Design and Evaluation of Intelligent Agent Prototypes for Assistance with Focus and Productivity at Work

Ted Grover
University of California, Irvine
Irvine, California, USA

Kael Rowan
Microsoft Research
Redmond, Washington, USA

Jina Suh
Microsoft Research
Redmond, Washington, USA

Daniel McDuff
Microsoft Research
Redmond, Washington, USA

Mary Czerwinski
Microsoft Research
Redmond, Washington, USA

ABSTRACT

Current research on building intelligent agents for aiding with productivity and focus in the workplace is quite limited, despite the ubiquity of information workers across the globe. In our work, we present a productivity agent which helps users schedule and block out time on their calendar to focus on important tasks, monitor and intervene with distractions, and reflect on their daily mood and goals in a single, standalone application. We created two different prototype versions of our agent: a text-based (TB) agent with a similar UI to a standard chatbot, and a more emotionally expressive virtual agent (VA) that employs a video avatar and the ability to detect and respond appropriately to users' emotions. We evaluated these two agent prototypes against an existing product (control) condition through a three-week, within subjects study design with 40 participants, across different work roles in a large organization. We found that participants scheduled 134% more time with the TB prototype, and 110% more time with the VA prototype for focused tasks compared to the control condition. Users reported that they felt more satisfied and productive with the VA agent. However, the perception of anthropomorphism in the VA was polarized, with several participants suggesting that the human appearance was unnecessary. We discuss important insights from our work for the future design of conversational agents for productivity, wellbeing, and focus in the workplace.

CCS CONCEPTS

• **Human-centered computing → Human computer interaction (HCI)**.

KEYWORDS

conversational agents, personal digital assistants, workplace productivity, focus, sensing, affective computing

IUI '20, March 17–20, 2020, Cagliari, Italy

ACM ISBN 978-1-4503-7118-6/20/03…$15.00
https://doi.org/10.1145/3377325.3377507

ACM Reference Format:
Ted Grover, Kael Rowan, Jina Suh, Daniel McDuff, and Mary Czerwinski. 2020. Design and Evaluation of Intelligent Agent Prototypes for Assistance with Focus and Productivity at Work. In *25th International Conference on Intelligent User Interfaces (IUI '20), March 17–20, 2020, Cagliari, Italy*. ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3377325.3377507

1 INTRODUCTION

The productivity of information workers can be significantly influenced by information overload and stress. Due to the ubiquity of multiple devices, including desktops, laptops, phones, surfaces, smart watches and speakers, notifications, messages and other kinds of disruptions have become a serious problem for keeping focused on high priority tasks at work. A large body of work in human-computer interaction (HCI) research has concentrated on better understanding how workers manage their tasks, attempt to stay focused, and deal with distractions and interruptions throughout the day. In any given day, information workers are typically faced with multiple tasks that they need to complete, and often they devise unique strategies to remember and make note of these tasks [12, 15, 25]. A diary study of information workers found that workers had an average of 50 task-shifts over an entire work week [9]. Furthermore, information workers are usually faced with numerous interruptions throughout their day.

At work, people average 40 seconds on a computer screen before switching [23], and it can take several minutes for an office worker to return to their original task after interruptions. The deleterious effects of notifications, email and face-to-face interruptions has been very well documented [9, 20] in terms of lowered productivity at work. Different factors like the type and duration of the interruption, the complexity of the task prior to interruption, and even the exact moment of the interruption can have a negative effect on workers' ability to resume a task, their perceived productivity and their satisfaction with their performance[2, 5, 25, 30, 38]. For instance, past work from Iqbal and Horvitz [14] found that when information workers were interrupted by conversation, they were more likely to return to work on more peripheral tasks like email and web searches, rather than resume their previous task.

Research suggests that more distractions can lead to higher reported stress and lower productivity in the workplace [20–22]. While there have been many attempts to design software to assist with this problem in order to reduce stress and improve focus (e.g., Freedom, Windows FocusAssist [29], Tracktime [4], see [16] for a review), few of these products have yet to be widely adopted in the

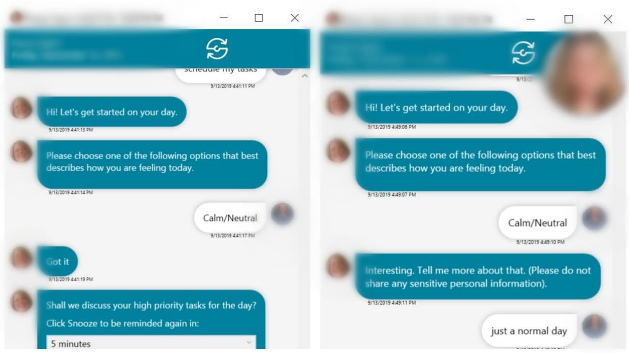

Figure 1: We present two productivity agent prototypes which help users schedule and block time on their calendar to focus on important tasks, monitor and intervene with distractions, and reflect on their daily mood and goals. Snapshots of the sample morning dialogue interactions for the text-based (left) and virtual agent (right) prototypes are shown. The face images are blurred to respect privacy.

workplace. Therefore, people continue to design their own methods and workarounds, but report having trouble nonetheless [16].

These findings suggests that maintaining focus in the face of constant shifting priorities and interruptions in the workplace is a complex and important problem. Yet, to our knowledge, only one system, from Kimani et al. [16], has yet been designed that helps knowledge workers with prioritizing their work, providing reminders on when to switch tasks or get back on task, to take breaks, as well as to reflect on how much they accomplished at the end of the day. While this effort was notable for its complex system design, the user interface was quite simple in its design, similar to a standard chatbot. In our work, we design two different agent prototypes which build upon this past work from Kimani et al. [16] (described further in the related work). Our paper offers three main contributions: 1) The design of two different conversational agents, one text-based (TB), similar to a chatbot, and one virtual, embodied, conversational agent that responds to the user's emotion (VA), 2) a longitudinal evaluation of these two agents against a shipping product that tries to help users schedule focus time in Microsoft Outlook, and 3) actionable insights from qualitative analysis of user feedback on agent design around anthropomorphism, user task scheduling, and the need for better back and forth negotiation and control with the user.

2 RELATED WORK

For task and time management, various different applications have been developed, most notably, MeTime [35], RADAR [10], TaskBot [34], and Calendar.help [8] that each aim to assist users be more efficient in managing their work through different approaches. MeTime aims to provide real-time awareness of how users are spending their time through graphic visualization of their application usage in order to promote efficiency of time use achieving their task goals [35]. The authors showed that exposure to meTime decreased time spent on distractions (e.g., social media), and increased their feelings of productivity [35]. RADAR is an AI system that was designed to help reduce email overload by automatically producing tasks and to-do lists directly from users incoming emails [10]. Through an experimental setup, users that were aided by RADAR that were confronted with overload performed better at completing email intitiated tasks than without RADAR. [10] TaskBot, a chatbot agent was designed to mediate the creation and management of tasks within project teams [34]. Users found TaskBot useful for naturally transferring conversations into actionable tasks, but found that the system struggled when having to deal with multi-threaded conversations [34]. Finally, Calendar.help used an AI interface integrated with the user's email to automatically schedule user's meetings, and allowed the user or another human assistant to make adjustments manually when needed [8].

