User-Centered Investigation of Features for Attention Management Systems in an Online Vignette Study

Dinara Talypova
University of Applied Sciences Upper Austria
Hagenberg, Austria
dinara.talypova@fh-hagenberg.at

Alexander Lingler
University of Applied Sciences Upper Austria
Hagenberg, Austria
alexander.lingler@fh-hagenberg.at

Philipp Wintersberger
University of Applied Sciences Upper Austria
Hagenberg, Austria
TU Wien
Vienna, Austria
philipp.wintersberger@fh-hagenberg.at

Figure 1: We used vignettes to immerse participants in different work-related roles such as air traffic controller, programmers, or communications officers but also private situations such as watching TV with friends. Participants rated the importance of various Attention Management System features considering these roles. Source: Getty Images, Shutterstock, Freepik.

ABSTRACT

Notifications and interruptions have shown to significantly impede task performance while causing stress. Attention management systems aim at mitigating these negative effects, for example, by delaying interruptions to task boundaries or times of low mental load. However, while the theoretical benefits of such an approach are well-documented, it is quite unclear how holding back information from users is accepted, especially in times of the "always-on-mentality". Thus, we conducted an online vignette experiment with N=163 participants, who were presented hypothetical private and work-related scenarios where interruptions are delayed by attention management systems. Participants rated how long they would allow particular interruptions to be delayed, as well as which data collection methods a system could use to perform these decisions. Our results show that interruption management is desired by potential users, provided they feel in control. We conclude with recommendations for the design of attention management systems.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in ubiquitous and mobile computing**; **User interface management systems**; *Interaction paradigms*; **Empirical studies in HCI**.

MUM '23, December 03–06, 2023, Vienna, Austria

ACM ISBN 979-8-4007-0921-0/23/12.
https://doi.org/10.1145/3626705.3627766

KEYWORDS

Attention Management, Human-Computer Interaction, Vignette Study, Attentive User Interface

ACM Reference Format:
Dinara Talypova, Alexander Lingler, and Philipp Wintersberger. 2023. User-Centered Investigation of Features for Attention Management Systems in an Online Vignette Study. In *International Conference on Mobile and Ubiquitous Multimedia (MUM '23), December 03–06, 2023, Vienna, Austria.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3626705.3627766

1 INTRODUCTION

Due to the ubiquitous presence of digital technology and smart devices, our attention is strained more than ever in human history. We constantly have to deal with interruptions and notifications (according to Pielot et al. [29], on average, more than 50 times per day), and we frequently fail to manage our response in an intelligent way. Imagine that while you are reading this introduction, you are receiving a notification from a short message service (such as WhatsApp or Signal) from a friend. Since such services affect autonomy and privacy, and blur the boundaries between work and private contexts [24] you will likely react or even respond. Afterward, you may continue reading this paragraph, but you have to re-attend the task mentally (i.e., freeing cognitive resources) and physically (i.e., putting away the smartphone). Overall, the detrimental effects of frequent task switches on human performance and comfort are well documented [6]. What if your friend's message had been managed more "attentively", for example, the notification could have been postponed until you finished reading this section or even the entire paragraph?

Unfortunately, most notifications and interruptions we receive are unmanaged, i.e., they are communicated to the user as soon as they appear. As a consequence, many users deactivate notifications from a wide range of apps or frequently put their devices into silent mode when they do not want to get disturbed [25]. Nevertheless, the calls for more intelligent management of users' attention date back to the origins of the "ubicomp vision" [3, 42], and it has been shown that interruptions work best at task boundaries or phases of low mental workload. Especially deferring interruptions is a powerful feature, since holding them back *"for a short time, i.e., just a few seconds, can lead to a large mitigation of disruption"* [6]. This is also frequently reflected in design guidelines. For example, Horvitz [19] argued to *"consider the status of a user's attention in the timing of services [...] while considering costs and benefits of deferring actions"*, and also in the more recent design guidelines on human-AI interaction by Amershi et al. [2] it is emphasized to *"time when to act or interrupt based on the user's current task and environment"*. Still, times and users have changed since the original proposals of attention management systems. Today, many users follow the "always-on mentality" and deliberately expose themselves to multiple conversations and media content simultaneously [41]. Further, despite the negative effects of multitasking, users may suffer the "fear of missing out" and query social media and messaging apps even without notifications [14]. Although we believe that it is more than urgent to develop suitable attention management systems for smart devices, we think it is also time to evaluate which associated features are valued and accepted by potential users.

Consequently, we have investigated these issues in an online vignette study. Since interruption receptivity is strongly based on context [3], we created multiple scenarios where attention management systems (AMSs) could be put into place (different workplace and private situations with varying risks and multitasking aspects) and evaluated the potential of AMS features. Specifically, we asked (a) how long different types of notifications (i.e., work and private-related) could be comfortably deferred in these scenarios, as well as (b) which observation parameters (such as screen recordings or physiological measurements) are accepted in these settings to optimize interruption delivery. Additionally, we asked our participants about acceptance and doubts regarding relevant AMS features. Our results show that many potential users welcome systems that better manage interruptions for them, as long as they feel in control of such systems. Finally, we contribute with a thorough set of recommendations for future AMSs.

