt like it could explain if I was*

available or not available". P2 and P11 also felt that auto-responses would help them take a break from their device, *P11: "At times I thought it was actually helpful to not feel the need to be connected to my phone because of that (auto) response. So I thought that was good"*.

5.5.4.2 Mistakes of the agent can increase users' effort and decrease their sense of control

Mistakes by the agent, such as sending an auto-response when it was not needed, resulted in an increased effort by participants to provide explanations to repair a social situation, *P11: "It would send a response, and then two seconds later, I would see it and have to explain that. That (it) was just a false alarm"*.

Reasons for misclassifications. We computed the overall false-positive rate (FPR) and false-negative rate (FNR) based on our logged data of (1) when a user received a message, (2) whether they attended to it within the expected response threshold (7.2 minutes), and (3) whether the agent sent an auto-response. The computed FPR of 0.03 was relatively low due to multiple factors, such as sending auto-responses only for known contacts, auto-response delay setting, contact interval setting, and contact blocking. Without these filters, the FPR would have been 0.21, which is still not very high and is comparable to the results of prior studies [157, 93]. The false-negative rate was 0.55, which was much higher than FPR. However, although the FPR was lower than FNR, our participants' perceptions of these misclassifications differed. All of our participants reported experiencing false positives, whereas only four mentioned experiencing false negatives, with only P2 and P5 reporting a high frequency of missed opportunities to send auto-responses. This indicates that most participants were more sensitive to the agent responding when not needed than not responding when it should.

How unavailability is defined can contribute to participants' perception of false positive incidences, i.e., how long of a delay in responding is acceptable to send an auto-response? As discussed in Section 5.2.1.1, we used a threshold of 7.2 minutes for labeling attentiveness based on prior works, which used the average median time in their respective datasets for evaluation [157, 93]. Multiple participants felt that not attending to a message within 7

minutes does not warrant an auto-response, *P8: "I'd say probably somewhere between 20 and 30 (minutes) is fast enough to not warrant an auto-response"*. Another reason could be the particular circumstances of our study, which took place during the work-from-home and stay-at-home period due to the COVID-19 pandemic. Multiple participants reported unusually greater attention to their devices due to classes and work taking place remotely from home, making it harder for the agent to detect instances of unavailability (resulting in greater FNR as well), *P8: "I think just because of the way my work in school is, I'm online most of the time, or, you know, I'm within ready access of my phone most of the time when I'm awake. And if I'm not, it's like I'm on a certain kind of call or running or driving. I don't think there were a lot of opportunities for it to send one where it didn't"*.

Sharing irrelevant or unwanted contexts also resulted in participants trying to explain that context while feeling more obligated to respond earlier than they would have. P13 described a situation where the agent sent an auto-response that they were listening to music which caused them to respond immediately, explaining themselves, *P13: "I was using an app, and I was playing music, and I saw an auto-response went out. I immediately got off the app and went into Messenger. And I told my mom. I'm like, Hey, I'm available to talk to you. I'm just, you know, listening to music"*. Similarly, P14 recalled when the agent sent an auto-response saying they were last using a communications app, making them respond quicker than they would have since the auto-response indicated they had been messaging recently. As mentioned in Section 5.2.1.4, our approach utilizes correlation with the availability state rather than causation. This can lead to sharing irrelevant context that the user or their contacts may not link to unavailability, leading to increased effort and loss of control over the interaction.

5.5.4.3 Uncertainty and lack of understanding of agent function negatively affects its usage

As mentioned earlier, about half the participants reported increased device engagement due to using the agent. Unfamiliarity with how the agent functioned was a significant reason for this behavior change. Some participants reported checking their phones more often to

prevent an auto-response from going out. For instance, P11 suspected that not using their phone for a certain time triggered an auto-response, *P11: "Sometimes I would check it even more frequently because I didn't want that auto-response to go through"*. Similarly, P8 described checking their phone more often in anticipation of an important message they did not want the agent to respond to. This was another example of increased use due to the belief that not using the phone will trigger an auto-response, *P8: "I was checking more constantly because I was worried that it would send him (landlord) something, and I'd have to explain it. We don't talk a lot. So I think it would be kind of weird"*.

On the other hand, P2's experience with the agent auto-responses was quite the opposite. P2 reported issues where they expected the agent to auto-respond, but it did not. They explained that they would often go into messaging app to check if the agent sent an auto-response upon getting a message. If not, they would respond themselves, reducing the agent's utility and increasing their device engagement, *P2: "My engagement would have probably went down. I don't want to engage with my phone as much. I was trying to practice that a little bit in terms by leaving my phone away from me for a bit, but then I will pick it back up If it was like five minutes and I didn't see anything (auto-response)"*. This behavior projects the gap between understanding the agent's function and expectations. Since the agent learns from messaging behavior of its users, opening a message within 5 minutes of arriving, P2 was inadvertently *attending* to it. This would cause the agent to prevent sending an auto-response (if the delay setting is greater than 5 minutes) while also learning that the user is available in that context. Some participants also tried to align the agent's behavior based on their understanding of the agent, e.g., by turning off the app, *P8: "I knew, I was going somewhere (and) the algorithm would notice that you know, doing something different, (or) at a different location and I didn't want it to notice that. I didn't want it to send auto-responses (at that location)"*.

The presence of the messaging agent also affected some participants' contacts. For instance, P12 reported that their contact sent multiple messages upon getting an auto-response, *P12: "A lot of times they said that when they messaged me like they weren't sure if I got it or not. They messaged me almost three times the same message. I don't think they were 100% if [I] got the message or if it went through. I think they felt like sometimes it was*

blocking them or something". P5 had to stop their app because some of their contacts were trying to trigger agent response out of curiosity, *P5: "I kind of had to stop it (app). Just because I know some people were starting to mess with the app and, like you know, purposely responding to stuff just to see what would happen. And like, I think it can get a bit too abusive with it"*.

5.6 Discussion and Summary

Intelligent Personal Assistants or IPAs are designed to assist users in their tasks by utilizing contextual information through sensors [123, 56]. We are seeing IPAs take on more proactive tasks without requiring initiation by their users [209]. Our work on the availability management agent advances our understanding of facilitating awareness in mobile messaging through a virtual assistant. We present design implications from our findings, followed by the limitations of this study.

5.6.1 Design Implications

5.6.1.1 Need for more cooperative human-agent interaction

As discussed in Section 5.5.4.2, mistakes made by the agent decreased users' sense of control and increased the effort to explain agent actions to their social contacts. These mistakes or misclassifications, as reported by the participants, took three forms: (1) false positives - situations where the user was available to respond, but an auto-response was still sent; (2) false negatives - situations where an auto-response was expected to be sent but was not; and (3) auto-responses shared irrelevant information to explain users' unavailability. While the model's intelligence can continually be improved as more data becomes available, as we explain below, there are also cases specific to unforeseen circumstances, such as the response from the user's contacts. Here, we argue that intelligent agents must be designed more as human partners, and their design should support user feedback.

Learning from the user. In addition to retraining the model daily, as mentioned in Sec-

tion 5.3.3, to reduce potential false positives, we introduced a delay setting in the CAR application to allow users to set up a delay before the agent sends an auto-response; however, we restricted the maximum delay setting to be 7 minutes to conform to the threshold used for labeling (7.2 minutes). Nine participants adjusted this delay setting, with 6 participants changing the delay at least twice. The general reason given by the participants for increasing the delay was to give them a greater chance to respond if they were to become available. Feedback from contacts also affected how participants adjusted the delay setting. For instance, P7 reduced the delay as it caused their contact to misread the agent response as theirs, *P7: "[My boyfriend] was just like, you know, I really don't like it when there's such a delay between the auto-responses. It makes me expect that you actually responded to my message".*

These interactions with the agent can provide helpful context about user preferences to the agent [65]. The agent can link each user interaction within its setting as a learning opportunity about the user. The agent uses past messaging behavior to create an availability model for their users. We learned that in some cases, users might be interested in pushing an auto-response even when the agent predicts them to be available correctly. Providing adequate controls to users in such cases while allowing the agent to learn the user-specific context for future incidences can improve human-agent interaction. This was also highly reflected in users' feedback about what context the agent should share. Most participants wanted to customize or add a personal touch to auto-responses. Allowing users to link or change auto-response presentations to specific contexts can help improve the agent's perception while simultaneously reducing ambiguity associated with specific sensor values. For instance, a user can be allowed to change the term 'noisy' to another term more applicable to their context, such as 'busy', as demonstrated in the case of P8, who described wanting to change the noise value phrasing, *P8: "It wasn't really telling them anything helpful about where I might be, um, maybe if the person knew like noisy environment equals busy. Maybe if I were like a construction worker or something, but I'm definitely not. It was kind of unhelpful information in that context".*

5.6.1.2 Intelligent Personal Assistants can teach their users about AI by being transparent

People typically appropriate technology to suit their needs [162]. In this study, as discussed in Section 5.5.4.3, our participants tried to use their exposure to the agent to understand how the agent was functioning and, in some cases, reverse engineer the behavior of the agent by altering their own behavior (e.g., turning off the agent when moving to a new location) or trial and error to decode the agents' behavior some of which resulted in increased device engagement contrary to the purpose of the agent. This demonstrates a significant opportunity to design intelligent personal assistants as a medium to teach users about intelligent algorithms. Previous literature on agent design has emphasized the importance of making AI actions and machine learning predictions explainable and transparent to users. It helps with improved system understanding [86] and can also help build trust in the system use [1]. Thus, for the design of the communication agent, it becomes essential to make the learned model open to the user and provide clarity towards agent actions and learning opportunities about the agent's behavior. Users can interact with the agent to ask questions about the agent's behavior in different contexts. Improved agent understanding and the addition of proper controls, such as modifying or removing any learned context (e.g., location) from the model, can help users make more meaningful appropriations of the agent and gain a higher level of awareness about intelligent agents.

5.6.2 Limitations

Our study was affected by the COVID-19 pandemic. As discussed earlier, our participants reported having higher than usual phone access due to working from home during the pandemic. Increased access and quick attendance to notifications limited the agents' opportunities to auto-respond and could have affected the results of our study. Further, the agent operated with the regular availability indicators in messaging applications. We did not ask our participants to disable these indicators since we wanted to support multiple messaging applications for this study. It would have required effort on the part of the participants to find and disable these indicators, which might not even be possible for all available applica-

tions. This could have affected our results as well. However, we did not receive any feedback from our participants on how auto-responses worked in combination with these indicators, which might be helpful to explore in the future.

As discussed in Section 5.5.2.1 some participants preferred auto-responses in urgent situations, while others preferred to handle urgent situations themselves. Further, there might be other situations where the agent's action could be undesired. For instance, P8 recalled a situation, *P8: "it's also sort of unpredictable, what kinds of responses warrant an auto-response versus not. But, um, I was asking someone for a letter of reference, who I primarily contact through text, and that person responded to me saying, I'm going to need a little extra time one of my parents died. And that's definitely the kind of message where I would want to respond personally and have some time to think about it. And so if the app is doing anything with message content, I would say like maybe scan for the message being kind of serious"*. While we developed a detection mechanism for the agent which prevented sampling of meta-messages such as reactions in Signal[6] app, we did not parse message content to detect emoticons or end-of-conversation behaviors [77] or messages such as *"goodbye", and "talk to you later"* [118]. This understanding of conversation will help prevent agent actions in these situations, potentially improving agent utility. However, parsing text messages can have privacy implications. Further research is needed to understand the balance between getting more context from conversations and user expectations of their privacy.

Finally, our participants used the agent for two weeks, within which the auto-responses were sent only for the second week. As noted in Section 5.5.4.1, users initially reported increased engagement due to curiosity about how the agent functioned. However, we might see more habituation and considerable decreases in device utilization once users are comfortable with the agent. A more extended study can provide quantitative evidence regarding how beneficial the agent can be for its users. In addition, it would be interesting to see how users and their contacts start sense-making of the information the agent shares in the long term as these might further raise privacy concerns [112]. A long-term study will also help us understand whether over-trust and over-reliance could be a potential issue for this agent type [55, 99]. Through personalization, as the agent gets better at its task, users may

[6] `https://signal.org/en/`

rely on it even more, potentially impacting how they utilize the agent and engage with their contacts.

6.0 Co-designing Explanations for the Messaging Agent

6.1 Introduction

As discussed in the previous chapter, our evaluation of the agent through an 'in-the-wild' study showed that users perceived the agent as helpful in signaling unavailability and reducing distractions associated with mobile messaging, which were the main goals we set for the design of this messaging agent. Although, as discussed in Sections 5.5.4.2 and 5.5.4.3, *mistakes* by the agent and the *lack of understanding* of how the agent functions negatively impacted the agent's use. Participants reported increased engagement with their devices to understand agent behavior and prevent agent actions. Further, multiple participants also reported putting in additional effort to explain inappropriate agent actions (such as sharing an undesired context). Lack of understanding of how the agent worked also affected how participants engaged with it, e.g., turning off the agent when moving to a new location for privacy reasons. While the agent did provide justifications for sending auto-responses, participants sometimes questioned the relation of that context to their unavailability. Further, the justification was missing when the agent did not take action, i.e., it did not send an auto-response. These observations indicate the need to augment the agent design to be more transparent and intelligible, allowing for better appropriation and intended use [162][1].

Indeed, the traditional black box design of AI systems can make it difficult for people to understand how they work [71]. 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 [171]—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 [208] and exerting unnecessary effort to use the system functions [96]. It may even harm the users through unexpected disclosure of sensitive information [36]. Explanations have been instrumental in improving user understanding of automated agents' actions and building trust in automated systems [146, 1, 190, 127].

[1]The material presented in this chapter was originally published as [97].