Applications developed within this domain have also aimed at helping block distractions to help users focus. One such example is Microsoft Focus Assist, which allows users to define "whitelisted" work related sites and applications and "blacklisted" non-work related sites and applications [29]. Focus Assist also then blocks access to non-work related sites and applications for a period of time determined by the user [29].

Finally, other applications have aimed at promoting workers to reflect on their emotions and goals throughout the workday, as well as to promote saving time for healthy and appropriate breaks. For instance, Robata was developed to be voice mediated agent that helped its users plan and organize their tasks, reflect on their motivation and satisfaction throughout the day, and also promote and reflect on their self-learning [18]. Users of Robata reported generally appreciating a way of reflecting on their planning and goal setting [18]. BreakSense is another application that was developed to help promote mobility for its users during their breaks in order to encourage healthy activity and lifestyle in office environments that can be often very sedentary [7].

Although each of these applications attends to a specific problem area within the broader domain of promoting workplace productivity and well-being, there has been little research on the development and evaluation of AI systems that incorporate task scheduling and management, distraction blocking and monitoring, and mood and goal reflection in a single, standalone application. To our knowledge, the only application that has attempted to integrate all these different functionalities into a single system is the work of Kimani et al. [16]. In their work, they presented *Amber*, a conversational desktop assistant whose purpose was to help workers in four main areas: (1) scheduling high priority tasks, (2) aiding workers in transitioning from one task to the next, (3) avoiding and intervening with distractions, and (4) reflecting on their work through a conversational AI interface [16]. In our work, we present our productivity agent, which builds upon the infrastructure and capabilities of this previous agent developed by Kimani et al. [16]. Similar to the work from [16], we gave our agent a female gender.

Our work extends this previous work in a number of ways. First, as Kimani et al. [16] found in their analysis that users desired an agent that took more control over the flow of interaction, we designed our agents to provide a series of dialogues that started at the beginning of each day with scheduling their tasks, helping users progress through these tasks until the end of the day. These dialogues were initiated automatically when appropriate without any extra user input. Second, in addition to incorporating the four major functionalities from Kimani et al. [16], we also wanted to investigate the impact that a more human-like agent with more emotional intelligence would have on user perceptions, as well as their focus and productivity. In past studies of virtual agents, research has found that emotionally appropriate responses from an agent contribute to a more positive and satisfying experience during interaction [11, 36]. Furthermore, emotional intelligence in agents has been shown to help alleviate negative emotions, like frustration [17]. In a study of the effectiveness of different agent characteristics in an organizational setting, anthropomorphic appearance in agents was also shown to increase users perceptions of the agents usefulness [31]. As little research has investigated the effect of more human-like agents in the context of workplace focus and productivity, a primary goal of our study was to evaluate and compare its effectiveness to a typical chatbot agent. Therefore, in this work, we build two different prototypes of our agent: a text-based conversational interface prototype (TB) that employs a similar UI to the previous version developed by Kimani et al. [16], and a prototype virtual agent conversational interface (VA) version with a video avatar that speaks to the user. The VA prototype incorporates the ability to detect user emotions through video input and adapt its responses to be appropriate and congruent with the users emotional state. We conduct a three week long within-subjects user study to evaluate and compare the VA prototype, the TB prototype, and a simple, non-conversational task scheduling tool integrated into users' email, similar in part to Calendar.help [8], as a control. To our knowledge, our work presents the first multi-week user study evaluating different intelligent agent prototypes that aim to increase focus and productivity at work.

3 RESEARCH QUESTIONS

Given the goals of our research, we constructed three primary guiding research questions to evaluate the effectiveness of our agent functionalities and compare each of our agent prototypes.

RQ1: *Are users more productive and less distracted during the time periods they scheduled for focused work through the agents?*

A primary functionality of both our agent prototypes is to take control of scheduling the users' high priority tasks in their calendar. The agents also help the user transition through these tasks, and intervene during the task period if the agent detects the user is distracted (described in more detail in the next section). We evaluate the effectiveness of this task scheduling and monitoring system by measuring the degree to which users appeared to be more productive and less distracted during their agent-scheduled task time periods compared to their their normal baselines outside of these periods.

RQ2: *With which agent prototype are users most productive, least distracted, and most satisfied?*

Based on the design differences between the text-based agent and the virtual agent prototypes we created, we aim to understand any differences in user experience between the agents in terms of productivity, avoiding distractions, and overall reported user satisfaction. We do so by examining a variety of data sources, including users' logged behaviors while interacting with both agent prototypes and survey responses.

RQ3: *What features of the agent prototypes are most useful, and are there design improvement opportunities?*

Given the novelty of our study, we are also interested in better understanding user perceptions of the agent in terms of the features that were most useful, and where improvements can be made on future iterations of our prototypes. To do so, we perform qualitative coding on the free response questions on both weekly and study exit surveys to identify emergent themes around our users subjective experiences when interacting with the agent prototypes.

More details on the research procedures are provided in the study design section of the paper.

4 FOCUS AGENT: CONVERSATIONAL PRODUCTIVITY ASSISTANT PROTOTYPES

In our work, we build and evaluate both a new TB conversational version and a VA conversational version of our productivity agent. First, we provide an overview of the basic functionality shared across both versions. Next, we outline the UI differences between the TB and VA agent versions. Finally, we provide an overview of the system design and architecture of both agent versions.

4.1 Daily User Workflow

Both versions of the productivity agent that we developed operated as applications that ran on Windows 10 operating systems. When not active on the user's screen, the agent appeared as an icon in the system tray of the desktop. Right-clicking on this icon allowed the user to exit out of the application, reconnect or reset the agent, or view help documents on how to use and interact with the agent. When the agent became active and notified the user of something, the application would pop-up to the foreground of the user's screen (usually as a panel to the right-hand side of their screen), and a notification would appear above the system tray in the lower-right corner of the screen.

Similar to the work from Kimani et al. [16], the purpose of the agents we developed was to design an intelligent assistant that could support the user's task-related goals and help them better focus on their high priority tasks, which would hopefully promote better productivity and emotional well-being at work. In their work, Kimani et al. [16] conducted an online survey of 70 technology information workers and literature review to better understand the design requirements for a conversational productivity assistant (see [16] for a full description of their requirements analysis). Through their research they determined that task scheduling, task switching and reminders, task re-attachment, and dealing with distractions were the most important skills to be incorporated in a productivity assistant. Given the similar research context of this past work to our own, we also focused on and incorporated these abilities into both our agents through a series of dialogue models that would be initiated throughout the user's work day. The dialogue models we created shared across both the TB and VA versions of our agent are described below:

First Time dialogue: When users first installed the agent, the application window would appear in the foreground of the screen. The agent would introduce herself, her role and capabilities.