2 BACKGROUND & THEORETICAL IMPLICATIONS

Cognitive science research has extensively examined the impact of interruptions on multitasking [7, 9, 10, 12, 18, 20, 23, 31]. According to Okoshi et al. [26] through Anderson et al. [3], an interruption is an "introduction to a new task or tasks on top of the ongoing activity, often unexpectedly, resulting in conflicts and loss of attention on the current activity, failing to resume the work where it was interrupted." Studies have consistently demonstrated that interruptions disrupt cognitive processes and can lead to performance decrements and increased error rates [6, 37, 45]. When individuals are interrupted from a primary task to engage in a secondary task, their cognitive resources are divided between the tasks, leading to decreased performance. Interruption costs are influenced by factors such as task complexity, interruption frequency, duration, and the nature of the secondary task [10, 12, 18]. Resumption lags — the time taken to return to the interrupted primary — also impact performance. The transition between the secondary and primary tasks during resumption requires cognitive resources for task-switching and reactivation of relevant information.

The Threaded Cognition theory [8] offers insights into these interruption-related phenomena. According to this theory, cognitive processes are not isolated but interconnected threads that can operate in parallel [34]. Resources execute processes exclusively in service of one task thread at a time. Interruptions can disrupt these threads, resulting in conflict for cognitive resources. The theory suggests that the cognitive system dynamically allocates resources to various threads, adapting to task demands and priorities [33]. When an interruption occurs, the cognitive system must shift resources from the primary task thread to the secondary task thread. Resuming the primary task requires the reallocation of resources back to the original thread, which involves cognitive effort and potential delays. Longer interruption and resumption lags can lead to interference between threads, resulting in performance decrements.

Delaying the delivery of notifications between subtasks as a potential solution draws upon cognitive psychology and theories related to attention management and multitasking. The core idea behind this approach is to strategically time notifications to minimize interruption costs and enhance overall task performance. Multiple studies revealed that it is better to be interrupted between (sub)tasks [1, 21, 32, 44] than in the middle of a task. One of the explanations comes from the Memory-for-problem-states theory, suggesting that individuals are more likely to have an active problem-solving state during the middle of a task, but not between distinct tasks [10].

To counter the negative effects of task switching, researchers have proposed to implement so-called "Attention Management Systems" or "Attentive User Interfaces" [3, 40]. Those can be defined as systems that *"computationally seek to balance a user's need for minimal disruption and the application's need to efficiently deliver information"* [6]. Anderson et al. [3] conducted a comprehensive evaluation of AMSs in ubiquitous computing environments. They summarize that an AMS consists of components for sensing (data from diverse sources), processing (to extract patterns), inferring (concepts from features), modeling (interruptibility), and managing (the attentional user states).

In exploring ways to enhance user experience, scholars have suggested two key improvements: (1) refining interruption timing based on user receptiveness (interruptibility), and (2) reconfiguring interfaces to make returning to tasks more straightforward. A wealth of studies has focused on pinpointing the perfect moments for notifications [48], employing sensors and machine learning to determine when users are most open to interruptions [17, 22, 30, 47]. Solutions extend beyond mere timing, proposing the display of user availability [46, 47] and tools that ease the transition back to tasks after an interruption occurs [38].

With respect to supporting task transitions, evidence shows that well-arranged interfaces facilitate smoother task resumption through cues [27]. These cues, aiding users in re-engaging with their tasks, can be explicit (giving precise task details) or implicit

(nudging user focus without direct instruction). They can be communicated before, during, and after interruptions using different modalities [35].

3 METHOD AND RESEARCH QUESTIONS

To gain additional insights, we have set up a study to investigate relevant features and parameters of AMS systems in multiple scenarios. In particular, we focus on the following aspects:

- **Notification deferral times**, where distinguish between work-related and private interruptions. In addition, we split between perceived critical- and non-critical work interruptions, as well as important and everyday interruptions emerging from private life.
- **Data collection methods**: AMS systems require complex (and potentially privacy-invasive) information about users and their behavior to work properly [3]. Thus, we want to know how comfortable participants feel when an AMS system uses (1) operation system parameters like opened and foreground applications, (2) screen content recordings including page views, mouse click locations, etc., (3) physiological measurements such as heart rate, body temperature, activity, (4) eye-tracking to determine where a user looks and how concentrated/distracted they are, and (5) video recordings of the user interacting with the system.
- **General AMS Features**, including but not limited to ones defined in [3]; (a) delaying notifications from private and work contexts in general, (b) emergency options/contacts not included in deferral, (c) assessment of the relative importance of interruptions based on context, (d) modifying the order of multiple notifications based on urgency, and (e) aids/interfaces helping to quickly reengage in a before suspended task.