Particularly in recommender systems, various explanation interfaces such as textual [110], visual [191], and interactive [114, 176] have been explored. These explanations usually aim to improve transparency, effectiveness, persuasiveness, scrutiny, trust, satisfaction, and efficiency of recommendations [189].

In this chapter, we explore how we can design explanations for actions taken by the proactive auto-response messaging agent we study in this thesis. 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 [96]. 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 [58] of how it works [2, 171]. 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 [96, 213]. 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 [96, 162].

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 [162] 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.

6.2 Methods

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

6.2.1 Study Design

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

6.2.1.1 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.

6.2.1.2 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 [94]. 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 unclear things 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

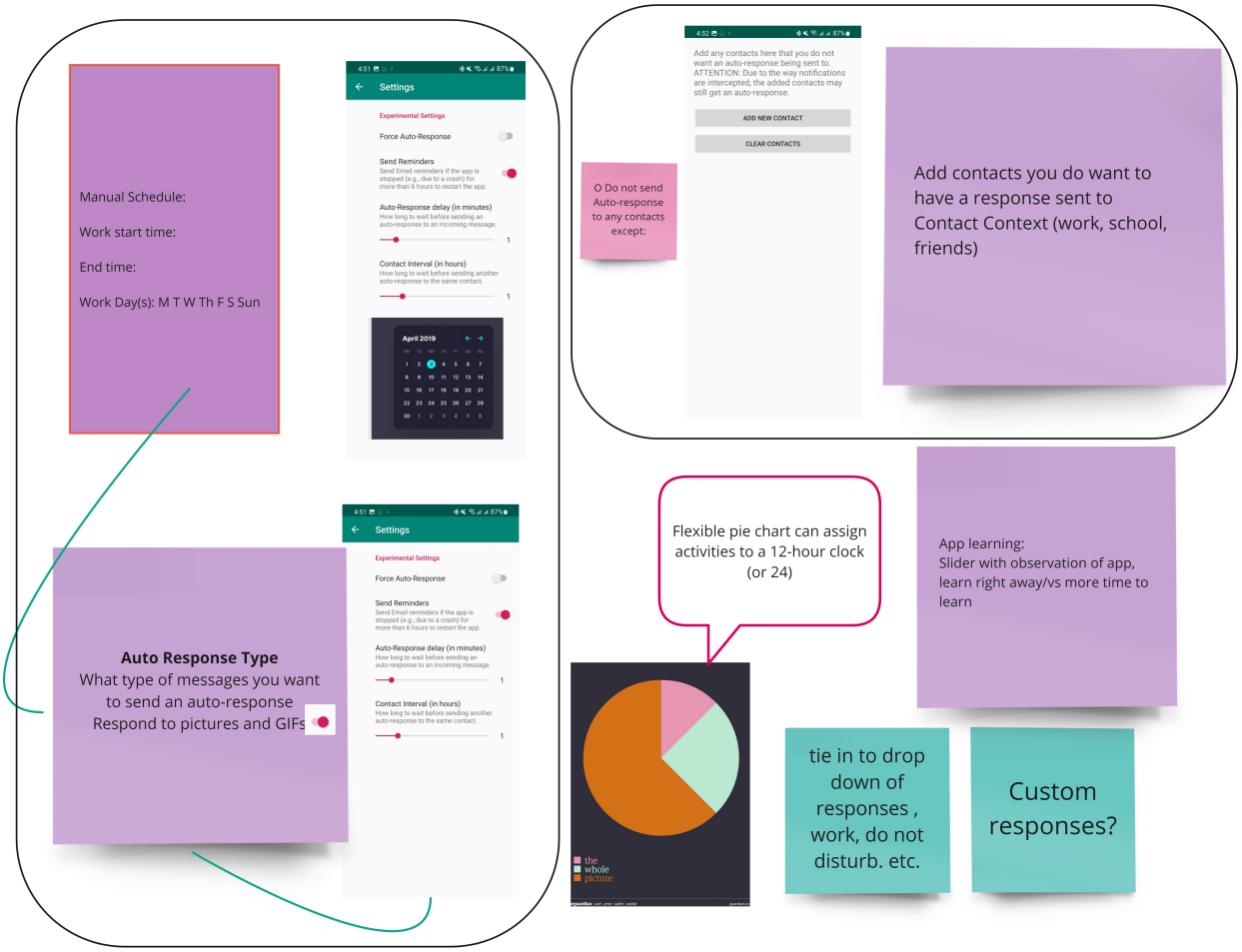


Figure 21: The design space for Dyad F with design sketches and sticky notes at the end of the session.

was not required, and participants could also email their notes before the scheduled design session.

6.2.1.3 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

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 [179, 203, 113]. 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 [113]. 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[2] 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 21 shows the completed design area for Dyad F for Task 1.

While some participatory design studies have researchers or external entities designing while the participant discusses their requirements [169], 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

[2] https://miro.com/

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.

6.2.1.4 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.

6.2.2 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 [52]. We used Nvivo to structure and categorize all codes[3]. 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

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

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 6.3).

6.2.3 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 difficulties and could not complete the study, and their data was not included in our analysis.

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*.

6.2.4 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.

6.3 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*: ($\mu = 4.79, \sigma = 0.41$), *I feel that my partner's opinion mattered and was incorporated into the design*: ($\mu = 4.50, \sigma = 0.63$), *I feel that the collaboration with my partner improved my designs*: ($\mu = 4.64, \sigma = 0.72$).

Our thematic analysis uncovered four key themes: (1) Exposure and observations of agent actions trigger reasoning about factors in its decisions (Section 6.3.1); (2) Curiosity about unexpected agent behavior motivated the desire to update initial mental models (Section 6.3.2); (3) Observations of agent actions and dyad interactions can support learning about the agent (Section 6.3.3); and (4) Users can strengthen agents' predictive models with rule-based heuristics (Section 6.3.4). We now explore each of these themes and their interrelations (Section 6.3.5).

6.3.1 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.

6.3.1.1 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 aware-

ness 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*".

6.3.1.2 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*".

6.3.2 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.

6.3.2.1 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*".

6.3.2.2 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*".

6.3.2.3 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*".

6.3.3 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.

6.3.3.1 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...*".

6.3.3.2 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.*"

6.3.4 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 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 22, "*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

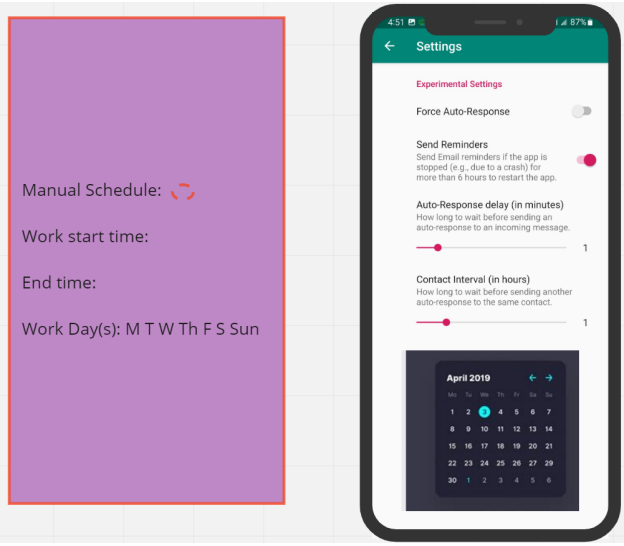


Figure 22: The design sketch by Dyad E to manually enter the user schedule to assist the agent in its predictions. Selecting a date on the calendar opens a new screen where the user can set their schedule.

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*".

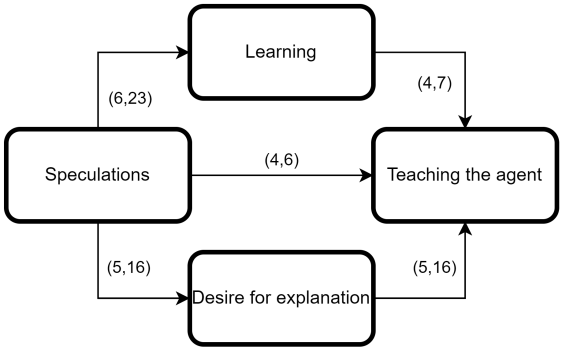


Figure 23: Four main concepts in the findings (Learning, Speculations, Desire for Explanations, and Teaching the agent) and how they are connected in the analysis. The first number represents unique dyads that transitioned from the source concept to the target concept. The second number represents the total number of times that transition happened in any discussion.

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

Figure 23 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 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 6.3.1), and teaching to the agent (Section 6.3.4) was always the end point of the discussion (end node). Below, we explain each transition in more detail.

6.3.5.1 From speculation to learning, desire for explanation, and teaching

As seen in Figure 23, 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 6.3.1, 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 6.3.5.1, 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*".

6.3.5.2 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 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*

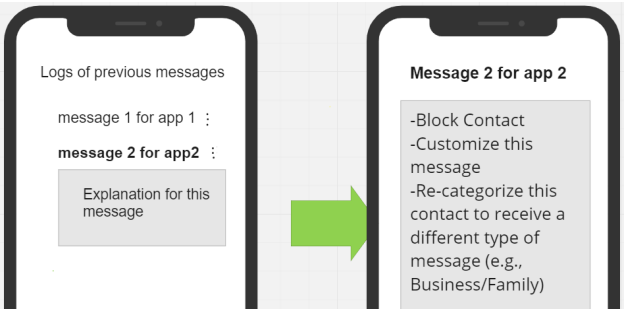


Figure 24: Design suggestion for actionable explanations by Dyad A.

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".

6.3.5.3 From Desiring explanation to Teaching the agent (5 dyads, 16 references)

While participants discussed wanting explanations for unexpected agent behavior (Section 6.3.2), 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 24 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.

6.4 Discussion and Summary

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.

6.4.1 Adaptable proactive agent design

We designed the messaging agent to be fully autonomous to reduce distractions associated with mobile messaging [96]. Our participants described multiple situations where they desired a specific behavior from the agent based on particular contexts (Section 6.3.4). 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 [173, 111]), 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 6.3.5.3). 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 6.3.5.1).

6.4.2 Understanding and augmenting social norms in agent-mediated interactions

In human-human conversations, people follow social norms such as being cooperative and polite [187]. 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) [76]. 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 6.3.4). 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 6.3.5.1). 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 6.3.3.2). **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 6.3.2.1).

This disruption in social norms of communication by the agent can increase the users' effort to justify agent actions to their contacts [96]. During the design sessions, participants described various ways such as providing feedback to the agent (Section 6.3.5.3) 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 6.3.4) 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 6.3.3.2 and 6.3.5.3). 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 6.3.5.3), where participants indicated a mismatch between their expectation of what constitutes *'not being receptive to communication'* compared to what the agent was coded with.

6.4.3 Leveraging user expertise towards desired behavior

6.4.3.1 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 6.3.1.2). 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 6.3.5.1). 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 [86] 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.

6.4.3.2 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 [93]. 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 [116, 138]. Finally, even if the agent's explanation is perfectly accurate, the user may not find it appropriate to share with specific contacts (Section 6.3.3.2). 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 6.3.5.3), customizing the content of the auto-response, and automatically triggering auto-responses based on specific contexts (Section 6.3.4). 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.

6.4.3.3 Community-based knowledge exchange

In Section 6.3.3.2, 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.

6.4.3.4 Engage with User curiosity

In terms of user-agent interaction, prior work has emphasized that explanations must be interactive to be more engaging [128, 137]. Allowing users to ask follow-up questions is one way of driving more natural interactions between humans and AI [122]. We also observed in Section 6.3.2 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 [6]. 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 [75]. Future work is needed to identify what factors could effectively enable human-like interactions between humans and AI to allow learning about the agent.

6.4.4 Limitations

There is a tendency to passively accept others' opinions in group-based discussions [202]. 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 [177]. 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 [113]. 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.

7.0 Discussion and Reflection

In this chapter, we discuss the implications of the results of the two user studies discussed in Chapter 5 and 6 towards improving user attention modeling. We also discuss how we can leverage the foundations of effective human communication into agent design to enhance user-agent interaction.

7.1 Improving user modeling

In Chapters 5 and 6, as part of our methodology, we collected participants' messaging activity and subsequent context related to that activity to build their attentiveness models. For the user study described in Chapter 5, we collected data from 12 participants over two weeks totaling *5782* messaging session instances. For the co-design study described in Chapter 6, we collected data from 17 participants for around two weeks, totaling *3473* messaging session instances. Table 10 summarizes the data collected from these two studies.

As seen in Table 10, the personalized modeling approach used in the two user studies had a comparatively lower performance than that used on the Pielot dataset we studied in Chapter 3. For the agent evaluation study (Chapter 5), even though the false-positive rate was not very high (0.21), as discussed in Section 5.5.4.2, multiple participants were critical of the agent's action when it was not needed. Similarly, in both user studies, false negatives also affected the perception of the agent's utility. For instance, in Section 6.3.2.2, multiple participants questioned agent utility and desired explanations when the agent failed to take action, i.e., send an auto-response when they were unavailable. Thus, while achieving perfect accuracy may not be possible, there is a need to improve the agent's modeling performance to improve its utility for its users and reduce the mistakes (misclassification) that can cause additional effort from message senders to explain these mistakes (Section 5.5.4.2).