Morning dialogue: When the user unlocks their computer for the first time each day, the agent initiates a conversation with the user. She first asks how the user is feeling wherein the user would be given six different options in a drop-down menu to choose from (*Happy, Sad, Stressed, Calm/Neutral, Focused, or Frustrated*). The agent then asks the user if they would like to schedule high priority tasks on their calendar for them. Users again are provided drop-down options of either 'Yes', 'No', or to remind them again in 5, 10, or 15 minutes. If the user chooses to schedule tasks, they are then asked what time they plan to head home for the day. Next, the agent prompts the user to enter their desired tasks in priority order and provide their estimated duration in minutes. Upon submission of these tasks, the agent attempts to schedule the users' desired tasks on their Outlook calendar between the current time and 30 minutes before the time they reported they planned to leave. The agent was programmed to have a preference for scheduling tasks with a 15 minute gap between tasks. If this was not possible, the agent would try a 10 minute gap, then a 5 minute gap, and, finally, no time gap between tasks. If necessary, the agent would also split tasks into separate chunks on their calendar. If the agent could not find any time on the user's calendar to schedule, she prompted the user to examine their schedule to see if they could make room for their tasks. If the user chose to make changes, the agent would attempt to schedule their desired tasks again. Upon successful scheduling of their tasks, the user was shown the time slots when their tasks were scheduled, and notified that their tasks were successfully scheduled on their calendar. Figure 1 shows part of an example user interaction for the morning dialogue for both agent prototype versions.

Task Ramp-up dialogue: Three minutes before a task that was scheduled through the agent began, the application would appear in the forefront of the user's screen, and the agent would inform the user of their next task (*"Your scheduled focus time for one of your high priority tasks is about to begin. Are you ready to switch to it?"*). As abrupt task-switching can be disruptive [2, 25, 30], this dialogue was designed to help users transition smoothly between tasks. Users would either select drop-down menu 'Yes' and transition to their task, have the agent remind them in 5, 10, or 15 minutes, or be given the option to reschedule the task given their next availability.

Task Ramp-down dialogue: Five minutes before a task that was scheduled through the agent ended, the application would appear in the forefront of the users screen and notify them that their scheduled focus time was about to end (*"Your scheduled focus time for this task ends in 5 minutes. Now might be a good time to wrap up for a smooth transition..."*). Again, this dialogue model was created to facilitate smooth transitions in and out of the users task goals. Next, either right away or five minutes later if they chose to let the agent remind them at the end of their task, the user was prompted with a question asking them how productive they felt during their scheduled task on a 5-point scale (*Not at all, Slightly, Moderately, Very, or Extremely*).

Distractions and Breaks dialogue: We also created a dialogue model that would be triggered when the agent determined that the user was supposed to be in the middle of an agent scheduled task, but the application detected that the user was distracted. We integrated a sensing application [24], described in further detail at the end of this section, into both our agent systems to enable the agent to monitor users' windowing activity to initiate the distraction dialogue when appropriate. Overuse of certain internet applications and sites at work, particularly social media sites, have been shown to be associated with lower work efficiency and day-end productivity [6, 19]. We therefore decided to incorporate social media sites (e.g., Facebook, Twitter, etc.) monitoring into our agent functionality, in addition to other types of internet sites and applications that were likely to be distractions if visited too frequently, including shopping (e.g., Amazon), news (e.g., New York Times, CNN), and music streaming sites (e.g., Spotify, Soundcloud). In total, we created a list of 45 unique site and application names to monitor for prolonged use. In the back-end of our agent, we kept a log of the windows a user visited and the time each window was in the foreground over the past five minutes, updated every second. If 50 percent or more of this window activity within this log was from one or more of the sites or applications that we monitored for, the distraction dialogue was initiated (*"It looks like you may be taking a break. Would you like me to set a timer and remind you to get back to your tasks after a short break?"*). Users would then be provided with a set of drop-down options to either set a timer for 5, 10, or 15 minutes, inform the agent that they are not taking a break, that they will get back to their task, or to 'let me be' (where the agent would not interrupt them again for the rest of their task).

We also designed our agents to encourage short breaks after periods of extended focus, as taking short mental breaks throughout the day has been shown to be beneficial for worker well-being and long-term productivity [18]. Our sensing application (described later in this section) also incorporated the ability to detect and classify the user's emotional state through video input from a webcam into four distinct categories (*Happy, Focused, Frustrated, or Other* (the default emotional state)). If the agent detected that the user had been in a *Focused* state continuously for the past hour, the user was prompted with a suggestion to take a short break.

End of Day dialogue: As research suggests that workers are better able to detach from work at day's end when they are able make future plans to finish any uncompleted goals for that day [33], we built a dialogue model to allow users to reflect on their day and schedule any unfinished tasks for the next day. 30 minutes before their reported departure time, the agent prompts the user, asking them to reflect upon their day (*"Hi again! Before you leave work, I would like to ask you to reflect on your day. Overall, how would you rate your day?"*), where the user would be given five options (*Very poor, Poor, Acceptable, Good, or Very good*). The user is then prompted to check off which tasks from the morning they did or did not complete, and if there are uncompleted tasks, the agent asks the user if they would like to save the tasks they did not complete for the next day, before wishing the user a good evening.

4.2 Differences Between Text-Based and Virtual Agent Versions

Although both our prototypes (TB and VA versions) had the same basic functionality and structure, the VA prototype included extra features that differed in the visual UI experience of the agent and in its emotional intelligence.

First, the VA prototype incorporated a video avatar of the agent speaking to the user in addition to the text output from the agent.

The words she would speak matched the text that was produced, providing more context in terms of emotional expressiveness and tone that is sometimes lost via text communication alone. To create the video clips used for the VA, we had an actor rehearse and film all 109 statements that the VA version would produce.

Second, the VA prototype was designed to be much more emotionally expressive and intelligent, in both text and video output. Although in many cases, the raw text between statements was almost identical, the emotional expression of these statements differed by the tone of voice, the use of contextual emojis, and the adaptation of responses based on sensed emotional state. In particular, we sought to create an agent personality that was encouraging, kind, and cheerful. The VA employed emojis quite frequently, while the TB never employed emojis. As Emoji use in text has been shown to strengthen the perceived affect of a message (either in a positive or negative direction) compared to the same text without accompanying emojis [32], we saw incorporating emojis into the VA prototype as an easy first step towards introducing more emotional expressiveness. For instance, we incorporated simple emojis into confirmations of reminders (*"Got it!* 😊 *Reminding in 5 minutes."*). In addition, in the VA prototype, we aimed to incorporate more emotional intelligence by having the agent adapt its responses to the current emotional state of the user, detected by our sensing software (*Happy, Focused, Frustrated, or Other*). For all agent messages in the Distraction dialogue, the emotional valence of the agent's text and corresponding video responses were designed to be congruent with the user's current emotional state. As an example, Table 1 provides an overview of the possible first utterances that the VA agent could produce depending on the users emotional state for distraction dialogue. Similarly, the emotional tone of the first message the VA agent produces in Ramp-down and End of Day dialogues was also adjusted to be concordant with the user's current emotional state (e.g., in the Ramp-down dialogue, if the user is in a *Happy* state: *"Woohoo!* 😄*. Your scheduled focus time for this task ends in 5 minutes..."*) In the TB version, the agent dialogue was always constant and never changed, regardless of the user's current emotional state.