To test these features and parameters, we conducted a vignette study. Vignettes are short descriptions of situations representing a characteristic or a combination of characteristics. They are used as an elicitation tool that facilitates the discovery of subjects' responses to presumptive situations [43]. Over the last 50 years, the vignette technique has found application across numerous disciplines [11]. It boasts a rich history in the exploration of various phenomena within the realms of behavioral, social, and health sciences, both in quantitative and qualitative research [15, 16]. According to Atzmüller and Steiner [4], the method is well-suitable *"for investigating respondents' beliefs, attitudes, or judgments"*. Besides their immersive quality, vignettes provide the opportunity to combine different sets of factors and levels to define the determining variables of the outcome [4, 5, 39]. Due to the nature of measured variables, we wanted to see how the answers vary depending on different factors of the vignettes. Thus, three factors with two levels each were defined: work/private **environment**, high/low **risk** and high/low **multitasking** demand (see Table 1 and section 3.1). The "level of multitasking demand" was only included for the well-defined work-related scenarios, as we cannot control potential users' levels in their private life situations. Driven by this approach, we have created six presumptive scenarios to assess the corresponding set of factors. To confirm the vignettes' validation, in the analysis, we report both - the comparison of the scenarios as a standalone story and the comparison of the effect of each factor/set of factors on the dependent variables. Using an online survey, we wanted to answer the following research questions:

- **RQ1:** What are the accepted notification delay times considering (a) the scenario context (work vs. private, high vs. low risk, and multitasking demand) and (b) the notification context (urgent vs. non-urgent and work vs. private)?
- **RQ2:** What are the data collection methods an AMS is accepted to use, considering the scenario context (work vs. private, high vs. low risk, and multitasking demand)?
- **RQ3:** Which AMS features are most valued by potential users, and how is this influenced by demographic variables?

Table 1: The six scenarios were presented as vignettes to study participants. We systematically created different situations to assess the varying requirements for AMSs.

	Work Scenario	**Private Life Scenario**
High Risk	Air Traffic Controller *(high multitasking demand)* Staff Nurse *(low multitasking demand)*	Automated Driving
Low Risk	Communications Officer *(high multitasking demand)* IT Programmer *(low multitasking demand)*	Home/Living Room

3.1 Materials and Measurements

The vignettes for the six dedicated scenarios contained both an image (see Figure 1) and a verbal description to let participants immerse themselves in the situation. Those were explained in the vignettes (slightly shortened) as follows (starting with "Imagine you are a/an"):

- **Air Traffic Controller**: Your tasks include directing the movement of aircraft on the ground & in the air, issuing plane takeoffs & landings, etc. using tools such as radios, radars, and computers. While some tasks can be planned, others require immediate attention. You typically handle multiple aircraft simultaneously - for example, by providing weather information while directing another through its landing approach.
- **Staff Nurse**: Your duties include recording details of patients' health. Aside from administering medication, you supervise nursing assistants and trainees. You have to react to emergencies and often you are the first person in a critical situation. Typically, you are notified when your help is needed. Often, this can happen while you are busy with other tasks and patients.
- **Communications Officer**: Your work is to navigate the news and monitor market updates. While developing communication and brand strategies, you also define channels, organize press events, plan promo campaigns, and budget.

You handle media inquiries, write content for social networks, manage reputational crises, and serve as a coordinating point between different departments. Multitasking is a part of your job because you have to be efficient and react quickly.

- **IT Programmer**: You create code and most of your work communication happens online. You do many concentration-demanding tasks (like programming), but also exchange updates with your team and react to various issues. For this, your team uses messengers, a professional task-tracking system, and online meetings. It is important to receive updates and be notified about problems.
- **Home/Living Room**: You have an active life both at work and in private, and you try to always be in touch and don't miss quality time with people you like. Sometimes you can feel overwhelmed by the amount of messages and notifications you receive. You may feel FOMO (Fear Of Missing Out) that doesn't let you fully relax. To have a chilled evening, you and your friends typically meet for dinner or watch a TV series together.
- **Automated Driving**: Imagine you have a highly automated vehicle. You are currently driving autonomously on a highway, and you have to take the next exit, which you will reach in roughly 5 minutes. At some point before the exit, the car has to inform you to take over control manually. In the meantime, you are reading an article in your favorite newspaper on your smartphone.

All scenario descriptions were followed by an introduction of an AMS system (*"Imagine an Attention Management System integrated into your workflow. You have used the system for some time and it seems to work as intended. Instead of interrupting you immediately, the AMS delays the notification. The system considers the urgency of the request, but interrupts you soon enough so that no negative side effects can occur."*) This text segment was accompanied by an example of the tasks in each scenario, i.e., issuing starting permission for a plane, taking blood samples for a patient, email communication, meeting request, short message from a friend, and in-vehicle notification, where it was always argued that this behavior allows the user to complete a task or sub-task before being interrupted. The vignettes (description and images) were integrated into an online survey (using the LimeSurvey platform).

After reading the vignette, the respondents faced the question block regarding accepting different data collection methods under the presented circumstances (scenario). Each of the data collection methods was assessed on a 7-point Likert scale from 1 (completely uncomfortable) to 7 (completely comfortable). Further, we asked about appropriate delay times for critical/non-critical work-related, as well as important/everyday private notifications/interruptions on a 6-point Likert scale with the answering options: should not be delayed at all; up to several seconds; up to several minutes; up to one hour; up to several hours; should not be received at all.

Finally, we included a set of general features for AMSs at the end of the survey, independent of the particular scenarios. Participants rated each of these features (see Section 3) on a 7-point Likert scale from 1 (not needed at all) to 7 (very much needed).