Why does the model underperform compared to the Pielot Dataset discussed in Chapter 3? Even though the evaluation results for the modeling process discussed

	Modeling Study	Evaluation Study	Co-design Study
Number of participants	274	12	17
Data collection period	∼3 weeks	∼2 weeks	∼2 weeks
Total Messaging instances	1,375,234	5,782	3,473
Unavailability instances	572,736	1,558	1,207
Availability instances	802,498	4,064	2,226
Imbalance Ratio	0.714	0.383	0.542
MCC score	0.603	0.154	0.146
F-measure	0.743	0.403	0.475
Accuracy	0.843	0.631	0.601
Random Oversampling			
MCC score		0.336	0.330
F-measure		0.693	0.651
Accuracy		0.657	0.654
SMOTE-NC			
MCC score		0.281	0.271
F-measure		0.662	0.631
Accuracy		0.632	0.627

Table 10: Comparison between the evaluation of personalized modeling applied to the Pielot dataset in the Modeling Study (Chapter 3) and the data collection from the two user studies (Evaluation Study (Chapter 5) and Co-design Study (Chapter 6).

in Chapter 3, and the two user studies are not directly comparable due to the difference in the number of participants, length of data collection, and the number of features used in modeling, it might still be helpful to understand the cause of lower model performance.

Further, we used sequential stratified k-fold cross-validation to evaluate the user study results compared to grouped k-fold cross-validation for the Pielot dataset since we could identify messaging sessions in our data collection, and these did not need to be grouped, unlike the Pielot dataset analysis. Sequential cross-validation generally underestimates the model performance for intrinsically ordered data [62, 164]. As discussed in Section 5.5.4.2, one of the potential reasons for reduced model performance could be the situational circumstances at the time of User Study 1 (agent evaluation). Due to the COVID-19 pandemic, multiple participants reported working from home and having more than usual access to their devices. This resulted in fewer unavailability instances, making it harder for the agent to learn and detect unavailability. This can also be observed in Table 10, as the number of unavailability instances was substantially lower than the number of availability instances for the first user study, resulting in an imbalance ratio of 0.714. The co-design study, on the other hand, took place from April 2022 to August 2022 and has a lower imbalance ratio than the evaluation study but is still much higher than the study on the Pielot dataset. Another potential reason for the mismatch in performance could be geographical and temporal differences. The Pielot dataset was collected in Europe in 2015-2016, whereas our data collection took place in North America, particularly in Pittsburgh, PA, between 2020-2022. Further investigation and evidence are required to ascertain whether these could have affected the modeling results.

To understand whether data imbalance indeed was the cause of lower modeling performance, we performed regression analysis on each participant's modeling evaluation. *Imbalance Ratio* was set as the independent variable, and the *number of sessions*, *median attend time*, participant *age*, and *gender* were set as covariates. F-measure (unavailability class) was set as the dependent variable. The regression analysis showed that Imbalance Ratio was a significant factor in model performance ($b = .391, p = 0.01$). The higher the imbalance ratio, the better the F-measure for detecting the unavailability class.

We also compared model performance using MCC (Matthew's correlation coefficient) classification metric. A high prediction score using MCC is only achieved when the model performs well in all four confusion matrix categories (true positives, false negatives, true negatives, and false positives) proportionally to the size of positive and negative elements in the dataset [25]. This makes using MCC preferable for imbalanced datasets, particularly for

binary classification over the F1 score, which does not account for True Negatives [25, 45]. The MCC score ranges from -1.0 to +1.0. When the score is close to +1.0, more predictions match the labels. If the score is closer to -1.0, more predictions disagree with labels. Finally, if the score is closer to 0, more predictions and labels do not have strong correlations, i.e., the predictions seem random. As shown in Table 10, the personalized modeling approach on the Pielot dataset yielded an MCC score of 0.603, substantially higher than the MCC score in User Studies 1 and 2.

To tackle the issue of data imbalance, generally, two approaches are utilized, (1) data undersampling and (2) oversampling. Undersampling in binary classification involves keeping all the instances associated with the minority class but decreasing the number of instances related to the majority class to balance the dataset. On the other hand, oversampling techniques are used to increase the number of minority class instances (for example, by duplicating them) while keeping the majority class instances unchanged. Since, with a personalized modeling approach, we already lack initial data for users, removing data points may not be feasible. Thus, we focus on oversampling to tackle the data imbalance issue in this thesis. We discuss evaluation results from random oversampling, where minority class instances are randomly picked and duplicated. Although, this duplication does not provide any new information to the model and, as discussed below, can even overestimate the model's performance. Another oversampling technique we explore is SMOTE [41]. It works by first selecting a minority class instance at random and finding its k nearest minority class neighbors. One of these neighbors is randomly chosen, and a new synthetic instance is generated as a convex combination of these two instances. While multiple implementations of SMOTE have been proposed, since our feature set includes a mix of categorical and numerical data, we used the SMOTE-NC, which works with this mix of data [91].

The result of applying the two oversampling approaches is shown in Table 10. With both oversampling approaches, we observed a substantial improvement in the F-measure (unavailability class) and the MCC scores for the User Study 1 and 2 datasets. While Random Oversampling seemingly performed better than the SMOTE-NC approach of generating synthetic samples, it is important to note that since samples are being duplicated and potentially being split across training and testing fold during evaluation, the performance of

the Random Oversampling approach would be overestimated to a higher degree compared to the performance of SMOTE approach. Thus, to conclude, we have shown that oversampling approaches can effectively tackle the issues of data imbalance in modeling messaging attentiveness using sensor data. We particularly recommend using SMOTE-NC for oversampling based on our evaluation results.

7.2 Improving the *quality* of agent responses

In the last section, we described the use of oversampling to improve modeling performance. Even when the agent's prediction is accurate regarding the state of the message recipient, the agent's outcome may still not match the user's desired outcome. For instance, a common occurrence with the use of the agent was the context shared with message senders was perceived as not useful or even inappropriate in some situations (Section 5.5.4.2). This was reported in both user studies 1 (Chapter 5) and 2 (Chapter 6).

In this section, we explore ways to improve the agent explanations based on the results of the co-design study discussed in Chapter 6. We present the issues reported in the two deployment user studies with how the agent communicates unavailability. We discuss how we can accommodate social constructs and lessons from human-human conversations that can be adapted into the agent's design to improve user interaction [187]. Rather than trying to emulate human-human conversations, we strive to learn from how humans communicate to minimize conversation breakdowns that the agent could cause as an intermediary.

It has previously been reported that when interacting with machines, users often have similar expectations, norms, and behaviors to that of interacting with humans [161, 187, 31]. Social norms followed in human-human conversations still apply when interacting with these systems, even when users know they are not interacting with actual humans [160]. We observed evidence of this phenomenon in our research as well. As discussed in Section 5.5.4.2, our participants reported explaining agent actions on multiple occasions to their contacts, for instance, when they felt that the agent response could be misinterpreted to avoid any negative consequences on their relationships. Next, we discuss situations where conversation

breakdowns could occur and potential ways to remedy these in agent design.

7.2.1 Achieving common ground between the user and the agent

When humans interact, they often have a shared understanding of each other, the situation, and the context of their communication [15, 51]. We try to establish this understanding as part of or before engaging in communication. Our conversations are generally tailored based on this shared understanding of the situation [187]. We need to understand each other's knowledge, abilities, beliefs, or emotions to gain this common ground. The current agent design inherently lacks the ability to achieve this common ground with its users.

Problem with asking about user preferences. One way the agent can try to achieve common ground with their users is to ask for their preferences directly. For instance, ask what responses are acceptable to share and which are not. We discussed context-sharing preferences in detail in Chapter 4. In the survey-based user study, we asked participants about their perceptions of the utility and comfort of sharing multiple categories of contextual information through auto-responses. We observed that user preferences varied, and we can, with some confidence, cluster users based on their preferences. However, as reported in Chapter 5, there were multiple instances where users reported discomfort sharing some environmental information such as noise or light value. In our study and prior work [109], when asked explicitly (e.g., through a survey), there were no indications of this shared context being misinterpreted. The negative misinterpretations of noise values were also related to the time this context was shared. For instance, sharing noise value at night was linked to 'partying' (Section 5.5.3.3). Another example where the perception of the shared context varied when asked explicitly was sharing the *'app status'*, which indicates what app was last running on the message recipient's phone. Sharing engagement with entertainment apps such as gaming or media playback was perceived negatively. In contrast, productivity or educational apps (office suite) were acceptable to be shared. When entertainment app use was given as a reason for unavailability, participants were concerned about being perceived as slacking off, particularly during the daytime, as some participants reported communicating via messaging for work (Section 5.5.3.2).

Hidden dimensions in context-sharing. The above observations point to hidden dimensions we did not consider in our survey study discussed in Chapter 4. We looked at context-sharing through the agent from the lens of only the communication context (urgency and social relation). The additional context or dimensions, such as time, location, and value of shared context itself, seems to influence users' perception of the agent's shared context. Although, it is hard to consider all such dimensions concerning the agent-shared context in a survey-based study. One solution could be to understand socially acceptable behavior for the agent in different situations and build that into its design. We have demonstrated the effectiveness of a co-design study in trying to understand social norms related to the use of the messaging agent. Further co-design sessions could be conducted with the agent users to understand their perspectives about the agent's behavior in several situations they encountered while using the agent. This could then be incorporated into the design of the agent.

7.2.2 Learning from mistakes

As mentioned before, errors or mistakes made by the agent could be due to inaccurate modeling and differences in user preferences, situations, and knowledge that the agent may not have or understand. When humans make a mistake while communicating, we tend to apologize and correct our errors [170]. Prior research has reported that humans prefer similar behavior from machines [211], i.e., incorporating apologies and explanations for errors followed by remedies to prevent future mistakes. In the co-design study, we also reported that our participants designed actionable explanations to understand the agent outcome for unexpected behavior (or mistakes) and remedy the situation by giving feedback to the agent (Section 6.3.5.3).

But how would the agent identify when it has made a mistake? One approach would be to rely on the user to report undesired agent behavior. However, this would require substantial effort on the user's part to report errors and teach the agent the correct behavior. This is particularly problematic in the early use of the agent when it is still learning about user preferences and is more likely to make mistakes. Thus, it becomes crucial to detect an

error and try to remedy the situation automatically. Below we discuss potential ways the agent can identify an error in its behavior and different actions it can take:

- Agent inaction, i.e., an auto-response was desired but was not sent. Participants in our two studies reported multiple instances where they expected the agent to respond on their behalf, but it missed these opportunities. If the agent predicted the user to be available when they were not, the agent would be able to identify this situation reliably as it would be able to check whether the user attended to the message within some threshold. The agent can remedy its mistake by sending an auto-response right after this threshold passes and updating the user model to learn from this data.
- Agent action, i.e., an auto-response was not desired but was sent. This situation can occur if the agent incorrectly predicted the user to be unavailable. The agent can recognize this mistake if the user attends to the message within the threshold after the agent sends an auto-response. The agent can again remedy this situation by updating the user model.
- Sharing undesired context. Another situation would be when the agent detects the unavailability correctly, but the context it shares to explain the user's unavailability might be considered inappropriate or irrelevant. Our participants in the two user studies noted their dissatisfaction with the shared context on multiple occasions. In this situation, the agent can detect its error if the user tries to remedy the situation immediately by attending to the message within the inattentiveness threshold. In this case, the agent can prompt the user and get feedback on whether the agent's response was undesired or the context it shared.

Recognizing its mistakes allows the agent to try and remedy the situation and learn more about their users' preferences. This would allow the agent to limit its inquiries to its users only in cases where the outcomes were undesired instead of inquiring about each of its actions which can be annoying for the user [17].

7.2.3 Interactive agent design

People often engage in back-and-forth exchanges to clarify and explain their perspectives rather than use disconnected remarks [76]. The current agent responses consist only of a

one-way signal of unavailability to the message senders. The agent was designed to inform unavailability and improve situational awareness of the message recipient's state. But as noted earlier, as the agent acts as an intermediary in human conversations, users expect the agent to conform to social norms. We discussed situations in Chapter 5, where these responses could be perceived as not useful or even be misinterpreted due to ambiguity of the shared information (Section 5.5.3.3). Augmenting the agent design to allow follow-up inquiries can improve the sense of utility for message senders and prevent misinterpretations related to these shared contexts. These follow-up inquiries can be used to explain the unclear context that the agent shared, which could be augmented with additional information or other contextual information. For instance, if the agent shares the ambient noise value and further information is requested, the agent can augment that by adding location information. Further research is required to understand how we can augment the shared context with additional information and the associated privacy considerations of this additional context.

Investigating further extending agent capabilities to increase utility might also be helpful. Prior research has looked into automated scheduling [48] and passing through notifications [46] depending on user interaction. This allows for the agent to have additional capabilities as an intermediary. For instance, depending on the communication context, e.g., in an urgent communication, the agent could alert or suggest ways to get in touch with message recipients. This was observed on multiple occasions where participants discussed wanting particular agent behavior in urgent situations (Sections 5.5.2.1 and 6.3.4). Further, humans like getting confirmatory feedback and acknowledgments as evidence of performing an action [204, 187]. As noted in Section 5.5.4.3, message senders might perceive the agent as a barrier to reaching the message recipient. In these cases, an acknowledgment from the agent that the message was received or scheduling an event for follow-up might assure message senders that they are not being 'blocked' by the agent. Although, further research is needed to develop effective scheduling mechanisms that do not create additional pressure and expectations on message recipients to respond based on agent scheduled times.

7.3 Privacy considerations in agent design

As mentioned in Section 4.5.2, there were multiple privacy considerations in the design of the agent. For instance, the agent only shared context when certain conditions were met. First, a message must be received on the agent owner's phone, i.e., communication needs to be *initiated*. Second, the communication initiation has to be a new messaging *session* that considers the time since the last message and the message sender (more details in Section 5.2.1.1). Third, the agent needs to predict the agent owner as *unavailable* to take action, i.e., share context through an auto-response. Fourth, the message must be from a known *contact* stored on the agent owner's phone. These considerations served two purposes, (1) since information is only being shared when message senders initiate communication, the recipient is **aware** of the information that has been shared and with whom [46, 144] and (2) even though more information is being shared about the recipients' context, it is limited in terms of **accessibility** compared to existing cues in messaging applications and awareness displays proposed in other works [57, 109].