Recent research also suggests that the psychological benefits of disclosure and reflection with an agent are similar to reflection with another human [13]. Therefore, we considered the ability for users to reflect on their feelings and sense of productivity to be a beneficial final extra feature in the VA prototype. After users reported how they were feeling during the morning dialogue, they were asked to reflect upon their feelings in an open-ended response. Similarly, in the Ramp-down and End of Day dialogues, after reporting their productivity or their feelings on how their day went (respectively, depending on the current dialogue) they were also asked to reflect on their response. If the user selected a negative option from the drop-down prompt (e.g., *Sad, Stressed, or Frustrated* in the morning feelings form, or *Very poor or Poor* in the end of day evaluation form), they were prompted with a sympathetic response to reflect more (*"I'm sorry. Thanks for sharing. Could you please tell me more about that?"*). If the user selected a positive option (e.g., *Happy or Focused* in the morning feelings form, or *Very good or Good* in the end of day evaluation form), they were prompted with a congratulatory response (*"That's great to hear! Can you tell me more about that?"*). If the user selected a neutral option, they would be prompted with a more neutral response. Figure 1 provides a side-by-side comparison of the TB and VA versions for a portion of the Morning dialogue for a sample user, wherein the VA prompts the user to reflect on their reported feeling choice.

4.3 System Design and Infrastructure

The system behind both the TB and VA prototypes is comprised of four main components: a sensing application, the agent client, a cloud-hosted back end that employs Microsoft Conversation Learner bot framework, and the user's Microsoft Outlook calendar. Figure 2 provides an overview of the system architecture used by both our agent prototypes.

Sensing Software: We leveraged a desktop sensing framework application that uses audio and video input (from a speaker and webcam) to continually detect the user's expressed emotional state (Figure 2A) [24]. The application uses a trained machine learning model to detect changes in the user's facial expression with high temporal resolution (multiple times per second). The sensing software was used to detect multiple different data streams from the users simultaneously, including the number of faces detected in the webcam view, and measures of the user's facial expression in different categories: *neutral, anger, contempt, surprise, disgust, sadness, and fear* using convolutional neural networks trained on hand-labeled images [3]. Facial expression measures varied between zero to one in one or more of these emotion categories. The software also logged user window activity including open window actions from the keyboard or mouse, hashed window titles and application names, and the current foreground window or application. The live streaming of information from the sensing software was used to help trigger appropriate dialogue models in the agent client application, which we describe next.

Agent Client Application: For both our prototypes (TB and VA), participants interacted with the agent through a standalone

Table 1: Agent utterances based on users' emotional state for first message of virtual agent (VA) distraction dialogue.

User is **Happy**	User is **Focused**	User is **Frustrated**	User is **Other** (default)
Hey! 😄 *It looks like you're really enjoying your break. Would you like me to set a timer to remind you to get back to your tasks after a short break?* 😉	*Sorry to interrupt.* 😳 *It looks like you needed a break. Would you like me to set a timer to remind you to get back to your tasks after a short break?* 😉	*Sorry to interrupt.* 😟 *It looks like you need a break. Would you like me to set a timer to remind you to get back to your tasks after a short break?* 😊	*Hi.* 😊 *it looks like you needed a break. Would you like me to set a timer to remind you to get back to your tasks after a short break?* 😊

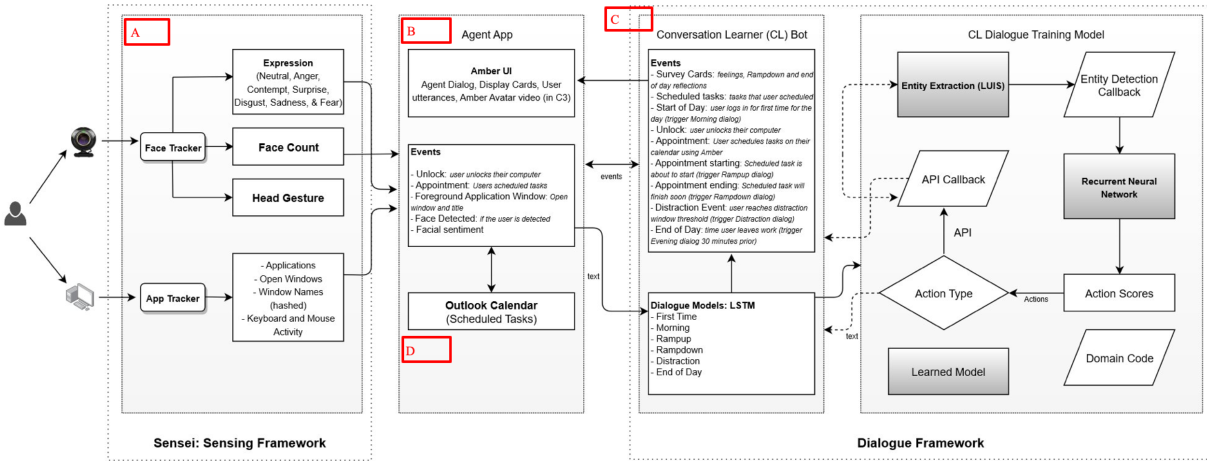

Figure 2: The Agent System Architecture. It contains four main components: A) the sensing framework collecting facial data, B) the client application that uses information gathered from sensing software to generate events used to initiate the appropriate dialogue, C) the dialogue framework, that uses client application events to determine and control the flow of conversation, D) the Outlook calendar.

desktop application (Figure 2B). The application interacts with the sensing software, the user's outlook calendar, and the backend system that provides dialogue models. The first time the user unlocked their computer for the day, the application prompted with the morning dialog; for all other unlocks, the application showed the user the message *"Welcome back!"*. When a user schedules tasks with the application at the start of the day, this list of tasks is sent to the Outlook calendar integration for scheduling (Figure 2D). Every minute, the calendar application provides an update of scheduled activities to the application which in turn sent appointment events to the backend system, that determined when to inform the user when it is time to switch to a different task. Similarly, when the user reports their approximate heading home time in the morning dialogue, this information is propagated to the backend system to initiate the End of Day dialogue at the appropriate time. If the agent did not parse the user's the end of day time correctly, the End of Day dialogue would initiate at 4pm (default value). Windowing activity from our sensing software would also be sent to the application in order to monitor for the user visiting distraction sites. For the VA version application, information gathered from our sensing software is also used by the application to continuously update the user's emotional state, and this information is then propagated to the backend system to initiate appropriate dialogue models. The logic was derived from past research analyzing the association between users' reported emotional states and their captured facial sentiment using the sensing software at time of their report [24].

Dialogue Framework: The client application communicates with the backend system that manages and trains the dialog, using the Microsoft Bot Framework (Figure 2C) [26]. This framework allows the client application to send and receive text and events from the backend dialogue system. This backend system employs Microsoft Conversation Learner framework [28] which allows us to build the conversation flow in the dialogue models using sample training interactions. The Conversation Learner framework employs an end-to-end recurrent neural network to train dialogue models that learn sample dialogues that involve multiple turns between the user and agent. Training dialogue models is done in two steps: entity extraction and action selection. To extract entities (e.g., labeling user response as a feeling or productivity reflection) conversation learner uses Microsoft's Language Understanding and Intent Service (LUIS) [27] that uses the text input's content and context in the dialogue flow to label appropriate entities. The action selection model also uses a recurrent neural network to take the sample training dialogue and current entities as inputs to determine the most appropriate response. Using the Conversation Learner we built and trained the six different dialogue models described in section 3.1. Functions built within the agent client application communicated with the dialogue framework to decide when each dialogue model should be initiated and terminated.