3.2 Participants and Procedure

In total, N=163 participants were included in the analyses. To acquire a diverse sample, we collected participants from two sources: (1) by email newsletters, social media, etc., and (2) through the survey platform Prolific. Participants provided demographic data and were briefly introduced to the concept of AMS systems. Afterward, they were presented with the scenarios (vignettes). After each scenario, we assessed the AMS features of notification deferral times and data collection methods. Completing the survey took about 20-30s minutes per participant. All data were collected over the course of two weeks. Two hundred sixty-five (N=265) individuals participated in the vignette study by answering three randomly selected vignettes. Since data quality is a serious issue in online studies [36], we put a high emphasis on getting reliable responses and included multiple attention-check questions (explicitly instructed response items). We considered only those participants in our final evaluation who successfully passed all attention checks (2-3 attention checks were randomly assigned per a set of vignettes). This left us with 163 participants (83 male, 75 female, 5 other genders, age 18-55 M=29.3, SD=7.6 years). Most participants live in or originate from Europe (82%) and have at least Bachelor level of education (68%). Respondents' places of residence hold different population levels, starting from <200,000 inhabitants to cities with more than 1,000,000 inhabitants. Slightly more than 15% of the sample (N=25) do not have children, >35% do not have and do not plan to have one (N=59), and >40% want to become parents in the future (N=70). Nine (9) respondents preferred not to share this information.

4 RESULTS

In the following, we present the results with respect to the individual research questions. For statistical analyses, we conducted Ordinal Logistic Regression with Cumulative Link Mixed Model (CLMM) to accommodate the ordinal nature of the data (Likert-scale) [5] and the random effects of different vignettes and individuals, see [4]. The CLMM analysis discerns potential differences in the odds ratios (OR). The ORs indicate the odds of choosing a higher response category on the Likert scale for each target item/question relative to the reference item/question. Upon completing the analysis for each research question, we established reference levels using, where relevant, the following criteria: readability (ensuring ORs are directed in one way for simpler understanding), logical structure (e.g., representing "child-free" individuals as one and "parents" as the other end of the scale), and/or sequencing (e.g., alphabetical order in binary categories). The analysis was conducted using the "ordinal" [13] package within the R programming environment. Table 2 shows the descriptive statistics (mean, median, mode) of the delay times and data collection methods.

4.1 Comfortable Delay Times

Critical work-related notifications got the biggest number of "Should not be delayed at all" scores, followed by the important notifications from private contacts. Respondents tend not to delay important notifications at all or not more than for several minutes (93% for work-related and 83% for private), while for everyday notifications, there is a high proportion of votes for delaying from an hour up to several hours - or not receiving at all (40% for work-related and

Table 2: Descriptive Statistics for the comfortable delay times and data collection methods assessed in the individual scenarios.

M (SD) Med (IQR) Mode	Air Traffic Controller	Staff Nurse	Comm. Officer	IT Progr.	At Home	Auto. Vehicle	Overall
Comfortable Delay Times							
Critical work-related	- 1 (1) 1	- 1 (1) 1	- 1 (1) 1	- 2 (2) 1	- 1 (2) 1	- 1 (2) 1	- 1 (1) 1
Non-critical work-related	- 3 (2) 3	- 3 (2) 3	- 4 (1) 4	- 4 (1) 4	- 4 (1.75) 3	- 3 (2) 3	- 3 (1) 3
Important private	- 2 (1.75) 2	- 2.5 (1) 3	- 2 (2) 3	- 3 (2.5) 1	- 2 (2) 1	- 2 (2) 1	- 2 (2) 1
Everyday Private	- 4 (2) 5	- 4 (2) 4	- 4 (1) 4	- 4 (2) 5	- 4 (2) 4	- 3 (1) 3	- 4 (2) 4
Data Collection Methods							
Op. system params	4.85 (1.6) 5 (2) 6	4.2 (1.73) 4 (3) 3	4.33 (1.91) 5 (3) 6	4.33 (1.77) 4 (3) 3	3.84 (1.99) 4 (3) 5	4.32 (1.67) 4 (3) 4	4.3 (1.8) 4 (3) 6
Screen content recordings	4.15 (1.8) 4 (3) 3	3.98 (1.81) 4 (4) 6	3.58 (1.84) 3 (3) 3	3.84 (1.78) 4 (2.5) 5	2.93 (1.74) 3 (3) 1	3.84 (1.84) 3 (3) 3	3.71 (1.83) 3 (3) 3
Physio. measurements	4.55 (1.77) 5 (3) 5	4.75 (1.87) 5 (3) 5	4.44 (2.03) 5 (3.5) 6	4.16 (2.04) 5 (3) 6	3.82 (1.96) 4 (3.75) 2	4.91 (1.71) 5 (2) 6	4.44 (1.93) 5 (3) 6
Eye tracking	4.29 (1.73) 5 (2.75) 5	3.81 (1.9) 4 (3.25) 4	3.52 (1.95) 3 (3) 3	3.81 (1.97) 4 (3) 5	3.15 (1.73) 3 (2) 3	4.39 (1.98) 5 (4) 5	3.81 (1.92) 4 (3) 3
Video rec. of user	3.3 (1.62) 3 (2) 3	3.1 (1.64) 3 (2) 3	2.62 (1.73) 2 (2) 1	2.85 (1.58) 3 (2) 3	2.54 (1.84) 2 (2) 1	3.2 (1.78) 3 (3) 3	2.93 (1.72) 3 (3) 1

62% for private, depending on context). The distribution of answers is shown in Figure 2. Due to the ordinal structure of the data, we provide only medians and modes in Table 2.