In addition to these considerations related to the accessibility of the agent owner's context, we ensured that we did not collect or store more information than we needed. For instance, the agent did not read or process the content of text messages as it was not required to predict a user's attentiveness to their messages (it may be influential in predicting responsiveness [135]). Only the metadata information, such as the time of the message, sender information, and which application the message arrived on, was captured. Further, the message sender information was only used to verify if the message came from a known contact and was used to identify a new messaging session. The information stored on the remote server included message metadata (except sender or contact information) and sensor and usage data representing user context at the time of the incoming message. Regarding sensor and usage data captured by the agent, we used mid-level sensed data as described in Section 5.2.2. For example, instead of storing the exact GPS coordinates, we store the semantic label the user assigned to a location. This prevented storing very granular data, which users could perceive as privacy-sensitive.

Further, we included controls to reduce undesired information disclosure through agent

actions. We included a control to add a delay to agent responses (adjustable from 0 to 7 minutes). This allowed users time to attend to incoming messages before the agent took action. We also included a control to prevent another auto-response to the same contact within a specific time (0 to 6 hours) if that contact has already received an agent response. This control was added to limit information disclosure for the same contact. More details about these controls can be found in Section 5.3.3.

7.3.1 Privacy concerns with agent shared context

In the agent evaluation study (Chapter 5), we noted that multiple participants were concerned about additional information disclosure due to the context the agent shared that could be perceived as speculative. For instance, P11 noted that sharing calendar event information can result in contacts inquiring about additional information about these events (Section 5.5.3.3). Further, in Section 5.5.3.3, we noted that agent-shared context could also be misinterpreted. For instance, sharing a *noisy environment* as an auto-response was often equated to being *at a party*. Thus, there is a need to have additional controls to allow participants to tune the agent model according to their preferences to avoid undesired information disclosure through the agent. At the same time, user privacy preferences may differ when asked about it compared to how they feel or behave in practice [140]. We reported a similar observation in Section 7.2.1. Thus, a more iterative design process involving users may be more appropriate to understand their context-sharing preferences.

7.3.2 Additional privacy controls moving forward

7.3.2.1 Improving awareness of agent actions

We can further reduce false positives, i.e., the agent sending auto-responses when not needed, by improving awareness of impending agent action. Figure 26 shows how the agent can generate notifications to improve user awareness of when it is about to take action, along with controls to prevent or even force agent actions (Figure 25).

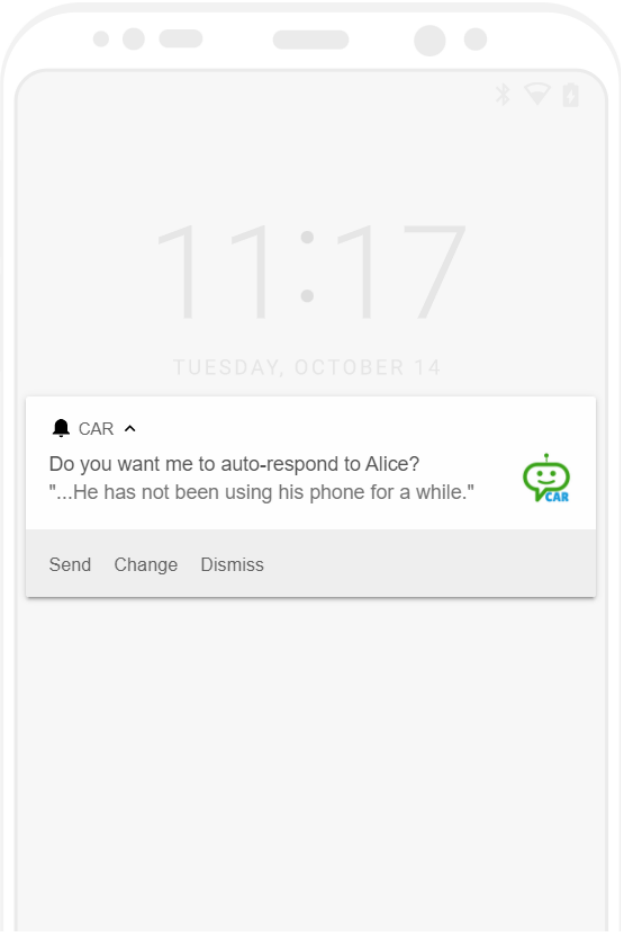


Figure 25: Forced action

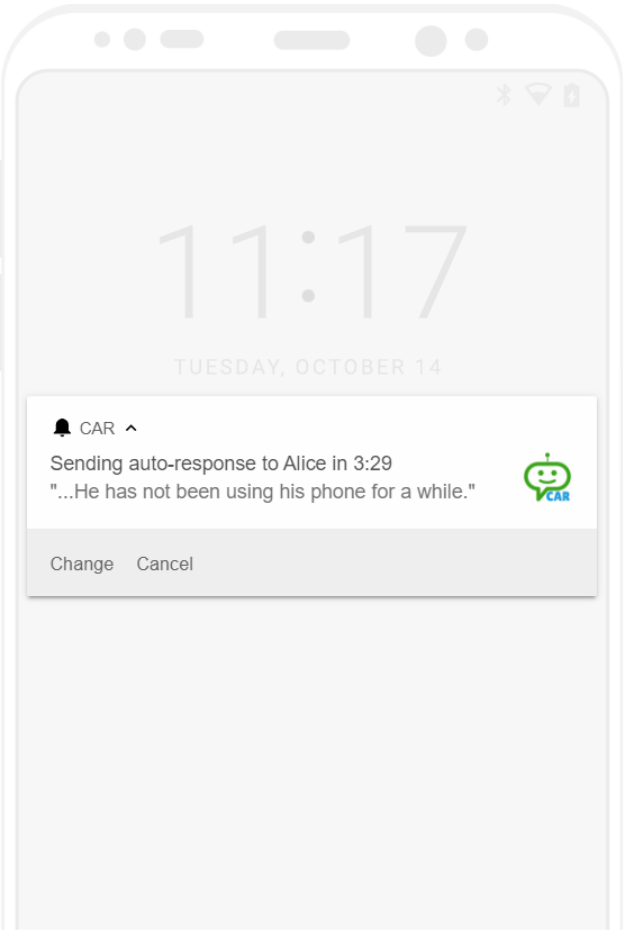


Figure 26: Preventive action

7.3.2.2 Selective information disclosure

In its current design, the agent does not allow users to selectively disable particular contexts, such as noise value, to be used in modeling or context sharing. Allowing users to selectively disable context they do not want the agent to utilize or share can further improve user privacy related to data collection and information disclosure. Although, as we noted in Section 5.5.3.2, the perception of comfort with sharing a category of auto-response sometimes also depended on the value of that category. For instance, sharing the 'productivity app' auto-response was perceived to be useful compared to sharing the

'entertainment app' auto-response, which the users were also uncomfortable with sharing. Thus, additional investigation is needed to identify the best approaches towards accounting for user preferences in context sharing while still keeping the utility of the categories for some value of auto-responses.

7.3.2.3 Gaining additional context from text messages

In the evaluation study (Chapter 5), there were multiple instances where participants reported that agent response was not needed due to the purpose or the content of the incoming message (e.g., emojis). As mentioned earlier, as a privacy measure, the agent does not read, store or send the contents of the user's text messages to gain additional context related to their messaging activities. Although allowing for the agent to parse user text messages will allow it to detect instances of messages where an auto-response is not needed, such as those that represent the end of a conversation or emojis where there is no expectation of a fast response. Allowing the agent to read text messages will also enable a richer communication context for the agent to leverage. For instance, detecting whether a conversation is urgent. This will allow for further response selection based on the identified communication context.

Reading the contents of the text message comes at the cost of user privacy. Further investigation is needed to understand how to balance user privacy and context for the agent. Potential solutions used in text processing include anonymizing the text message before processing [143], processing on-device rather than on a remote server [61], and removing sensitive information before processing [167]. Processing text messages on-device, similar to how we process sensor data on the user's phone, can be particularly useful to gain *enough* context (e.g., urgency or end of conversation) while mitigating the issue of sharing and storing sensitive information on a remote server. Further, controls can be provided to the user to have more agency in how the agent processes text messages. For instance, in addition to making users aware of the agent's ability to read message contents, the users can be presented with controls to enable message reading for selected contacts only.

7.4 How much engagement can we expect from the user to align agent behavior to their expectations?

In the co-design study (Chapter 6), we presented several controls our participants designed to increase agency over agent behavior. However, prior research has shown that user engagement with system controls is generally low [84]. Further, user interfaces may become challenging to understand when too many controls are included, increasing the cognitive load [98]. This load is further exaggerated with multiple applications on users' phones, each with complex controls of their own [217]. While we want users to have increased agency over agent behavior, it is also undesirable to have too much engagement with agent controls as the broad goal of the agent is also to reduce distractions from device use. If the user spends a substantial amount of time trying to tweak the agent's behavior, it reduces the agent's effectiveness in reducing the user's device engagement. Thus, there is a need to understand where the tipping point exists where users are unwilling to nor should be expected to engage further with agent controls.

We speculate that this tipping point would highly depend on individual users as they may have substantially different situations in how they use messaging and engage with their devices. For instance, we noted in Section 5.5.2.1, that some participants interacted with other people as part of their jobs and thus had different expectations of how they wanted the agent to function for these interactions. Further, the frequency of engagement with the agent will differ during early use compared to once the novelty wears off. We reported this in the evaluation study in Chapter 5. Indeed the novelty factor is prominent as users engage with new technology [141]. Thus, when determining this tipping point, we need to account for the use period for the agent.

Additionally, we need to consider the impact of long-term control engagement from the users. Prior work has reported that a clear benefit of the controls, along with ease of use and access, is important in the design of the control mechanism [84]. The design of controls by our participants in the co-design study (Chapter 6) also included considerations for easy access to these controls (e.g., through persistent notifications or localized in the agent responses). Thus, in determining the tipping point, we may need to account for how much engagement

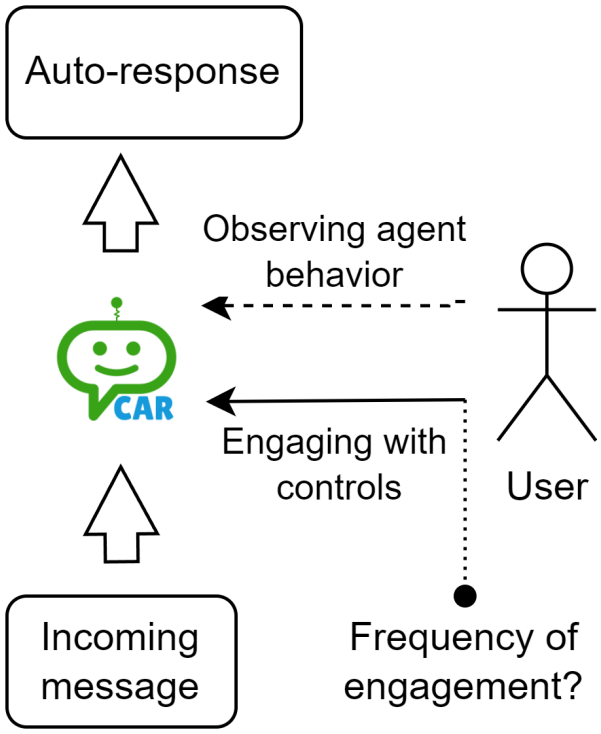


Figure 27: Balancing user engagement with the agent controls is crucial to improve agent utility for its users.

is required from the user to use that control.

7.4.1 Issues with user literacy of the agent design and function

In Section 6.3.5.1, we discussed situations where participants developed an inaccurate mental model of how the agent works, resulting in them requesting controls for aspects of the agent that do not exist. Based on this finding, we suggested using operational and mechanistic explanations to allow users to understand the agent's workflow and how the agent controls work during the early stages of the agent use.

From the co-design study, we also reported that users mental models evolved as they observed more agent actions and interactions with their partners. Through this improved learning about the agent, the users requested more meaningful controls than speculating on

agent behavior based on limited observations.

Thus, one way we can ensure user literacy would be to allow users to familiarize themselves with the agent function. As suggested in the co-design study chapter, the agent can start at lower levels of automation, such as proactive suggestions to allow users to understand how it works and, at the same time, provide feedback to it. Once the agent reaches a certain level of confidence in understanding user preferences, it can start acting with a high level of proactivity. Similarly, user understanding should be higher at this stage as the user would have observed multiple agent outcomes to understand its function better.

8.0 Conclusion

Mobile messaging has become increasingly popular in the age of smartphones. We are expected to be constantly connected and ready to respond. At any time, a large number of applications and other stimuli, such as social media, are competing for our attention. Given the limits of our attention economy, we need to design technology to gain control back of how much attention we give to these applications.

In this thesis, we demonstrated that it is possible to automatically acquire rich context from smartphone sensors to build highly accurate user attention models. We also showed that it is possible to design an intelligent agent that can leverage these user models to predict and share unavailability-related context automatically. Our two-week-long evaluation study of this agent showed that the participants found the agent helpful in communicating unavailability to their contacts and reducing their perceived obligation to respond. Participants also reported the agent's potential to help them reduce their device engagement and interruptions from incoming messages. Finally, through a co-design study, we explored important design considerations for making the messaging agent more intelligible for its users to allow for its better appropriation.