5 USER STUDY

In order to effectively evaluate the TB and VA prototypes, we conducted a 3 week, within-subjects user study with 45 information workers across a variety of work roles in a large organization. The study was approved by authors' internal ethics review board. In the following section we describe the study design, participant pool, survey measures, and our guiding research questions for our study.

5.1 Study Design

Our within-subjects study design had three different conditions:

- **Condition 1 (C1)**: This was our control condition, which was a product integrated into users' Microsoft Outlook accounts. Clicking on the product icon allowed users to schedule blocks of time for 'Focused' work on their calendar at the beginning of each day. However, this was the product's only functionality. It did not have any conversational or agent-like properties, and did not include any other functionalities shared by our two prototype agents.
- **Condition 2 (C2)**: The text-based (TB) agent prototype.
- **Condition 3 (C3)**: The virtual agent (VA) agent prototype.

The study began on a Friday afternoon, when we asked participants to install the sensing software on their primary computers, and provided them with instructions on how to activate the C1 product into their Outlook accounts. We asked participants to run the sensing software continuously on their computer in order to collect behavioral data from study start to finish. All participants began by spending the first week of the study (starting on the following Monday) in C1, using the Microsoft Outlook product. C1 acted as a control condition for all participants, in order to have a baseline productivity tool for all users to compare and evaluate the agent prototypes against. On Friday afternoon at the end of the first week, participants were given instructions to disable the Outlook widget and install one of the two prototype agents. For the second and third weeks, participants were counterbalanced such that 50% of the participants started the second week in C2 and 50% started in C3. On Friday afternoon at the end of the second week, participants were again given instructions to uninstall their current agent prototype and install the prototype version they had not yet interacted with.

Upon receiving instructions via email for the next week of the study on the preceding Friday afternoon, participants were given until the following Monday at noon to complete the instructions for the next phase of the study. Although most participants adhered to this timeline, some participants were a bit delayed with installing the software due to technical problems that we had to trouble shoot. However, for these participants, we staggered their study end dates such that each participant had at least 4.5 full days in each condition.

5.2 Participants

We recruited participants through an email advertisement at a large technology organization. In order to aim for gender parity in our participant pool, we pre-sampled the organization employee database to obtain a roughly equal number of male and female employees to include in our email distribution list. Interested participants completed a demographics survey that asked them about their work role, work space setup, and the typical work schedule. We filtered interested participants to recruit only those that worked out of the organizational campus where our study was being ran, and those that reported spending less than 25 hours a week in meetings, as we wanted participants to have a significant amount of interaction time in each condition. In total, we initially recruited 50 participants, however, over the course of the study, five of these participants either dropped out of the study or left the Organization prior to the the study concluding. Furthermore, another five participants reported severe technical difficulties with one or more of the agent systems not being able to integrate properly with their Outlook calendar. Therefore, we omitted these participants from our analysis, for a final participant pool of 40. This final pool was 58% male and 42% female, and the average age was 33 years. 53% worked in Engineering and Development roles, 20% worked in Research or Design roles, 10% worked in Administrative Assistant or Human Resources roles, 10% worked in Sales, Strategy, or Marketing roles, and 7% worked in Program Management roles. 65% of participants used a desktop computer with an external webcam for the study, while the remainder used a laptop with a built in webcam.

5.3 Survey Measures

Over the course of the three week user study, we administered a series of surveys to our participants:

Intake Survey: Prior to the beginning of the study, users were asked to complete a survey containing demographic questions.

Weekly End-of-Interaction Surveys: At the end of each week of interaction with the agent, in weeks two and three of the study, users were asked to complete two short surveys to rate their experience with the agent for that week. The first asked 15 Likert (5-point) scale questions on their rating of the agent (e.g., trust, satisfaction, smoothness of conversation). The second survey contained 15 Likert (5-point) scale questions asking the users about their interaction experience (e.g. *"I enjoyed scheduling tasks with the agent"*). Participants were also given the opportunity to provide their thoughts on their experience with each agent through an open-ended response question. The content and questions in the surveys were the same for both C2 and C3.

Exit Survey: One week after the study concluded, users were asked to reflect on their overall experience partcipating in the study: (*"Please describe your overall experience participating in this study, including any ideas about the future of conversational agents, focus time blocking, and the nature of tasks and interruptions:"*). Last, participants were also asked four multiple choice questions asking them to judge each of the three conditions (with an option for 'None of the above'): *"With which of the following systems did you feel the most productive?", "With which of the following systems did you feel was the most helpful in avoiding web-based distractions?", "With which of the following systems are you most satisfied with?"*, and *"With which of the following systems would you continue using if given the opportunity?"*.

5.4 Research Procedures

We used a combination of behavioral logs from our sensing software and survey responses to address our research questions:

RQ1: To determine if users were more productive and less distracted during the time periods they scheduled tasks with the agents (which we refer to as **focused tasks**), we use windowing activity data collected from our sensing software to determine the proportion of time where productivity and distraction applications (apps) and websites (sites) were the active foreground window during users focused tasks, and in time periods outside of focused tasks (which we refer to as **outside focused tasks**). In addition, we also used the face detection data from our sensing software to determine the proportion of time that the user was detected to be present at their workspace during their focused tasks and outside focused tasks.

For our analysis of distraction, we used the collection of 45 distraction apps and site names described earlier. For productivity, we composed a list of all apps that were accessed over the course of the study, and coded this list for apps that we deemed to be very likely to be mostly used for productivity purposes (e.g., Microsoft Word, Integrated Development Environments (IDE)). As there was a wide variety of different apps that were used across work roles, we constructed a list of 225 different productivity apps and sites. For both our distraction and productivity analyses, we took a conservative approach, and only included apps and sites that we were quite certain were distractions or used for productivity.

For each condition, and for productivity and distraction measures respectively, we use the Wilcoxon Rank Sum test to determine if the means of the proportion of time spent between focused tasks and outside focused tasks were significantly different. Similarly, for the face detection data, we again used the Wilcoxon Rank Sum test to determine if the means of the proportion of time the user's face was detected at their work station between focused tasks and outside focused tasks, were significantly different. We used the Wilcoxon Rank sum test as we cannot assume that the distribution of the time proportions will be normally distributed.

RQ2: To determine if there were significant differences in the proportion of face detections, productivity, and distraction between the three conditions, we applied the Kruskal-Wallis H test, for each proportion metric described above, with the condition as the independent variable. We also investigated if there were significant differences in the total amount of focused time users scheduled with each of the three conditions, using the same test.

In addition to analyzing log data from our sensing software, and to better understand any differences in the user experience between C2 and C3, we investigated significant differences using t-tests between C2 and C3 for the weekly end-of-interaction surveys, and examined the results of the reflective multiple choice questions from the exit survey.

RQ3:Finally, to better understand the subjective aspects of users experience interacting with the agents, we performed qualitative coding of the free response prompts across all surveys and extract key themes from the participants' experiences in the study.