The CLMM analysis revealed notable differences in odds ratios among all the questions. As stated above, the ORs indicate the odds of choosing a higher response category on the Likert scale for each target question relative to the reference question. In our case, this means "reversed" grading: "up to several hours" and "should not be received at all" are noted with 5 and 6, while "should not be delayed at all" and "up to several seconds" with 1 and 2. All ORs exhibit significant differences from the reference. Thus, everyday notifications from private contacts, non-critical work-related tasks, and important notifications from private contacts are 57.3, 20.4, and 4 times more likely, respectively, to be delayed for longer. The ORs and corresponding CIs (95%) are summarized in Table 3.

We explored pairwise differences of estimated marginal means (EMM) in coefficients between the questions with post-hoc Tukey analysis. The findings reveal significant coefficients' differences ($p<.001$) among all the question pairs, indicating distinctness in participants' responses across the questions. As expected, everyday notifications from private contacts demonstrated the largest (-4.05, SE=0.17) coefficient difference from the reference (critical work-related notifications), indicating that participants were more likely to delay longer everyday private notifications relative to critical work-related ones. Interestingly, important messages from private contact demonstrate a significant but relatively low coefficient difference from the non-critical work-related tasks compared to the opposite pair (critical work-related vs. everyday private) described above. The coefficients for compared pairs are presented in the second part of Table 3.

Ordinal regression analysis was further applied to each of the questions separately to define the factors that influence the odds of higher scores (longer delays). We analyzed three variations/layers of each question: scenario variation (treating each presented scenario as a standalone story rather than a combination of factors within a vignette), factors set variation (exploring the impact of three factors - high/low risk, high/low multitasking, and private/working environment - on the question's final score), and demographic variation (investigating how participants' traits may influence the score). The Air Traffic Controller scenario was selected as a reference level for scenario variation (which is a risky working environment with a high level of multitasking load). For the factors set variation, we chose an opposite set combination as reference levels: low risk, low multitasking private environment. In demographics measurements, reference levels were: childless individuals without the intention of becoming a parent, female, with a residence place of <200,000

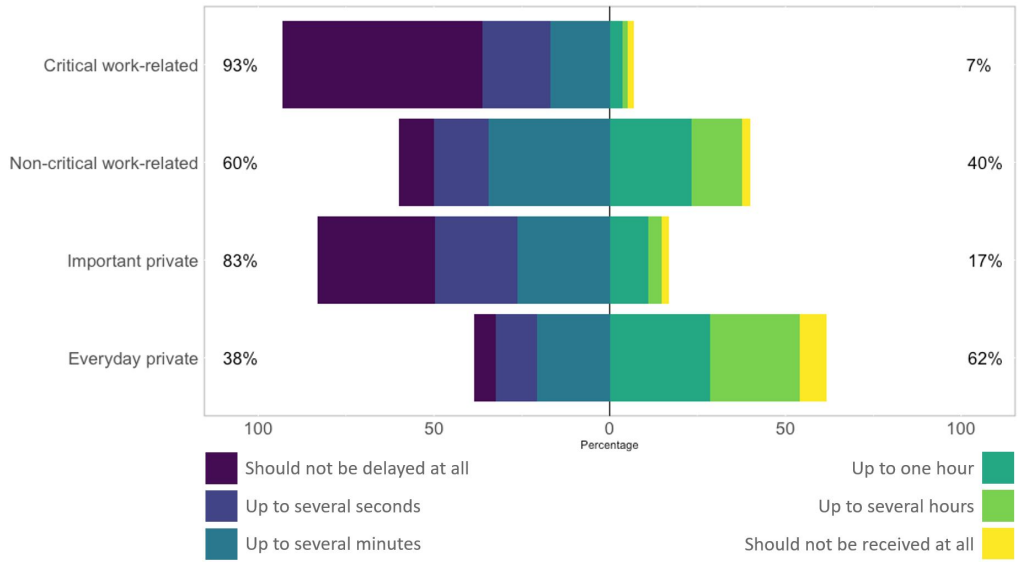

Figure 2: Distribution of answers on a 6-point Likert scale on the question "How long do you think the AMS should delay the following interruptions?"

Table 3: Ordinal Regression with CLMM for Comfortable Delay Times.

Questions		OR	95% CI
(ref.) Critical work-related		—	—
Everyday from private contacts		**57.3*****	41.5, 79.1
Non-critical work-related tasks		**20.4*****	15.2, 27.5
Important from private contacts		**4.00*****	3.03, 5.27
Contrast (EM means pairwise comparisons)		**Coef.**	**SE**
Critical work-related	Everyday from private contacts	**-4.05*****	0.17
Critical work-related	Important from private contacts	**-1.39*****	0.14
Critical work-related	Non-critical work-related tasks	**-3.02*****	0.15
Everyday from private contacts	Important from private contacts	**2.66*****	0.14
Everyday from private contacts	Non-critical work-related tasks	**1.03*****	0.13
Important from private contacts	Non-critical work-related tasks	**-1.63*****	0.13

inhabitants. See Table 4 for the summarized results of different variations regression analyses with ORs and correspondent CIs.