Through this work, we took steps towards realizing the goal of designing a virtual assistant that we can rely on to mediate our messaging interactions. A successful design of the agent has the potential to reduce interruptions from incoming messaging notifications while improving situational awareness and re-establish the asynchronous nature of mobile messaging.

8.1 Contributions

Through this thesis, we contribute to the knowledge body on user modeling, context sharing, designs of proactive agents, and explainable AI. In particular, we make the following contributions:

1. (**User modeling**) Through empirical evaluation, we highlight the importance of utilizing a personalized modeling approach to detect unavailability in mobile messaging accurately. We also provide evidence of variability in what these individual models are learning and what are the data requirements for effective use of the personalized modeling approach.
2. (**User modeling**) We compare and evaluate the modeling performance for different approaches to identifying user groups. We present evidence and interpretation of device usage-based user clustering in predicting unavailability.
3. (**User modeling**) We propose a novel ensemble modeling approach to modeling messaging attentiveness and show its effectiveness in tackling the cold-start problem on a large-scale notification dataset. We also show the importance of incorporating group models in this ensemble approach.
4. (**Context-sharing**) We identify and report the different categories of contextual information a messaging agent can share through automated responses through literature review and analysis of a real-world messaging corpus.
5. (**Context-sharing**) We show the importance of information type (device or user), social tie strength, and urgency of communication in the perception of utility and comfort with agent-shared responses.
6. (**Context-sharing**) We propose and show the effectiveness of initializing user preferences related to context-sharing through a tree-based model.
7. (**Agent design**) We present a novel design of a fully automated messaging agent that learns from users messaging behavior to identify and share relevant context related to their unavailability state. We discuss essential design guidelines that can improve the perception of a communication agent as a representative of its users.
8. (**Human-agent interaction**) We describe ways in which this agent can be helpful for its users in informing unavailability and improving situational awareness and what factors affect its utility for its users. We provide insights by empirically evaluating the agent's role in both positive and negative user behavior changes.
9. (**Explainable-AI**) Our work contributes to understanding how people reason about a proactive messaging agent's design and actions and how that informs their motivations for desiring explanations and affecting change in the agent's behavior.

10. (**Explainable-AI**) We provide directions for designing explanation interfaces for a proactive messaging agent and the importance of enabling user-agent interactions that consider social norms employed in human-human communication.

The results of this research and methodology for evaluation, for example, bringing users into the design process for agent explanations, can serve as the foundation for other researchers and practitioners to understand essential aspects in designing future agents that are highly proactive, interactive, and intelligible.

8.2 Summary

Referring back to our thesis statement presented in Section 1.1.1:

> *It is possible to design an intelligible virtual assistant through user-centered design that can leverage mobile usage and sensor data to improve situational awareness in mobile messaging by predicting user unavailability and sharing relevant unavailability context.*

Through contributions (1-3), we showed that we can build highly accurate models that can be used to predict unavailability even in the face of a lack of data for new users. From contributions (4-6), we identified and reported on the utility and comfort associated with different contextual information that can be used for improving situational awareness in messaging. Through contributions (7-8), we designed and implemented a fully-functional messaging agent and did a real-world evaluation providing evidence that the agent can improve situational awareness in messaging, reduce the perceived obligation to respond, and subsequently has the potential to reduce device engagement and distractions from incoming messaging notifications. Finally, with contributions (9-10), by involving users directly in the design process, we presented ways to make this agent more intelligible for its users by designing explanations targeting the gaps in their knowledge, accommodating aspects of socially acceptable behavior as part of the agent design to be accepted as an intermediary in human-human communication.

8.3 Future Work

Next, we discuss some directions for future work related to this research.

8.3.1 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 [22]. 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 [96]. With more smart-home systems and intelligent agents integrated into our daily lives, bystander privacy has recently been an active area of research [3, 22]. While in this dissertation, 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.

8.3.2 Incorporating social norms in the agent design

As discussed in Section 6.4.2, since the agent acts as an intermediary in human-human communication, it can potentially disrupt the social norms people follow in interpersonal communication. We observed multiple cases where our participants in both the user studies tried to explain and justify the agent's actions to their contacts as a repair mechanism whenever they felt that the agent outcome was inappropriate. Even a single mistake by the agent, including taking action when not needed, could negatively affect their social relationships, which our participants reported in multiple instances. In addition to affecting

their interpersonal relationships, there is also potential for other significant consequences of agent action, for example, affecting someone in a professional setting by disclosing harmful context.

So how do we incorporate social norms in the agent design? Future studies should consider co-design methodology to identify and discuss various situations in the use of the agent and how do people expect the agent to handle those situations. A comprehensive set of these situations can then be used to identify the context (e.g., location and time) associated with these situations of agent use. It can help create feedback loops to tailor the agent behavior to their user's preferences in these contexts. The limitation of this research is also our exploration of only limited dimensions related to agent use and context sharing. For instance, we only considered the communication context (urgency and social relation) when understanding the perception of utility and comfort associated with sharing different categories of auto-responses. Once we can identify additional dimensions that can affect user perception of the agent, future studies can aim to evaluate the significance of these additional dimensions and how to design agents to accommodate these other dimensions. For instance, if the *time of the day* is a significant factor in how people interpret the agent-shared environmental information, then future agent designs can augment the agent knowledge base with rules related to context-sharing at different times of the day.

Finally, another possibility worth exploring is conducting these co-design sessions with familiar or related dyads who regularly interact through text messaging. 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 information exploration than pairing strangers [113]. Social norms are complex and can depend on multiple factors, such as culture, communities, and relationships. Thus, setting up varied pairings or groups during these co-design sessions may help identify additional pointers related to social norms when discussing and negotiating interactions.

8.3.3 Agent-agent interaction in human communication

The capabilities of virtual assistants continue to improve, and they are being integrated with more applications. We demonstrated one virtual assistant design that can communicate, albeit just a single message, on their user's behalf for an incoming communication attempt. But what if an agent initiated the communication instead of a human? Let's consider a scenario. User A wants to schedule a meeting with User B. They ask their agent to send a message to User B asking about their availability for an appointment. The agent on User B's phone auto-responds, sharing User B's unavailability to respond. How does User A's agent use that information? What does User B think about their agent sharing their context with another agent? Future studies can explore context-sharing preferences for these agent-agent interactions and user privacy concerns related to potential agent-agent interactions.

8.3.4 Generalizability of our results for other agent domains

Through the results of this research, we contributed to the research on user modeling, context-sharing, agent design, and explainable AI.

Regarding **user modeling**, we showed how to leverage automatically acquired data (sensor and device usage) from a user's smartphone to model their messaging behavior. In particular, we showed the value of a personalized modeling approach in utilizing smartphone data. The approaches explored in this work towards modeling user behavior through smartphone data can be applied to other classification tasks of predicting certain user states or interactions with mobile devices. Further, we showed the value of usage-based group modeling in tackling the cold-start problem in personalized modeling. Future studies can aim to explore the use of group modeling and combining different modeling approaches to other domains, such as recommender systems.

For **context-sharing**, our results showed that perception of utility and comfort varied based on the communication context (urgency and social relation) and the information type (device or user state). More applications are now employing context-sharing in their design. We already have automated burglar alarm systems that can share context with law enforcement and smart devices (phones or watches) which share medical data with emergency

personnel. Future studies can evaluate user perception of sharing additional smartphone context in various settings and confirm whether the factors identified in this research hold significance in these varied settings.

Design of proactive agents: In this dissertation, we explored a proactive agent design tasked with mediating messaging interactions on behalf of their users. Our results showed that participants found the agent helpful in reducing their obligations to respond and thus reducing device engagement and distractions caused by incoming messaging notifications. The process we followed to design a highly automated proactive agent can be applied to other domains, such as mhealth and persuasive systems. Leveraging user models to predict user states can help determine when to automate agent behavior for a specific purpose. For instance, we can design proactive agents to nudge users to make healthier diet, exercise, and mental well-being choices.

Further. the results of our co-design study showed that due to the social aspects associated with the agent acting as an intermediary in human communication, there was a heightened sense of having more control over the agent outcomes, e.g., to avoid negative implications of inappropriate context shared through the agent. Future research can explore using the co-design methodology for designing explanations for proactive agent designs in other domains, such as recommender systems. These future studies can help designers develop a generalized framework that can account for task criticality in the design of future proactive agents to present guidelines for incorporating explanations that would be helpful for users to understand agent behavior that they care about the most based on its tasks.

8.3.5 Understanding long terms effects of mobile agent usage on user behavior and engagement with device

Regular interaction with technology can invoke behavior change in users [205]. Prior research and our research findings suggest that mobile virtual assistants can support positive behavior change [107, 199, 53]. Although, it is unclear how long-term use of this agent will affect user behavior. Prior research has shown that people often appropriate technology to suit their needs [162], often in unexpected ways [142], which can even have negative

consequences such as privacy implications [157]. Investigating the agent use over the long term would help us understand how users appropriate the agent over extended usage and what influences these appropriations. New technologies, applications, and even extended capabilities of existing virtual agents can all potentially overwhelm the user. Considering the exogenous changes in society thus becomes very important. Understanding extended usage can help us determine unintended uses and consequences detrimental to agent acceptance and trust. Investigating long-term use can also help us design adaptive approaches towards accommodating variations in user behavior, subsequently maintaining or improving agent utility for their users.

Bibliography

[1] Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y. Lim, and Mohan Kankanhalli. 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*, CHI '18, page 1–18, New York, NY, USA, 2018. Association for Computing Machinery.

[2] Ankit Agrawal and Jane Cleland-Huang. 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):15–26, Oct. 2021.

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

[4] Ionut Andone, Konrad Błaszkiewicz, Mark Eibes, Boris Trendafilov, Christian Montag, and Alexander Markowetz. How age and gender affect smartphone usage. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct*, UbiComp '16, page 9–12, New York, NY, USA, 2016. Association for Computing Machinery.

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

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

[7] Hagai Attias. A variational baysian framework for graphical models. In *Advances in neural information processing systems*, pages 209–215, 2000.

[8] Daniel Avrahami, Darren Gergle, Scott E Hudson, and Sara Kiesler. Improving the match between callers and receivers: A study on the effect of contextual information on cell phone interruptions. *Behaviour & Information Technology*, 26(3):247–259, 2007.

[9] Daniel Avrahami and Scott E. Hudson. Qna: Augmenting an instant messaging client to balance user responsiveness and performance. In *Proceedings of the 2004 ACM Conference on Computer Supported Cooperative Work*, CSCW '04, page 515–518, New York, NY, USA, 2004. Association for Computing Machinery.

[10] Daniel Avrahami and Scott E. Hudson. Responsiveness in instant messaging: Predictive models supporting inter-personal communication. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '06, page 731–740, New York, NY, USA, 2006. Association for Computing Machinery.

[11] Naveen Farag Awad and M. S. Krishnan. The personalization privacy paradox: An empirical evaluation of information transparency and the willingness to be profiled online for personalization. *MIS Quarterly*, 30(1):13–28, 2006.

[12] Brian P Bailey, Joseph A Konstan, and John V Carlis. The effects of interruptions on task performance, annoyance, and anxiety in the user interface. In *Human-Computer Interaction: IFIP TC13 International Conference on Human-Computer Interaction, July 9-13*, INTERACT '01, pages 593–601. IOS Press, 2001.

[13] João Balsa, Isa Félix, Ana Paula Cláudio, Maria Beatriz Carmo, Isabel Costa e Silva, Ana Guerreiro, Maria Guedes, Adriana Henriques, and Mara Pereira Guerreiro. Usability of an intelligent virtual assistant for promoting behavior change and self-care in older people with type 2 diabetes. *Journal of Medical Systems*, 44(7):1–12, 2020.

[14] Jakob E Bardram and Thomas R Hansen. Context-based workplace awareness. *Computer Supported Cooperative Work (CSCW)*, 19(2):105–138, 2010.

[15] Simon Baron-Cohen. *Mindblindness: An essay on autism and theory of mind*. MIT press, 1997.

[16] Eric Bates and Lynda R. Wiest. Impact of personalization of mathematical word problems on student performance. *The Mathematics Educator*, 14:17–26, 2004.

[17] Dennis Becker, Vincent Bremer, Burkhardt Funk, Joost Asselbergs, Heleen Riper, and Jeroen Ruwaard. How to predict mood? delving into features of smartphone-based data. *Twenty-second Americas Conference on Information Systems*, pages 1–10, 2016.

[18] James "Bo" Begole, Nicholas E. Matsakis, and John C. Tang. Lilsys: Sensing unavailability. In *Proceedings of the 2004 ACM Conference on Computer Supported*

Cooperative Work, CSCW '04, page 511–514, New York, NY, USA, 2004. Association for Computing Machinery.

[19] James Bo Begole and John C Tang. Incorporating human and machine interpretation of unavailability and rhythm awareness into the design of collaborative applications. *Human–Computer Interaction*, 22(1-2):7–45, 2007.

[20] Victoria Bellotti and Sara Bly. Walking away from the desktop computer: Distributed collaboration and mobility in a product design team. In *Proceedings of the 1996 ACM Conference on Computer Supported Cooperative Work*, CSCW '96, page 209–218, New York, NY, USA, 1996. Association for Computing Machinery.

[21] Paul N. Bennett, Ryen W. White, Wei Chu, Susan T. Dumais, Peter Bailey, Fedor Borisyuk, and Xiaoyuan Cui. Modeling the impact of short- and long-term behavior on search personalization. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '12, page 185–194, New York, NY, USA, 2012. Association for Computing Machinery.

[22] Julia Bernd, Ruba Abu-Salma, and Alisa Frik. 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, August 2020.

[23] Reuben Binns, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. '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*, CHI '18, page 1–14, New York, NY, USA, 2018. Association for Computing Machinery.