6 RESULTS

6.1 RQ1

Figure 3 presents boxplots of the distributions for the proportion of time spent using productivity apps or sites, distraction apps or sites, and proportion of time a face was detected at the user's work station (from left to right) for both focused tasks and outside focused tasks, for each condition. Table 2 provides the results of the Wilcoxon Rank Sum test applied to focused tasks and all time periods outside of focused tasks, for each condition and each metric. We see that for productivity apps and site usage, participants spent more time on productivity apps and sites during focused tasks than outside of focus tasks in C1 (V=443, p <.01), C2 (V=618, p<.01), and C3 (V=668*, p<.01). Looking at the proportion of all time spent in productivity sites or apps across all conditions, shows that participants spent more time in productivity apps and sites during focused tasks compared to outside of focused tasks (V=1043, p<.01). However, for distraction site usage, there were no significant differences in time spent on distraction sites during focused tasks compared to outside focused tasks. For face detections, there were no significant differences in proportion of time the users face was detected during focused tasks compared to outside focused tasks for C1 and C2, however for C3 users had more face detections during focused tasks (V=537, p=.024). Looking at the proportion of time faces were detected across all conditions shows that participants had more face detections during focused tasks compared to outside of focused tasks (V=825, p=.032). Across all conditions participants spent 69.6% more of their time on productivity apps or sites, and 19.8% more of their time at their workspace during focused time compared to outside focused time (i.e. their usual baselines).

6.2 RQ2

We next perform a Kruskal-Wallis test using the proportion of time during focused tasks as the dependent variable, and the condition as the independent variable for each metric. We find that the effect of condition on proportion of time in productivity apps or sites (χ^2=5.87, df=2, p=.07), proportion of time in distraction sites (χ^2=1.79, df=2, p=.41), and proportion of time with their face detected at their workstation (χ^2=.63, df=2, p=.73) during focused tasks to be non-significant. However, for the same test applied to the total focused time scheduled in each condition, the effect of condition is significant (χ^2=11.38, df=2, p<.01). Pairwise wilcoxon rank sum tests between the conditions (with bonferroni correction) show that both users scheduled more focus time with C2 (p<.01) and C3 (p<.01) than C1, although there was no siginficant difference between C2 and C3 (p =.93). Figure 5 presents boxplots of the total focused time scheduled across each condition.

Next, we analyze differences in the agent ratings between C2 and C3 from the weekly surveys. Participants rated the statement *'My conversation with the agent was natural.'* higher in C2 than in C3 (t=2.36, df=40, p=.024), and rated the statement *'I like scheduling tasks with the agent.'* higher in C3 than in C2. There were no other significant differences in survey items between C2 and C3.

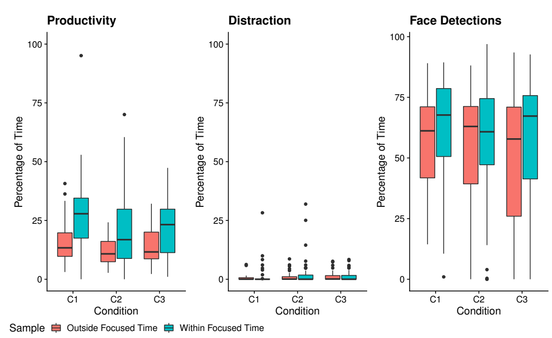

Figure 3: Boxplots for perecentage of time spent in productivity apps or sites (left), distraction apps or sites (middle), and proportion of time with face detected at work station (right) for both within focused task (blue), and outside focused task (red) periods for each condition.

Table 2: Results (V-statistic and p-values) of Wilcoxon Rank Sum test applied to Productivity, Distraction, and Face Detection metrics during focused time and outside of focused time for each Condition. '↑' indicates that the population mean rank of proportion of time spent is higher in focused tasks than outside of focused tasks. * = p < .05, ** = p < .01

	C1	**C2**	**C3**	**All**
Productivity	443^{**} ↑	618^{**} ↑	668^{**} ↑	1043^{**} ↑
Distraction	120	218	131	427
Face Detections	321	521	537^{*} ↑	825^{*} ↑

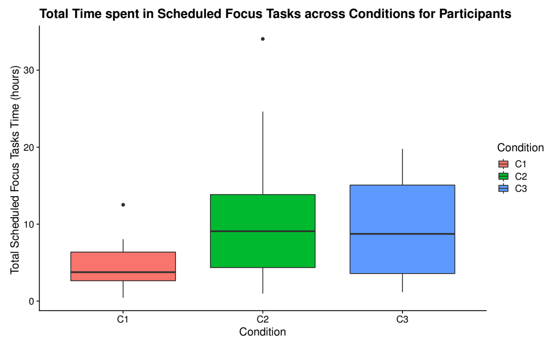

Figure 4: Boxplots for the total focused time scheduled for each condition.

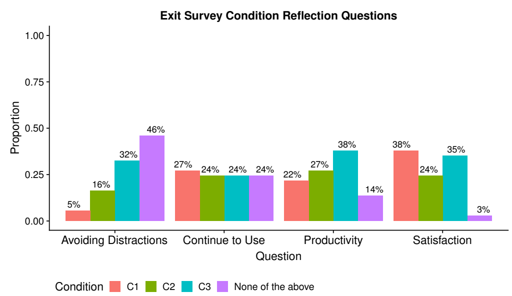

Figure 5: Percentage of responses for each condition for exit survey multiple choice questions. Participants were given descriptions of each of the conditions to choose from, for brevity we refer to the options as C1, C2, and C3.

Figure 6 shows the results of the multiple choice questions asked in our exit survey. 46% of participants reported that none of the conditions were most helpful for avoiding distractions, however C3 was best compared to rest of the conditions, with 32% picking this option. Participants reported a quite even split with respect to which system or agent they would like to continue using (between 24% and 27% across each of the four options). However, 38% of participants reported that they felt most productive in C3, the highest percentage of any of the options. Finally, participants reported being most satisfied with C1 (38%), however a close second was C3 with 35% of participants reporting being most satisfied with the VA prototype.

6.3 RQ3

Finally, through qualitative coding on the free response questions from the weekly and exit surveys we found three primary emergent themes that arose most from our analysis, which we present here:

Polarized perceptions of anthropomorphism in C3: Some participants appreciated the anthropomorphic design of the VA prototype (C3), and thought it provided a more 'natural' experience: *"P24: The one major change in this version was the add on of a question asking me to say more about my productivity or day...it made the conversation feel more natural than before."*. However, other participants thought that the anthropomorphism was unnecessary for the context the agent was designed for, and set the wrong expectations: *"P1: found the anthropomorphization of the agent unnecessary and setting the wrong expectations. P8: She asked me about my emotions which I didn't enjoy...I would have rather focus on my tasks then reflect on my feelings."*.

Desire for more intelligent automatic task scheduling: For both the TB and VA prototypes (C2 and C3), users repeatedly expressed a desire for the agents to schedule tasks in a more intelligent and flexible manner: *"P15: when I had conflicts that came up I could not easily move things with the agent's help. P12: she can't really deal with having to switch priorities on the fly."*. Users also repeatedly reported that the agent would schedule their tasks over lunch breaks if they didn't have a placeholder for lunch on their calendar. Other participants suggested that the task scheduling logic should be personalized to the user: *"P21: The agent should understand (or ask) more about the properties of the user i.e., when they would like to do the stressful work, how often they like to take break etc."*.