Critical work-related notifications. Ordinal regression analysis with an independent variable as scenario revealed significant differences in odds ratios (p<.05) from the reference scenario (Air Traffic Controller) with all the scenarios with the exception of Staff Nurse (the latter belongs to the high-risk*working environment just like Air Traffic Controller). The remaining scenarios showed significantly bigger odds of choosing longer delays (responses "up to several hours/should not be received at all") than the reference one, with the biggest OR (7.16, p<.001) in the IT Programmer scenario. Analysis conducted for factors set variations confirmed the findings on the scenario layer: interaction of high risk and working environment predict significantly (p<.001) lower odds to get longer delayed answers (i.e., participants are more likely to choose no/short delays) compared to a combination of lower risk in the private environment. Notably, each factor alone (risk or environment) does not have a significant effect. Lastly, demographic factors analysis showed that males have significantly (p<.001) lower chances (OR=0.18) of choosing longer delays for critical work-related notifications compared to females. In other words, males are less prone to postpone critical work messages than females.

Important notifications from private contacts. Both "At Home" and "Autonomous Vehicle" scenarios showed significantly (p<.05) lower ORs for the higher response category compared to the reference. That is, in these scenarios, participants are prone to receive important private messages sooner compared to scenarios at work. Since both scenarios have a characteristic of a private environment, this result could be due to respondents' association that private messages should be left for after-work time. The factors set model repeated the results from the scenario layer findings: people with the working environment conditions have 3.22 (p<0.05) higher chances to choose longer delays for important private messages. Demographic data did not manifest significant findings (p<.05); however, males again showed lower chances (OR=0.39) of giving a

Table 4: Ordinal Regression with CLMM for Comfortable Delay Times: scenarios, leafs sets and demographic data

	Critical work-related		Important from private contacts		Non-critical work-related		Everyday from private contacts	
	OR	95% CI	OR	95% CI	OR	95% CI	OR	95% CI
Questions' variables								
Vignettes								
Air Traffic Controller	—	—	—	—	—	—	—	—
Staff Nurse	1.15	0.48, 2.76	1.1	0.55, 2.23	0.87	0.44, 1.73	1.35	0.68, 2.68
Communications Officer	**3.65****	1.54, 8.67	0.8	0.39, 1.65	**2.2***	1.11, 4.35	1.4	0.71, 2.79
IT Programmer	**7.16*****	2.96, 17.3	1.22	0.57, 2.59	**2.66****	1.29, 5.47	1.09	0.54, 2.23
At Home Scenario	**3.73****	1.50, 9.33	**0.38****	0.17, 0.82	**2.28***	1.10, 4.75	0.59	0.29, 1.23
Autonomous Car	**3.83****	1.55, 9.48	**0.39****	0.18, 0.83	0.94	0.46, 1.92	0.49'	0.24, 1.01
Factors								
Risk								
(ref.) Low	—	—	—	—	—	—	—	—
High	1.18	0.36, 3.90	1.14	0.41, 3.20	**0.36***	0.13, 0.96	1.12	0.41, 3.00
Environment								
(ref.) Private	—	—	—	—	—	—	—	—
Work	1.92	0.85, 4.32	**3.22****	1.47, 7.07	1.17	0.57, 2.41	1.85	0.90, 3.81
Multitasking								
(ref.) Low	—	—	—	—	—	—	—	—
High	0.51	0.24, 1.09	0.66	0.32, 1.35	0.83	0.42, 1.62	1.28	0.65, 2.55
Risk * Environment								
high risk * work	**0.14*****	0.04, 0.45	0.79	0.27, 2.29	0.91	0.34, 2.48	1.11	0.41, 3.01
Risk * Multitasking								
high risk * high multi	1.71	0.54, 5.41	1.38	0.50, 3.77	1.38	0.52, 3.64	0.58	0.22, 1.52
Respondents' variables								
Age	1.03	0.94, 1.13	1.01	0.91, 1.12	1.04	0.95, 1.14	**1.18****	1.06, 1.31
Children								
(ref.) No, and don't plan to have	—	—	—	—	—	—	—	—
No, but plan to have	1.09	0.37, 3.19	1.55	0.52, 4.62	1.18	0.45, 3.11	0.85	0.29, 2.53
Yes	2.81	0.39, 20.3	3.45	0.49, 24.4	1.01	0.18, 5.76	1.25	0.18, 8.75
Gender								
(ref.) Female	—	—	—	—	—	—	—	—
Male	**0.18*****	0.07, 0.47	0.39'	0.15, 1.02	0.53	0.23, 1.23	0.52	0.20, 1.34
Population								
<200,000 inhabitants	—	—	—	—	—	—	—	—
<500,000 inhabitants	2.42	0.55, 10.6	0.7	0.16, 3.08	0.65	0.18, 2.40	0.64	0.15, 2.79
<1,000,000 inhabitants	1.61	0.37, 6.94	1.07	0.24, 4.72	0.42	0.11, 1.62	0.23'	0.05, 1.03
>1,000,000 inhabitants	1.62	0.51, 5.15	0.39	0.12, 1.26	0.74	0.26, 2.10	0.37	0.12, 1.21
Age*children								
age * No, but plan to have	0.86	0.72, 1.02	0.99	0.83, 1.18	0.91	0.78, 1.06	**0.75*****	0.62, 0.89
age * Yes	0.81'	0.66, 1.00	0.9	0.74, 1.11	0.94	0.78, 1.12	**0.82***	0.66, 1.00

higher score (bigger delays), although the p-value slightly exceeded the conventional threshold (p=0.055).