[24] David M Blei, Michael I Jordan, et al. Variational inference for dirichlet process mixtures. *Bayesian analysis*, 1(1):121–143, 2006.

[25] Sabri Boughorbel, Fethi Jarray, and Mohammed El-Anbari. Optimal classifier for imbalanced data using matthews correlation coefficient metric. *PloS one*, 12(6):e0177678, 2017.

[26] Clara Bove, Jonathan Aigrain, Marie-Jeanne Lesot, Charles Tijus, and Marcin Detyniecki. 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*, IUI '22, page 807–819, New York, NY, USA, 2022. Association for Computing Machinery.

[27] Michael Boyle, Carman Neustaedter, and Saul Greenberg. Privacy factors in video-based media spaces. In *Media Space 20+ Years of Mediated Life*, pages 97–122. Springer, 2009.

[28] Dawn O Braithwaite and Paul Schrodt. *Engaging theories in interpersonal communication: Multiple perspectives*. Routledge, 2021.

[29] Virginia Braun and Victoria Clarke. *Successful qualitative research: A practical guide for beginners*. sage, 2013.

[30] Alex Braunstein, Laura Granka, and Jessica Staddon. Indirect content privacy surveys: Measuring privacy without asking about it. In *Proceedings of the Seventh Symposium on Usable Privacy and Security*, SOUPS '11, pages 1–14, New York, NY, USA, 2011. Association for Computing Machinery.

[31] Cynthia Breazeal. Emotion and sociable humanoid robots. *International journal of human-computer studies*, 59(1-2):119–155, 2003.

[32] Barry Brown and Louise Barkhuus. Leisure and cscw: Introduction to special edition. *Computer Supported Cooperative Work (CSCW)*, 16(1-2):1–10, 2007.

[33] Dariusz Brzeziński and Jerzy Stefanowski. Accuracy updated ensemble for data streams with concept drift. In *International conference on hybrid artificial intelligence systems*, pages 155–163. Springer, 2011.

[34] Andreas Buchenscheit, Bastian Könings, Andreas Neubert, Florian Schaub, Matthias Schneider, and Frank Kargl. Privacy implications of presence sharing in mobile messaging applications. In *Proceedings of the 13th International Conference on Mobile and Ubiquitous Multimedia*, MUM '14, page 20–29, New York, NY, USA, 2014. Association for Computing Machinery.

[35] Nikolay Burlutskiy, Miltos Petridis, Andrew Fish, Alexey Chernov, and Nour Ali. An investigation on online versus batch learning in predicting user behaviour. In *International Conference on Innovative Techniques and Applications of Artificial Intelligence*, pages 135–149. Springer, 2016.

[36] Alexandra Burton, Claudia Cooper, Ayesha Dar, Lucy Mathews, and Kartikeya Tripathi. Exploring how, why and in what contexts older adults are at risk of financial cybercrime victimisation: A realist review. *Experimental Gerontology*, 159:111678, 2022.

[37] Daniel Buschek, Mariam Hassib, and Florian Alt. Personal mobile messaging in context: Chat augmentations for expressiveness and awareness. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 25(4):1–33, 2018.

[38] Jeff K. Caird, Kate A. Johnston, Chelsea R. Willness, Mark Asbridge, and Piers Steel. A meta-analysis of the effects of texting on driving. *Accident Analysis & Prevention*, 71:311 – 318, 2014.

[39] Yung-Ju Chang, Yi-Ju Chung, and Yi-Hao Shih. I think it's her: Investigating smartphone users' speculation about phone notifications and its influence on attendance. In *Proceedings of the 21st International Conference on Human-Computer Interaction with Mobile Devices and Services*, pages 1–13, 2019.

[40] Yung-Ju Chang and John C. Tang. Investigating mobile users' ringer mode usage and attentiveness and responsiveness to communication. In *Proceedings of the 17th International Conference on Human-Computer Interaction with Mobile Devices and Services*, MobileHCI '15, page 6–15, New York, NY, USA, 2015. Association for Computing Machinery.

[41] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[42] Nitesh V Chawla and Jared Sylvester. Exploiting diversity in ensembles: Improving the performance on unbalanced datasets. In *International Workshop on Multiple Classifier Systems*, pages 397–406. Springer, 2007.

[43] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 785–794, New York, NY, USA, 2016. Association for Computing Machinery.

[44] Zhenyu Chen, Mu Lin, Fanglin Chen, Nicholas D. Lane, Giuseppe Cardone, Rui Wang, Tianxing Li, Yiqiang Chen, Tanzeem Choudhury, and Andrew T. Campbell. Unobtrusive sleep monitoring using smartphones. In *Proceedings of the 7th International Conference on Pervasive Computing Technologies for Healthcare*, PervasiveHealth '13, page 145–152, Brussels, BEL, 2013. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering).

[45] Davide Chicco and Giuseppe Jurman. The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC genomics*, 21:1–13, 2020.

[46] Hyunsung Cho, Jinyoung Oh, Juho Kim, and Sung-Ju Lee. 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):1–25, 2020.

[47] Sungjoon Choi, Eunwoo Kim, and Songhwai Oh. Human behavior prediction for smart homes using deep learning. In *RO-MAN*, volume 2013, page 173, 2013.

[48] Anand Chowdhary. Email-based intelligent virtual assistant for scheduling (eiva). B.S. thesis, University of Twente, 2020.

[49] Michael Chromik, Malin Eiband, Felicitas Buchner, Adrian Krüger, and Andreas Butz. I think i get your point, ai! the illusion of explanatory depth in explainable ai. In *26th International Conference on Intelligent User Interfaces*, IUI '21, page 307–317, New York, NY, USA, 2021. Association for Computing Machinery.

[50] Karen Church and Rodrigo de Oliveira. What's up with whatsapp? comparing mobile instant messaging behaviors with traditional sms. In *Proceedings of the 15th International Conference on Human-Computer Interaction with Mobile Devices and Services*, MobileHCI '13, page 352–361, New York, NY, USA, 2013. Association for Computing Machinery.

[51] Herbert H Clark and Susan E Brennan. Grounding in communication. *Perspectives on socially shared cognition*, 1991.

[52] Victoria Clarke and Virginia Braun. *Successful qualitative research: A practical guide for beginners*. Sage publications ltd, 2013.

[53] Drew Clinkenbeard, Jennifer Clinkenbeard, Guillaume Faddoul, Heejung Kang, Sean Mayes, Alp Toygar, and Samir Chatterjee. What's your 2%? a pilot study for encouraging physical activity using persuasive video and social media. In Anna Spagnolli, Luca Chittaro, and Luciano Gamberini, editors, *Persuasive Technology*, pages 43–55, Cham, 2014. Springer International Publishing.

[54] Sunny Consolvo, Ian E. Smith, Tara Matthews, Anthony LaMarca, Jason Tabert, and Pauline Powledge. Location disclosure to social relations: Why, when, & what people want to share. In *Proceedings of the SIGCHI Conference on Human Factors*

in Computing Systems, CHI '05, page 81–90, New York, NY, USA, 2005. Association for Computing Machinery.

[55] Mary L Cummings. Automation bias in intelligent time critical decision support systems. In *Decision Making in Aviation*, pages 289–294. Routledge, 2017.

[56] Allan de Barcelos Silva, Marcio Miguel Gomes, Cristiano André da Costa, Rodrigo da Rosa Righi, Jorge Luis Victoria Barbosa, Gustavo Pessin, Geert De Doncker, and Gustavo Federizzi. Intelligent personal assistants: A systematic literature review. *Expert Systems with Applications*, 147:113193, 2020.

[57] Edward S. De Guzman, Moushumi Sharmin, and Brian P. Bailey. Should i call now? understanding what context is considered when deciding whether to initiate remote communication via mobile devices. In *Proceedings of Graphics Interface 2007*, GI '07, page 143–150, New York, NY, USA, 2007. Association for Computing Machinery.

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

[59] Anind K Dey. Understanding and using context. *Personal and ubiquitous computing*, 5(1):4–7, 2001.

[60] Anind K Dey, Katarzyna Wac, Denzil Ferreira, Kevin Tassini, Jin-Hyuk Hong, and Julian Ramos. Getting closer: an empirical investigation of the proximity of user to their smart phones. In *Proceedings of the 13th international conference on Ubiquitous computing*, pages 163–172, 2011.

[61] Sauptik Dhar, Junyao Guo, Jiayi Liu, Samarth Tripathi, Unmesh Kurup, and Mohak Shah. A survey of on-device machine learning: An algorithms and learning theory perspective. *ACM Transactions on Internet of Things*, 2(3):1–49, 2021.

[62] Thomas G Dietterich. Machine learning for sequential data: A review. In *Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR)*, pages 15–30. Springer, 2002.

[63] Djellel Difallah, Elena Filatova, and Panos Ipeirotis. Demographics and dynamics of mechanical turk workers. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, WSDM '18, page 135–143, New York, NY, USA, 2018. Association for Computing Machinery.

[64] Tilman Dingler and Martin Pielot. I'll be there for you: Quantifying attentiveness towards mobile messaging. In *Proceedings of the 17th International Conference on Human-Computer Interaction with Mobile Devices and Services*, pages 1–5. ACM, 2015.

[65] Paul Dourish. What we talk about when we talk about context. *Personal and ubiquitous computing*, 8(1):19–30, 2004.

[66] Paul Dourish and Sara Bly. Portholes: Supporting awareness in a distributed work group. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '92, page 541–547, New York, NY, USA, 1992. Association for Computing Machinery.

[67] Kelsey Kathleen Earle. *Attributions Online: An Examination of Time Stamps, Read Receipts, and Ellipses in Text-Based Communication.* PhD thesis, North Dakota State University, 2018.

[68] Ryan Elwell and Robi Polikar. Incremental learning of concept drift in nonstationary environments. *IEEE Transactions on Neural Networks*, 22(10):1517–1531, 2011.

[69] Richard Emanuel, Rodney Bell, Cedric Cotton, Jamon Craig, Danielle Drummond, Samuel Gibson, Ashley Harris, Marcus Harris, Chelsea Hatcher-Vance, Staci Jones, et al. The truth about smartphone addiction. *College Student Journal*, 49(2):291–299, 2015.

[70] Mica R Endsley. Situation awareness misconceptions and misunderstandings. *Journal of Cognitive Engineering and Decision Making*, 9(1):4–32, 2015.

[71] Ethan Fast and Eric Horvitz. Long-term trends in the public perception of artificial intelligence. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, page 963–969. AAAI Press, 2017.

[72] Denzil Ferreira, Vassilis Kostakos, and Anind K Dey. Aware: mobile context instrumentation framework. *Frontiers in ICT*, 2:6, 2015.

[73] Joel E. Fischer, Nick Yee, Victoria Bellotti, Nathan Good, Steve Benford, and Chris Greenhalgh. Effects of content and time of delivery on receptivity to mobile interruptions. In *Proceedings of the 12th International Conference on Human Computer Interaction with Mobile Devices and Services*, MobileHCI '10, page 103–112, New York, NY, USA, 2010. Association for Computing Machinery.

[74] Robert Fisher and Reid Simmons. Smartphone interruptibility using density-weighted uncertainty sampling with reinforcement learning.

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

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

[77] Rebecca Grinter and Margery Eldridge. Wan2tlk? everyday text messaging. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 441–448, 2003.

[78] Ted Grover, Kael Rowan, Jina Suh, Daniel McDuff, and Mary Czerwinski. Design and evaluation of intelligent agent prototypes for assistance with focus and productivity at work. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*, pages 390–400, 2020.

[79] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5):1–42, 2018.

[80] Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *Journal of machine learning research*, 3(Mar):1157–1182, 2003.

[81] Jeff Hancock, Jeremy Birnholtz, Natalya Bazarova, Jamie Guillory, Josh Perlin, and Barrett Amos. Butler lies: Awareness, deception and design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '09, page 517–526, New York, NY, USA, 2009. Association for Computing Machinery.

[82] Mark Handel and James D. Herbsleb. What is chat doing in the workplace? In *Proceedings of the 2002 ACM Conference on Computer Supported Cooperative Work*, CSCW '02, page 1–10, New York, NY, USA, 2002. Association for Computing Machinery.

[83] Mark Handel and Graham Wills. Teamportal: Providing team awareness on the web. In *Proceedings of the International Workshop on Awareness and the WWW*, ACM CSCW 2000, pages 3–12, 2000.

[84] Jaron Harambam, Dimitrios Bountouridis, Mykola Makhortykh, and Joris Van Hoboken. Designing for the better by taking users into account: A qualitative evaluation of user control mechanisms in (news) recommender systems. In *Proceedings of the 13th ACM Conference on Recommender Systems*, pages 69–77, 2019.

[85] Steve Harrison. *Media Space 20+ Years of Mediated Life*. Springer Publishing Company, Incorporated, 1st edition, 2009.

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

[87] Eric J Horvitz and Johnson T Apacible. Use of a bulk-email filter within a system for classifying messages for urgency or importance, July 21 2009. US Patent 7,565,403.

[88] Roberto Hoyle, Srijita Das, Apu Kapadia, Adam J. Lee, and Kami Vaniea. Was my message read? privacy and signaling on facebook messenger. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, page 3838–3842, New York, NY, USA, 2017. Association for Computing Machinery.

[89] Scott Hudson, James Fogarty, Christopher Atkeson, Daniel Avrahami, Jodi Forlizzi, Sara Kiesler, Johnny Lee, and Jie Yang. Predicting human interruptibility with sensors: a wizard of oz feasibility study. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 257–264. ACM, 2003.