Desire for better distraction monitoring: Some participants reported that the distraction logic didn't always initiate at the appropriate times. *"P15: For me, I never really had [agent] engaged when I was distracted. P21: Breaks that I set for myself (using a variation of the Pomodoro technique) were detected by the agent, which I found intrusive."*.

7 SUMMARY OF RESULTS AND DISCUSSION

In sum, we found that agent assisted scheduling of focused time on users calendars appeared to be quite beneficial, as we observed that in focused time users spent approximately 70% more of their time on productivity sites, and 20% more of their time present at their work station compared to their normal working patterns outside of these time blocks (RQ1). We see this functionality in conjunction with helping users transition smoothly through these tasks, and avoid distractions during their tasks to be crucial components to include for future iterations of intelligent agents in this domain. After the study concluded (and users had completed the exit survey), we provided participants the option to continue using either agent prototype. After three weeks two participants continued to use the agent daily (one with the VA and one with the TB prototype). We see this adoption of the agents as encouraging, as these agents are first iterations of research prototypes. However, from our analysis it is also clear that there remains significant room for improvement and refining these functionalities going forward.

Although our results were not wholly conclusive for RQ2, in our exit survey we saw that users *felt* they were more productive and less distracted when using the VA prototype compared to the TB prototype or the control product. From our qualitative analysis in RQ3, it became clear that users thought the general idea behind our agents was beneficial and that we were aiming towards addressing a very important need that many information workers face. However the users expected more from the agents. From our perspective, this is encouraging as it provides motivation to build upon and refine the underlying agent framework we have. Given the qualitative results of our user study, we provide three major insights for the design and development of future intelligent agents for assisting with focus, wellbeing, and productivity at work:

Congruence between agent anthropomorphism and abilities: Past work on intelligent agents for personal well-being at work has found that incorporating anthropomorphism and emotional intelligence into the agents makes the agents more likeable and effective at fostering self-reflection [11, 18]. However, in our work, where the primary purpose of our agent was to assist with productivity and focus at work, the perceptions of the human-like appearance of the VA prototype was polarized. Many participants saw the human-like appearance of the VA prototype as setting the wrong expectations in terms of its capabilities, and they were disappointed when the agent's intelligence only extended towards responding to the their emotion and prompting more self-reflection. Although we think that self-reflection is an important component to include in future productivity agents, the usefulness of self-reflection in this context may be very dependent on various factors like the user's personality and organizational work role. Future research should investigate these factors further. We believe incorporating human-like qualities and emotional intelligence into future agents to be worthwhile; however, intelligence should also extend into other aspects of the agent's capabilities in order to better help users be as efficient as possible in achieving their work goals. Future research should investigate the anthropomorphic aspects of agent design that simultaneously improve likeability of the agent and also improve user efficiency at work. As one participant (P24) stated in their exit survey: *"There is a delicate balance to strike here between helping someone be more productive, and wasting time. Just make sure the agents are really efficient and use up very little time."*

Improved Task Scheduling Intelligence: We see one immediate opportunity for improving upon other aspects of our agent intelligence is to make the agent-assisted task scheduling more flexible and adaptive. In the work from Kimani et al. [16], they found that users desired more control and guidance from their productivity agent. In our work, we created a more rigid and structured task scheduling system that began at the beginning of each day, however, our participants reported that they found this system to not be flexible enough to deal with the frequent need to switch priorities on the fly. Some participants also reported that it was difficult sometimes to estimate ahead of time how much time should be blocked off on their calendar for a given tasks. Incorporating an additional dialog model that could be initiated by the user to reschedule their tasks, or increase the time of a task if necessary, could be very beneficial in helping users navigate the sometimes turbulent and unpredictable nature of a work day. Including the ability for users to communicate and configure their task-scheduling preferences (e.g., usual lunch time or time of the day they like to do stressful work) would likely be very beneficial as well.

Create more robust and personalized distraction detection logic: The distraction detection logic we incorporated into our agent was quite basic, as we only monitored users' windowing activity for common sites that we judged to be distractions. However, it is difficult to determine how comprehensive our distraction site list was, and it is likely that participants may have had different interpretations of what constitutes a distraction for their own task-related goals. Furthermore, we did not include any way for our agent to detect and intervene during face-to-face human interruptions. This is a challenging problem to tackle, and past work using chatbot initiated reminders has revealed similar desires for these reminders to be more context-sensitive [34]. Future work should aim to construct a more robust distraction site list and incorporate face detection and voice sensing data to detect human interruptions. In addition, future systems should enable users to personalize the apps and websites that they know are personal distractors for the agent to monitor for, rather than relying solely on a universal preset list.

7.1 Limitations

Our work has a few important limitations that should be noted. First, our agent prototypes, particularly the VA prototype, used users' webcams to continuously monitor their emotional state. User surveillance like this may result in some users feeling anxious that they are being monitored. These challenges are important and relevant, and we recommend future work to investigate the nature of user privacy concerns in this context. Second, a major dependency for our prototypes to be effective was the use of desktop notifications to help users organize and navigate their daily schedule. Therefore, the insights into the usefulness of these prototypes are likely only applicable to 'desk based' information workers. Furthermore, in our work we did not isolate the effect that these reminders and notifications that were present in the TB and VA prototypes had on user productivity, compared to other features of the agent. It is possible that the existence of these reminders may be the primary feature that resulted in greater perceived productivity in C2 and C3 compared to C1. Future work should disentangle and evaluate the effect of reminders from other agent features (e.g., anthropomorphism and emotional intelligence). Finally, our agent was designed to have a female gender through appearance and voice (for the VA prototype). We did this in order to mirror past work in this domain, where researchers also designed a female agent [16]. As research suggests that the combination of an agent's gender and personality can play an important role in user perceptions and expectations [1, 37], employing a male agent instead may result in some significant differences in user perceptions or ratings of the agent. We encourage future work to investigate how gender and personality of a workplace productivity agent might influence user experience.

8 CONCLUSION

In this paper, we describe our productivity agent designed to aid information workers be more productive and focused over the course of the work day. We designed two different agent prototypes: a text-based (TB) agent with a similar UI to a standard chatbot, and a more emotionally expressive virtual agent (VA) that employs a video avatar and the ability to detect and respond appropriately to users' emotions. Through a 3-week user study of these two agent prototypes and a baseline, we found that during their agent scheduled tasks, participants spent more time present at their workstation and on productivity apps and sites compared to their normal baselines. Participants reported that they felt more productive and satisfied with the VA prototype compared to the TB prototype, however they reported valuable feedback on how the agent could be improved, particularly through more personalized and intelligent task scheduling and distraction monitoring systems. We see our results as encouraging, and hope they can provide actionable insights for the development of future workplace productivity agents.

REFERENCES

[1] Fatema Akbar, Ted Grover, Gloria Mark, and Michelle X Zhou. 2018. The Effects of Virtual Agents' Characteristics on User Impressions and Language Use. In *Proceedings of the 23rd International Conference on Intelligent User Interfaces Companion*. ACM, 56.

[2] Anja Baethge and Thomas Rigotti. 2013. Interruptions to workflow: Their relationship with irritation and satisfaction with performance, and the mediating roles of time pressure and mental demands. *Work & Stress* 27, 1 (2013), 43–63.