Non-critical work-related notifications. In the context of non-critical working notifications, non-risky scenarios obtained significant differences (p<.05) with the reference level of high risk: we found that "Communications Officer", "IT- Programmer", and "At Home" scenarios have more than two times higher chance of getting "more delayed" answer score compared to reference scenario "Air Traffic Controller" (and other risky scenarios - Staff Nurse and Autonomous Car - did not show a significant difference to the reference). As predicted, the high-risk level demonstrated a significantly lower OR (0.36, p<.05) for bigger delays compared to the low-risk level. The rest of the factors did not show any significant findings. We also did not detect any significant results in the demographic data.

Everyday notifications from private contacts. For this question, no significant effects that influence delay score were found between scenarios or scenarios' (vignettes') factors. Yet, we revealed a significant correlation between respondents' age as well as the interaction between age and parental status/attitude. The OR for age centered around the sample mean (M = 29.3) was found to be 1.18. The findings indicate that for each year change, holding other variables constant, the odds of the "more delayed" score increase by approximately 18%. In other words, we could say that the older are

participants the more they tend to delay everyday private messages. Interesting is the effect of the combination of age and parental status. Thus, parents and those who want to have children in the future, with increasing age, have respectively 18% (p<.05) and 25% (p<.001) lower odds of choosing a higher response category (longer delays) than those without plans of having children in the future. Still, the model did not reveal any significant result for parent status alone but only with the interaction with age.

4.2 Data Collection Methods Acceptance

Regarding potential parameters of AMSs, video recordings (of the user) got the biggest number (71%) of negative scores across different scenarios, with the smallest number (19%) of positive (comfortable) scores. At the same time, the physiological measurements question got the biggest number of above-neutral scores (55%). The distribution of answers is shown in Figure 3 (left). From modes in Table 2, we can define three groups of (un)comfortable data collection methods: operation system parameters and physiological measurements have mode 6 (mostly comfortable), screen content recordings and eye tracking have mode 3 (slightly uncomfortable), while video recordings get the most common score 1 (completely uncomfortable).

Since this type of response category is frequently treated as continuous in Likert-scale surveys, we additionally report the means. Figure 3 (right) illustrates the scores' means distinguished per question and scenario. It visualizes similarities and differences of the average participant's path through all the questions in each scenario. Additionally, symbols differentiate high-risk from low-risk scenarios. Hence, the graph demonstrates similar overall preferences among data collection methods as described at the beginning of the section. The "At Home" scenario has the lowest mean scores among all the questions, while the rest of the scenarios have more elaborate ratings depending on the question. At the same time, high-risk scenarios tend to be assessed higher (or not lower) in the comfort scale among all the questions. To assess the relevance of these first findings, we conducted ordinal regression analysis following the same principle as in the "Comfortable Delay Times" section: comparisons of questions' differences, then scenario variation, factors set variation, and demographic variation to estimate whether these variables determine the final score. Table 6 summarized the results of different variations regression analyses with ORs and correspondent CIs.

CLMM compared the ordinal responses obtained from 5 Likert-scale questions with the reference level of "Operation system parameters". The analysis demonstrated no significant differences in odds ratios between reference level and physiological measurements (p=0.15), while all other questions got significantly lower (p<.001) ORs. Hence, the regression analysis result confirmed our intuitive first findings. Post-hoc Tukey analysis of pairwise differences in EMMs revealed significant coefficients' differences (p<.001) among all the question pairs, besides the above-mentioned pair of "Operation system parameters" - "Physiological measurements" as well as "Screen content recordings" - "Eye Tracking" - again, in line with the previous observations of the descriptive data. Other pair comparisons also confirm our comfortable methods' rating derived from the modes. The ORs and corresponding CIs (95%) from CLMM, as well as coefficients with SEs for compared pairs, are summarized in Table 5.

Operation system parameters. Ordinal regression analysis with scenario as an independent variable revealed that the reference scenario (Air Traffic Controller) had significantly higher odds ratios for choosing a higher score in the operation system (OS) parameters question among all, i.e., OS tracking is perceived to be more comfortable in Air Traffic Controller context compared to the rest scenarios. The exception is the Communications officer scenario with a p-value of 0.055. Surprisingly, divergent trends emerged when we delved into factors set variations and demographic data. In these analyses, the anticipated significant trends did not materialize. This could be due to the interplay of other hidden variables not covered in the model since our scenarios do not totally resemble classical vignettes with minimum variations in the texts. However, another explanation is the absence of a two-level multitasking private environment to conduct a full comparison. Digging deeper, post-hoc Tukey analysis highlighted specific scenarios of interest. Two specific factor combinations stood out prominently: low-risk low-multitasking private environment (corresponds to "At Home" scenario) and high-risk high-multitasking private environment (corresponds to "Autonomous Car" scenario). These factor combinations showcased notably reduced coefficients when contrasted with the high-risk high-multitasking working environment (corresponds to the "Air Traffic Controller" scenario), exhibiting coefficient differences of -1.38 (SE = 0.37, $p < 0.01$) and -1.13 (SE = 0.36, $p < 0.05$), respectively. Upon closer examination through post-hoc pairwise comparisons with the application of Tukey correction, the distinction among scenario variations initially noted by CLMM was reaffirmed exclusively for the "Autonomous Car" and "At Home" scenarios. Briefly, the "Air Traffic Controller"-"Autonomous Car" and "Air Traffic Controller"-"At Home" scenario pairs sustained their significance in terms of odds ratios after accounting for Tukey correction.