[90] Ellen Isaacs, Alan Walendowski, Steve Whittaker, Diane J. Schiano, and Candace Kamm. The character, functions, and styles of instant messaging in the workplace. In *Proceedings of the 2002 ACM Conference on Computer Supported Cooperative Work*, CSCW '02, page 11–20, New York, NY, USA, 2002. Association for Computing Machinery.

[91] Wahyu Wibowo Islahulhaq and Iis Dewi Ratih. Classification of non-performing financing using logistic regression and synthetic minority over-sampling technique-nominal continuous (smote-nc). *Int. J. Adv. Soft Comput. Appl*, 13:115–128, 2021.

[92] Jacob Jacoby and Michael S. Matell. Three-point likert scales are good enough. *Journal of Marketing Research*, 8(4):495–500, 1971.

[93] Pranut Jain, Rosta Farzan, and Adam J. Lee. Adaptive modelling of attentiveness to messaging: A hybrid approach. In *Proceedings of the 27th ACM Conference on User*

Modeling, Adaptation and Personalization, UMAP '19, page 261–270, New York, NY, USA, 2019. Association for Computing Machinery.

[94] Pranut Jain, Rosta Farzan, and Adam J. Lee. Are you there? identifying unavailability in mobile messaging. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI EA '19, page 1–6, New York, NY, USA, 2019. Association for Computing Machinery.

[95] Pranut Jain, Rosta Farzan, and Adam J. Lee. Context-based automated responses of unavailability in mobile messaging. *Computer Supported Cooperative Work (CSCW)*, pages 307–349, 2021.

[96] Pranut Jain, Rosta Farzan, and Adam J. Lee. Laila is in a meeting: Design and evaluation of a contextual auto-response messaging agent. In *Designing Interactive Systems Conference*, DIS '22, page 1457–1471, New York, NY, USA, 2022. Association for Computing Machinery.

[97] Pranut Jain, Rosta Farzan, and Adam J. Lee. Co-designing with users the explanations for a proactive auto-response messaging agent. In *Proceedings of the 25th International Conference on Mobile Human-Computer Interaction*, MobileHCI '23, New York, NY, USA, 2023. Association for Computing Machinery.

[98] Yucheng Jin, Nava Tintarev, and Katrien Verbert. Effects of personal characteristics on music recommender systems with different levels of controllability. In *Proceedings of the 12th ACM Conference on Recommender Systems*, pages 13–21, 2018.

[99] Jason D. Johnson, Julian Sanchez, Arthur D. Fisk, and Wendy A. Rogers. Type of automation failure: The effects on trust and reliance in automation. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 48(18):2163–2167, 2004.

[100] Ian Jolliffe. Principal component analysis. In *International encyclopedia of statistical science*, pages 1094–1096. Springer, 2011.

[101] J Wolfgang Kaltz, Jürgen Ziegler, and Steffen Lohmann. Context-aware web engineering: Modeling and applications. *Revue d'intelligence artificielle*, 19(3):439–458, 2005.

[102] Ruogu Kang, Stephanie Brown, Laura Dabbish, and Sara Kiesler. Privacy attitudes of mechanical turk workers and the u.s. public. In *Proceedings of the Tenth USENIX*

Conference on Usable Privacy and Security, SOUPS '14, page 37–49, USA, 2014. USENIX Association.

[103] Ashraf Khalil and Kay Connelly. Context-aware telephony: Privacy preferences and sharing patterns. In *Proceedings of the 2006 20th Anniversary Conference on Computer Supported Cooperative Work*, CSCW '06, page 469–478, New York, NY, USA, 2006. Association for Computing Machinery.

[104] Bitna Kim, Kyung-Shik Shin, and Sangmi Chai. How people disclose themselves differently according to the strength of relationship in sns? *Journal of Applied Business Research*, 31(6):2139, 2015.

[105] Minhyung Kim, Inyeop Kim, and Uichin Lee. Beneficial neglect: Instant message notification handling behaviors and academic performance. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 5(1):1–26, 2021.

[106] Sang Chon Kim, Doyle Yoon, and Eun Kyoung Han. Antecedents of mobile app usage among smartphone users. *Journal of marketing communications*, 22(6):653–670, 2016.

[107] Everlyne Kimani, Kael Rowan, Daniel McDuff, Mary Czerwinski, and Gloria Mark. A conversational agent in support of productivity and wellbeing at work. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–7. IEEE, 2019.

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

[109] Johannes Knittel, Alireza Sahami Shirazi, Niels Henze, and Albrecht Schmidt. Utilizing contextual information for mobile communication. In *CHI '13 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '13, page 1371–1376, New York, NY, USA, 2013. Association for Computing Machinery.

[110] Pigi Kouki, James Schaffer, Jay Pujara, John O'Donovan, and Lise Getoor. Personalized explanations for hybrid recommender systems. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, IUI '19, page 379–390, New York, NY, USA, 2019. Association for Computing Machinery.

[111] Matthias Kraus, Nicolas Wagner, Zoraida Callejas, and Wolfgang Minker. The role of trust in proactive conversational assistants. *IEEE Access*, 9:112821–112836, 2021.

[112] Albrecht Kurze, Andreas Bischof, Sören Totzauer, Michael Storz, Maximilian Eibl, Margot Brereton, and Arne Berger. *Guess the Data: Data Work to Understand How People Make Sense of and Use Simple Sensor Data from Homes*, page 1–12. Association for Computing Machinery, New York, NY, USA, 2020.

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

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

[115] Airi Lampinen, Sakari Tamminen, and Antti Oulasvirta. All my people right here, right now: Management of group co-presence on a social networking site. In *Proceedings of the ACM 2009 International Conference on Supporting Group Work*, GROUP '09, page 281–290, New York, NY, USA, 2009. Association for Computing Machinery.

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

[117] Bjorn Lantz. Equidistance of likert-type scales and validation of inferential methods using experiments and simulations. *The Electronic Journal of Business Research Methods*, 11(1):16–28, 2013.

[118] Christine Lee. How does instant messaging affect interaction between the genders. *Stanford, CA: The Mercury Project for Instant Messaging Studies at Stanford University. Retrieved August*, 11:2006, 2003.

[119] Min-Joong Lee and Chin-Wan Chung. A user similarity calculation based on the location for social network services. In *International Conference on Database Systems for Advanced Applications*, pages 38–52. Springer, 2011.

[120] Uichin Lee, Joonwon Lee, Minsam Ko, Changhun Lee, Yuhwan Kim, Subin Yang, Koji Yatani, Gahgene Gweon, Kyong-Mee Chung, and Junehwa Song. Hooked on smartphones: an exploratory study on smartphone overuse among college students. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 2327–2336, 2014.

[121] Kung-Yee Liang and Scott L. Zeger. Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22, 04 1986.

[122] Q. Vera Liao, Daniel Gruen, and Sarah Miller. Questioning the ai: Informing design practices for explainable ai user experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–15, New York, NY, USA, 2020. Association for Computing Machinery.

[123] Yuting Liao, Jessica Vitak, Priya Kumar, Michael Zimmer, and Katherine Kritikos. Understanding the role of privacy and trust in intelligent personal assistant adoption. In *International Conference on Information*, pages 102–113. Springer, 2019.

[124] Ruoyun Lin and Sonja Utz. Self-disclosure on sns: Do disclosure intimacy and narrativity influence interpersonal closeness and social attraction? *Computers in Human Behavior*, 70:426 – 436, 2017.

[125] Scott M. Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M. Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. From local explanations to global understanding with explainable ai for trees. *Nature Machine Intelligence*, 2(1):2522–5839, 2020.

[126] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 4768–4777, Red Hook, NY, USA, 2017. Curran Associates Inc.

[127] Deborah Lupton, Sarah Pink, Christine Heyes LaBond, and Shanti Sumartojo. 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.

[128] Prashan Madumal, Tim Miller, Liz Sonenberg, and Frank Vetere. A grounded interaction protocol for explainable artificial intelligence. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '19, page 1033–1041, Richland, SC, 2019. International Foundation for Autonomous Agents and Multiagent Systems.

[129] Lisa M Mai, Rainer Freudenthaler, Frank M Schneider, and Peter Vorderer. "i know you've seen it!" individual and social factors for users' chatting behavior on facebook. *Computers in Human Behavior*, 49:296–302, 2015.

[130] Naresh K. Malhotra, Sung S. Kim, and James Agarwal. Internet users' information privacy concerns (iuipc): The construct, the scale, and a causal model. *Information Systems Research*, 15(4):336–355, 2004.

[131] Helia Marreiros, Mirco Tonin, Michael Vlassopoulos, and M.C. Schraefel. "now that you mention it": A survey experiment on information, inattention and online privacy. *Journal of Economic Behavior & Organization*, 140:1 – 17, 2017.

[132] Judith Masthoff. Group recommender systems: Combining individual models. In *Recommender systems handbook*, pages 677–702. Springer, 2011.

[133] Daniel McDuff and Mary Czerwinski. Designing emotionally sentient agents. *Communications of the ACM*, 61(12):74–83, 2018.

[134] Abhinav Mehrotra, Mirco Musolesi, Robert Hendley, and Veljko Pejovic. Designing content-driven intelligent notification mechanisms for mobile applications. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '15, pages 813–824, New York, NY, USA, 2015. ACM.

[135] Abhinav Mehrotra, Veljko Pejovic, Jo Vermeulen, Robert Hendley, and Mirco Musolesi. My phone and me: Understanding people's receptivity to mobile notifications. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, page 1021–1032, New York, NY, USA, 2016. Association for Computing Machinery.

[136] Mehdi Mekni, Zakaria Baani, and Dalia Sulieman. A smart virtual assistant for students. In *Proceedings of the 3rd International Conference on Applications of Intelligent Systems*, pages 1–6, 2020.

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

[138] Jun-Ki Min, Afsaneh Doryab, Jason Wiese, Shahriyar Amini, John Zimmerman, and Jason I. Hong. Toss 'n' turn: Smartphone as sleep and sleep quality detector. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '14, page 477–486, New York, NY, USA, 2014. Association for Computing Machinery.

[139] Christian Montag, Konrad Błaszkiewicz, Rayna Sariyska, Bernd Lachmann, Ionut Andone, Boris Trendafilov, Mark Eibes, and Alexander Markowetz. Smartphone usage in the 21st century: who is active on whatsapp? *BMC research notes*, 8(1):331, 2015.

[140] Jill Mosteller and Amit Poddar. To share and protect: Using regulatory focus theory to examine the privacy paradox of consumers' social media engagement and online privacy protection behaviors. *Journal of Interactive Marketing*, 39(1):27–38, 2017.

[141] Adity U Mutsuddi and Kay Connelly. Text messages for encouraging physical activity are they effective after the novelty effect wears off? In *2012 6th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth) and Workshops*, pages 33–40. IEEE, 2012.

[142] Bonnie A. Nardi, Steve Whittaker, and Erin Bradner. Interaction and outeraction: Instant messaging in action. In *Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work*, CSCW '00, page 79–88, New York, NY, USA, 2000. Association for Computing Machinery.

[143] Hoang-Quoc Nguyen-Son, Minh-Triet Tran, Hiroshi Yoshiura, Noboru Sonehara, and Isao Echizen. Anonymizing personal text messages posted in online social networks and detecting disclosures of personal information. *IEICE TRANSACTIONS on Information and Systems*, 98(1):78–88, 2015.

[144] Karin Niemantsverdriet, Harm Van Essen, Minna Pakanen, and Berry Eggen. Designing for awareness in interactions with shared systems: the dass framework. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 26(6):1–41, 2019.

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

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

[147] Daniel R O'Day and Ricardo A Calix. Text message corpus: applying natural language processing to mobile device forensics. In *IEEE International Conference on Multimedia and Expo Workshops*, ICMEW 2013, pages 1–6, 2013.

[148] Harri Oinas-Kukkonen and Marja Harjumaa. Persuasive systems design: Key issues, process model, and system features. *Communications of the Association for Information Systems*, 24(1):28, 2009.

[149] Bauyrzhan Ospan, Nawaz Khan, Juan Augusto, Mario Quinde, and Kenzhegali Nurgaliyev. Context aware virtual assistant with case-based conflict resolution in multi-

user smart home environment. In *2018 international conference on computing and network communications (coconet)*, pages 36–44. IEEE, 2018.

[150] Motoyuki Ozeki, Shunichi Maeda, Kanako Obata, and Yuichi Nakamura. Virtual assistant: enhancing content acquisition by eliciting information from humans. *Multimedia Tools and Applications*, 44(3):433–448, 2009.

[151] Lindsay C Page and Hunter Gehlbach. How an artificially intelligent virtual assistant helps students navigate the road to college. *AERA Open*, 3(4):2332858417749220, 2017.

[152] Veljko Pejovic and Mirco Musolesi. Interruptme: Designing intelligent prompting mechanisms for pervasive applications. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '14, page 897–908, New York, NY, USA, 2014. Association for Computing Machinery.

[153] Stano Pekár and Marek Brabec. Generalized estimating equations: A pragmatic and flexible approach to the marginal glm modelling of correlated data in the behavioural sciences. *Ethology*, 124(2):86–93, 2018.

[154] Mark Perry, Kenton O'hara, Abigail Sellen, Barry Brown, and Richard Harper. Dealing with mobility: understanding access anytime, anywhere. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 8(4):323–347, 2001.

[155] Martin Pielot. Large-scale evaluation of call-availability prediction. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 933–937. ACM, 2014.

[156] Martin Pielot, Bruno Cardoso, Kleomenis Katevas, Joan Serrà, Aleksandar Matic, and Nuria Oliver. Beyond interruptibility: Predicting opportune moments to engage mobile phone users. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(3):1–25, 2017.