[3] Emad Barsoum, Cha Zhang, Cristian Canton Ferrer, and Zhengyou Zhang. 2016. Training deep networks for facial expression recognition with crowd-sourced label distribution. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*. ACM, 279–283.

[4] Fabrica Binaria. 2019. TrackTime. https://www.fabbricabinaria.it/solutions/tracktime

[5] Jelmer P Borst, Niels A Taatgen, and Hedderik van Rijn. 2015. What makes interruptions disruptive?: A process-model account of the effects of the problem state bottleneck on task interruption and resumption. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*. ACM, 2971–2980.

[6] Stoney Brooks. 2015. Does personal social media usage affect efficiency and well-being? *Computers in Human Behavior* 46 (2015), 26–37.

[7] Scott A Cambo, Daniel Avrahami, and Matthew L Lee. 2017. BreakSense: Combining physiological and location sensing to promote mobility during work-breaks. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 3595–3607.

[8] Justin Cranshaw, Emad Elwany, Todd Newman, Rafal Kocielnik, Bowen Yu, Sandeep Soni, Jaime Teevan, and Andrés Monroy-Hernández. 2017. Calendar.help: Designing a workflow-based scheduling agent with humans in the loop. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 2382–2393.

[9] Mary Czerwinski, Eric Horvitz, and Susan Wilhite. 2004. A diary study of task switching and interruptions. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 175–182.

[10] Andrew Faulring, Brad Myers, Ken Mohnkern, Bradley Schmerl, Aaron Steinfeld, John Zimmerman, Asim Smailagic, Jeffery Hansen, and Daniel Siewiorek. 2010. Agent-assisted task management that reduces email overload. In *Proceedings of the 15th international conference on Intelligent user interfaces*. ACM, 61–70.

[11] Asma Ghandeharioun, Daniel McDuff, Mary Czerwinski, and Kael Rowan. 2018. EMMA: An Emotionally Intelligent Personal Assistant for Improving Wellbeing. *arXiv preprint arXiv:1812.11423* (2018).

[12] Jacek Gwizdka. 2000. Timely reminders: a case study of temporal guidance in PIM and email tools usage. In *Chi'00 extended abstracts on human factors in computing systems*. ACM, 163–164.

[13] Annabell Ho, Jeff Hancock, and Adam S Miner. 2018. Psychological, relational, and emotional effects of self-disclosure after conversations with a chatbot. *Journal of Communication* 68, 4 (2018), 712–733.

[14] Shamsi T Iqbal and Eric Horvitz. 2007. Conversations amidst computing: A study of interruptions and recovery of task activity. In *International Conference on User Modeling*. Springer, 350–354.

[15] William Jones, Harry Bruce, and Susan Dumais. 2001. Keeping found things found on the web. In *Proceedings of the tenth international conference on Information and knowledge management*. ACM, 119–126.

[16] Everlyne Kimani, Kael Rowan, Daniel McDuff, Mary Czerwinski, and Gloria Mark. 2019. A Conversational Agent in Support of Productivityand Wellbeing at Work. In *Proceedings of the 7th Intenrational Conference on Affective Computing and Intelligent Interaction*. IEEE.

[17] Jonathan Klein, Youngme Moon, and Rosalind W Picard. 2002. This computer responds to user frustration: Theory, design, and results. *Interacting with computers* 14, 2 (2002), 119–140.

[18] Rafal Kocielnik, Daniel Avrahami, Jennifer Marlow, Di Lu, and Gary Hsieh. 2018. Designing for workplace reflection: a chat and voice-based conversational agent. In *Proceedings of the 2018 Designing Interactive Systems Conference*. ACM, 881–894.

[19] Gloria Mark, Mary Czerwinski, and Shamsi T Iqbal. 2018. Effects of Individual Differences in Blocking Workplace Distractions. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 92.

[20] Gloria Mark, Daniela Gudith, and Ulrich Klocke. 2008. The cost of interrupted work: more speed and stress. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. ACM, 107–110.

[21] Gloria Mark, Shamsi Iqbal, and Mary Czerwinski. 2017. How blocking distractions affects workplace focus and productivity. In *Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2017 ACM International Symposium on Wearable Computers*. ACM, 928–934.

[22] Gloria Mark, Shamsi Iqbal, Mary Czerwinski, and Paul Johns. 2015. Focused, aroused, but so distractible: Temporal perspectives on multitasking and communications. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*. ACM, 903–916.

[23] Gloria Mark, Shamsi T Iqbal, Mary Czerwinski, Paul Johns, and Akane Sano. 2016. Neurotics can't focus: An in situ study of online multitasking in the workplace. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. ACM, 1739–1744.

[24] Daniel McDuff, Kael Rowan, Piali Choudhury, Jessica Wolk, ThuVan Pham, and Mary Czerwinski. 2019. A Multimodal Emotion Sensing Platform for Building Emotion-Aware Applications. *arXiv preprint arXiv:1903.12133* (2019).

[25] Daniel C McFarlane and Kara A Latorella. 2002. The scope and importance of human interruption in human-computer interaction design. *Human-Computer Interaction* 17, 1 (2002), 1–61.

[26] Microsoft. 2017. Bot Framework. https://dev.botframework.com/

[27] Microsoft. 2018. Language Instruction (LUIS). https://www.luis.ai/home

[28] Microsoft. 2018. Project Conversation Learner. https://labs.cognitive.microsoft.com/en-us/project-conversation-learner

[29] Microsoft. 2018. Turn Focus assist on or off in Windows 10. https://support.microsoft.com/en-us/help/4026996/windows-10-turn-focus-assist-on-or-off

[30] Christopher A Monk, J Gregory Trafton, and Deborah A Boehm-Davis. 2008. The effect of interruption duration and demand on resuming suspended goals. *Journal of Experimental Psychology: Applied* 14, 4 (2008), 299.

[31] Tim Rietz, Ivo Benke, and Alexander Maedche. 2019. The Impact of Anthropomorphic and Functional Chatbot Design Features in Enterprise Collaboration Systems on User Acceptance. (2019).

[32] Monica A Riordan. 2017. Emojis as tools for emotion work: Communicating affect in text messages. *Journal of Language and Social Psychology* 36, 5 (2017), 549–567.

[33] Brandon W Smit. 2016. Successfully leaving work at work: The self-regulatory underpinnings of psychological detachment. *Journal of Occupational and Organizational Psychology* 89, 3 (2016), 493–514.

[34] Carlos Toxtli, Andrés Monroy-Hernández, and Justin Cranshaw. 2018. Understanding chatbot-mediated task management. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 58.

[35] Steve Whittaker, Vaiva Kalnikaite, Victoria Hollis, and Andrew Guydish. 2016. 'Don't Waste My Time': Use of Time Information Improves Focus. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 1729–1738.

[36] Yang Yang, Xiaojuan Ma, and Pascale Fung. 2017. Perceived emotional intelligence in virtual agents. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. ACM, 2255–2262.

[37] Michelle X Zhou, Gloria Mark, Jingyi Li, and Huahai Yang. 2019. Trusting Virtual Agents: The Effect of Personality. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 9, 2-3 (2019), 10.

[38] Manuela Züger, Sebastian C Müller, André N Meyer, and Thomas Fritz. 2018. Sensing interruptibility in the office: A field study on the use of biometric and computer interaction sensors. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 591.