[157] Martin Pielot, Rodrigo de Oliveira, Haewoon Kwak, and Nuria Oliver. Didn't you see my message? predicting attentiveness to mobile instant messages. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '14, page 3319–3328, New York, NY, USA, 2014. Association for Computing Machinery.

[158] Martin Pielot, Tilman Dingler, Jose San Pedro, and Nuria Oliver. When attention is not scarce - detecting boredom from mobile phone usage. In *Proceedings of the*

2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing, UbiComp '15, page 825–836, New York, NY, USA, 2015. Association for Computing Machinery.

[159] Martin Pielot, Amalia Vradi, and Souneil Park. Dismissed! a detailed exploration of how mobile phone users handle push notifications. In *Proceedings of the 20th international conference on human-computer interaction with mobile devices and services*, pages 1–11, 2018.

[160] Martin Porcheron, Joel E Fischer, Stuart Reeves, and Sarah Sharples. Voice interfaces in everyday life. In *proceedings of the 2018 CHI conference on human factors in computing systems*, pages 1–12, 2018.

[161] Byron Reeves and Clifford Nass. How people treat computers, television, and new media like real people and places, 1996.

[162] Ana Paula Retore and Leonelo Dell Anhol Almeida. Understanding appropriation through end-user tailoring in communication systems: A case study on slack and whatsapp. In Gabriele Meiselwitz, editor, *Social Computing and Social Media. Design, Human Behavior and Analytics*, pages 245–264, Cham, 2019. Springer International Publishing.

[163] Lindsay Reynolds, Madeline E. Smith, Jeremy P. Birnholtz, and Jeff T. Hancock. Butler lies from both sides: Actions and perceptions of unavailability management in texting. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*, CSCW '13, page 769–778, New York, NY, USA, 2013. Association for Computing Machinery.

[164] David R Roberts, Volker Bahn, Simone Ciuti, Mark S Boyce, Jane Elith, Gurutzeta Guillera-Arroita, Severin Hauenstein, José J Lahoz-Monfort, Boris Schröder, Wilfried Thuiller, et al. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, 40(8):913–929, 2017.

[165] Stephanie Rosenthal, Anind K. Dey, and Manuela Veloso. Using decision-theoretic experience sampling to build personalized mobile phone interruption models. In *Proceedings of the 9th International Conference on Pervasive Computing*, Pervasive'11, page 170–187, Berlin, Heidelberg, 2011. Springer-Verlag.

[166] Antti Salovaara, Antti Lindqvist, Tero Hasu, and Jonna Häkkilä. The phone rings but the user doesn't answer: Unavailability in mobile communication. In *Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices*

and Services, MobileHCI '11, page 503–512, New York, NY, USA, 2011. Association for Computing Machinery.

[167] David Sánchez, Montserrat Batet, and Alexandre Viejo. Detecting sensitive information from textual documents: an information-theoretic approach. In *Modeling Decisions for Artificial Intelligence: 9th International Conference, MDAI 2012, Girona, Catalonia, Spain, November 21-23, 2012. Proceedings 9*, pages 173–184. Springer, 2012.

[168] Iqbal H Sarker. Silentphone: Inferring user unavailability based opportune moments to minimize call interruptions. *arXiv preprint arXiv:1810.10958*, 2018.

[169] Cristiele A Scariot, Adriano Heemann, and Stephania Padovani. Understanding the collaborative-participatory design. *Work*, 41(Supplement 1):2701–2705, 2012.

[170] Emanuel A Schegloff, Gail Jefferson, and Harvey Sacks. The preference for self-correction in the organization of repair in conversation. *Language*, 53(2):361–382, 1977.

[171] Tim Schrills and Thomas Franke. How to answer why – evaluating the explanations of ai through mental model analysis, 2020.

[172] Hinrich Schütze, Christopher D Manning, and Prabhakar Raghavan. *Introduction to information retrieval*, volume 39. Cambridge University Press, 2008.

[173] Thomas B Sheridan and William L Verplank. Human and computer control of undersea teleoperators. Technical report, Massachusetts Inst of Tech Cambridge Man-Machine Systems Lab, 1978.

[174] Jeremiah Smith, Anna Lavygina, Jiefei Ma, Alessandra Russo, and Naranker Dulay. Learning to recognise disruptive smartphone notifications. In *Proceedings of the 16th international conference on Human-computer interaction with mobile devices & services*, pages 121–124. ACM, 2014.

[175] Barry Smyth, Jill Freyne, Maurice Coyle, Peter Briggs, and Evelyn Balfe. I-spy—anonymous, community-based personalization by collaborative meta-search. In *International Conference on Innovative Techniques and Applications of Artificial Intelligence*, pages 367–380. Springer, 2003.

[176] Kacper Sokol and Peter Flach. 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*, IJCAI'18, page 5868–5870. AAAI Press, 2018.

[177] Clay Spinuzzi. The methodology of participatory design. *Technical communication*, 52(2):163–174, 2005.

[178] Nili Steinfeld. "i agree to the terms and conditions":(how) do users read privacy policies online? an eye-tracking experiment. *Computers in human behavior*, 55:992–1000, 2016.

[179] Wolfgang Stroebe and Michael Diehl. Why groups are less effective than their members: On productivity losses in idea-generating groups. *European review of social psychology*, 5(1):271–303, 1994.

[180] Gail M Sullivan and Anthony R Artino Jr. Analyzing and interpreting data from likert-type scales. *Journal of graduate medical education*, 5(4):541–542, 2013.

[181] Maxwell Szymanski, Martijn Millecamp, and Katrien Verbert. Visual, textual or hybrid: The effect of user expertise on different explanations. In *26th International Conference on Intelligent User Interfaces*, IUI '21, page 109–119, New York, NY, USA, 2021. Association for Computing Machinery.

[182] Kar Yan Tam and Shuk Ying Ho. Understanding the impact of web personalization on user information processing and decision outcomes. *MIS Quarterly*, 30(4):865–890, 2006.

[183] John C Tang. Approaching and leave-taking: Negotiating contact in computer-mediated communication. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 14(1):5–es, 2007.

[184] John C. Tang, Nicole Yankelovich, James Begole, Max Van Kleek, Francis Li, and Janak Bhalodia. Connexus to awarenex: Extending awareness to mobile users. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '01, page 221–228, New York, NY, USA, 2001. Association for Computing Machinery.

[185] Jaime Teevan and Alexander Hehmeyer. Understanding how the projection of availability state impacts the reception incoming communication. In *Proceedings of the*

2013 Conference on Computer Supported Cooperative Work, CSCW '13, page 753–758, New York, NY, USA, 2013. Association for Computing Machinery.

[186] Alexander Thayer, Matthew J. Bietz, Katie Derthick, and Charlotte P. Lee. I love you, let's share calendars: Calendar sharing as relationship work. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*, CSCW '12, page 749–758, New York, NY, USA, 2012. Association for Computing Machinery.

[187] Paul Thomas, Mary Czerwinksi, Daniel McDuff, and Nick Craswell. Theories of conversation for conversational ir. *ACM Transactions on Information Systems (TOIS)*, 39(4):1–23, 2021.

[188] Lauren Thomson, Adam J Lee, and Rosta Farzan. Ephemeral communication and communication places. In *International Conference on Information*, pages 132–138. Springer, 2018.

[189] Nava Tintarev and Judith Masthoff. A survey of explanations in recommender systems. In *2007 IEEE 23rd International Conference on Data Engineering Workshop*, pages 801–810. IEEE Computer Society, April 2007.

[190] Nava Tintarev and Judith Masthoff. *Designing and Evaluating Explanations for Recommender Systems*, pages 479–510. Springer US, Boston, MA, 2011.

[191] Chun-Hua Tsai and Peter Brusilovsky. Evaluating visual explanations for similarity-based recommendations: User perception and performance. In *Proceedings of the 27th ACM Conference on User Modeling, Adaptation and Personalization*, UMAP '19, page 22–30, New York, NY, USA, 2019. Association for Computing Machinery.

[192] Joshua R. Tyler and John C. Tang. When can i expect an email response? a study of rhythms in email usage. In *Proceedings of the 2003 European Conference on Computer-Supported Cooperative Work*, ECSCW 2003, pages 239–258. Springer Netherlands, 2003.

[193] Amy Voida, Wendy C Newstetter, and Elizabeth D Mynatt. When conventions collide: the tensions of instant messaging attributed. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 187–194, 2002.

[194] Daniel T Wagner, Andrew Rice, and Alastair R Beresford. Device analyzer: Understanding smartphone usage. In *International Conference on Mobile and Ubiquitous Systems: Computing, Networking, and Services*, pages 195–208. Springer, 2013.

[195] Tanja Walsh, Piia Nurkka, and Rod Walsh. Cultural differences in smartphone user experience evaluation. In *Proceedings of the 9th International Conference on Mobile and Ubiquitous Multimedia*, page 24. ACM, 2010.

[196] Joseph B. Walther and Judee K. Burgoon. Relational communication in computer-mediated interaction. *Human Communication Research*, 19(1):50–88, 1992.

[197] Haixun Wang, Wei Fan, Philip S Yu, and Jiawei Han. Mining concept-drifting data streams using ensemble classifiers. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 226–235. AcM, 2003.

[198] May Wang, Stella Cho, and Trey Denton. The impact of personalization and compatibility with past experience on e-banking usage. *International Journal of Bank Marketing*, 35(1):45–55, 2017.

[199] Steve Whittaker, Vaiva Kalnikaite, Victoria Hollis, and Andrew Guydish. 'don't waste my time' use of time information improves focus. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 1729–1738, 2016.

[200] Mikael Wiberg and Steve Whittaker. Managing availability: Supporting lightweight negotiations to handle interruptions. *ACM transactions on computer-human interaction (TOCHI)*, 12(4):356–387, 2005.

[201] Jason Wiese, Patrick Gage Kelley, Lorrie Faith Cranor, Laura Dabbish, Jason I. Hong, and John Zimmerman. Are you close with me? are you nearby? investigating social groups, closeness, and willingness to share. In *Proceedings of the 13th International Conference on Ubiquitous Computing*, UbiComp '11, page 197–206, New York, NY, USA, 2011. Association for Computing Machinery.

[202] Jennifer Wiley and Jeannine Bailey. *Effects of collaboration and argumentation on learning from web pages*. Lawrence Earlbaum, Mahwah, New Jersey, 2006.

[203] Jennifer Wiley and Cara Jolly. When two heads are better than one expert. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 25, 2003.

[204] Deanna Wilkes-Gibbs and Herbert H Clark. Coordinating beliefs in conversation. *Journal of memory and language*, 31(2):183–194, 1992.

[205] Michael Winikoff, Jocelyn Cranefield, Jane Li, Cathal Doyle, and Alexander Richter. The advent of digital productivity assistants: The case of microsoft myanalytics. In *HICSS*, pages 1–10, 2021.

[206] Ting-Wei Wu, Yu-Ling Chien, Hao-Ping Lee, and Yung-Ju Chang. Im receptivity and presentation-type preferences among users of a mobile app with automated receptivity-status adjustment. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2021.

[207] Qiang Xu, Jeffrey Erman, Alexandre Gerber, Zhuoqing Mao, Jeffrey Pang, and Shobha Venkataraman. Identifying diverse usage behaviors of smartphone apps. In *Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference*, pages 329–344. ACM, 2011.

[208] Michael Yeomans, Anuj Shah, Sendhil Mullainathan, and Jon Kleinberg. Making sense of recommendations. *Journal of Behavioral Decision Making*, 32(4):403–414, 2019.

[209] Neil Yorke-Smith, Shahin Saadati, Karen L Myers, and David N Morley. The design of a proactive personal agent for task management. *International Journal on Artificial Intelligence Tools*, 21(01):1250004, 2012.

[210] Fengpeng Yuan, Xianyi Gao, and Janne Lindqvist. How busy are you?: Predicting the interruptibility intensity of mobile users. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, pages 5346–5360, New York, NY, USA, 2017. ACM.

[211] Sihan Yuan, Birgit Brüggemeier, Stefan Hillmann, and Thilo Michael. User preference and categories for error responses in conversational user interfaces. In *Proceedings of the 2nd Conference on Conversational User Interfaces*, pages 1–8, 2020.

[212] R Yáñez Cortés. The problem of interpretation in psychology. *Acta psiquiatrica y psicologica de America latina*, 21(2):84—89, June 1975.

[213] Nima Zargham, Leon Reicherts, Michael Bonfert, Sarah Theres Voelkel, Johannes Schoening, Rainer Malaka, and Yvonne Rogers. Understanding circumstances for desirable proactive behaviour of voice assistants: The proactivity dilemma. In *Proceedings of the 4th Conference on Conversational User Interfaces*, CUI '22, New York, NY, USA, 2022. Association for Computing Machinery.

[214] Chen Zhao, Pamela Hinds, and Ge Gao. How and to whom people share: The role of culture in self-disclosure in online communities. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*, CSCW '12, page 67–76, New York, NY, USA, 2012. Association for Computing Machinery.

[215] Sha Zhao, Julian Ramos, Jianrong Tao, Ziwen Jiang, Shijian Li, Zhaohui Wu, Gang Pan, and Anind K Dey. Discovering different kinds of smartphone users through their application usage behaviors. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 498–509. ACM, 2016.

[216] Xujuan Zhou, Yue Xu, Yuefeng Li, Audun Josang, and Clive Cox. The state-of-the-art in personalized recommender systems for social networking. *Artificial Intelligence Review*, 37:119–132, 2012.

[217] Yun Zhou, Marta Piekarska, Alexander Raake, Tao Xu, Xiaojun Wu, and Bei Dong. Control yourself: on user control of privacy settings using personalization and privacy panel on smartphones. *Procedia Computer Science*, 109:100–107, 2017.