Screen content recordings. CLMM results indicated that all scenarios had significantly lower ORs of selecting a higher comfort category in "Screen content recordings" compared to the reference level (Air Traffic Controller). The lowest OR was for the "At Home" scenario and was equal to 0.13 (p<.001). The factors set model additionally revealed that scenarios with a working environment had almost three (2.85) times higher chances of getting a higher comfort score than the private environment, holding other variables constant. Demographic variables did not demonstrate any significant findings.

Physiological measurements. We found that the Staff Nurse scenario (high-risk low-multitasking working environment) had significantly (but not drastically) higher odds (9% higher chance) of getting a more comfortable response for physiological measurements compared to the reference level (high-risk high-multitasking working environment). While Communications Officer, IT Programmer, At Home scenarios (all are low-risk scenarios) revealed significantly lower chances of getting high comfort scores. Autonomous car statistics did not reach significant results. Comparing odds on factors set variation confirmed that high-risk scenarios had 3.14 higher

Table 5: Ordinal Regression with CLMM for or Data Collection Methods Acceptance: "Please rate how comfortable you are with the system to use..."

Questions		OR	95% CI
(ref.) Operation system parameters		—	—
Screen content recordings		**0.48*****	0.37, 0.60
Physiological measurements		**1.2**	0.94, 1.53
Eye Tracking		**0.54*****	0.42, 0.69
Video recordings		**0.17*****	0.14, 0.23
Contrast (EM means pairwise comparisons)		Coef.	SE
Operation system parameters	Screen content recordings	**0.743*****	0.123
Operation system parameters	Physiological measurements	-0.181	0.125
Operation system parameters	Eye Tracking	**0.617*****	0.123
Operation system parameters	Video recordings	**1.744*****	0.129
Screen content recordings	Physiological measurements	**-0.925*****	0.124
Screen content recordings	Eye Tracking	-0.126	0.121
Screen content recordings	Video recordings	**1.001*****	0.125
Physiological measurements	Eye Tracking	**0.799*****	0.124
Physiological measurements	Video recordings	**1.925*****	0.131
Eye Tracking	Video recordings	**1.127*****	0.125

Table 6: Ordinal Regression with CLMM for Data Collection Methods Acceptance: scenarios, factors sets and demographic data

	Operation system parameters		**Screen content recordings**		**Physiological measurements**		**Eye Tracking**		**Video recordings**	
	OR	95% CI	OR	95% CI	OR	95% CI	OR	95% CI	OR	*95% CI*
Questions' variables										
Vignettes										
Air Traffic Controller	—	—	—	—	—	—	—	—		
Staff Nurse	**0.39****	0.20, 0.76	**0.5***	0.26, 0.98	**1.09*****	1.08, 1.09	**0.43***	0.22, 0.84	0.58	0.29, 1.14
Communications Officer	0.52'	0.27, 1.01	**0.3*****	0.15, 0.60	**0.84*****	0.83, 0.84	**0.33****	0.17, 0.66	**0.27*****	0.14, 0.55
IT Programmer	**0.43***	0.21, 0.86	**0.36****	0.18, 0.72	**0.49*****	0.49, 0.49	**0.38****	0.19, 0.78	**0.33****	0.16, 0.68
At Home Scenario	**0.25*****	0.12, 0.51	**0.13*****	0.06, 0.26	**0.32*****	0.19, 0.54	**0.2*****	0.09, 0.40	**0.13*****	0.06, 0.28
Autonomous Car	**0.32****	0.16, 0.65	**0.36****	0.18, 0.74	0.94	0.54, 1.63	0.89	0.43, 1.82	0.51'	0.25, 1.03
Factors										
Risk										
(ref.) Low	—	—	—	—	—	—	—	—		
High	0.5	0.19, 1.30	1.46	0.57, 3.75	**3.14***	1.19, 8.31	1.94	0.73, 5.12	2.23	0.83, 6.04
Environment										
(ref.) Private	—	—	—	—	—	—	—	—		
Work	1.71	0.85, 3.44	**2.85****	1.42, 5.74	1.53	0.76, 3.08	1.97'	0.97, 4.01	**2.55***	1.21, 5.36
Multitasking										
(ref.) Low	—	—	—	—	—	—	—	—		
High	1.21	0.62, 2.35	0.85	0.44, 1.65	1.7	0.86, 3.35	0.86	0.43, 1.72	0.82	0.41, 1.65
Risk * Environment										
high risk * work	1.81	0.69, 4.76	0.96	0.37, 2.52	0.7	0.26, 1.85	0.57	0.21, 1.55	0.78	0.28, 2.13
Risk * Multitasking										
high risk * high multi	2.12	0.83, 5.46	2.35	0.91, 6.02	0.55	0.21, 1.44	**2.71***	1.02, 7.16	2.11	0.80, 5.57
Respondents' variables										
Age	0.96	0.88, 1.05	0.98	0.90, 1.06	**0.9***	0.81, 1.00	1	0.91, 1.11	0.97	0.88, 1.06
Children										
(ref.) No, and don't plan to have	—	—	—	—	—	—	—	—		
No, but plan to have	1.08	0.43, 2.71	1.35	0.56, 3.23	1.01	0.34, 2.96	1.02	0.35, 2.95	1.46	0.54, 3.97
Yes	1.29	0.25, 6.70	1.61	0.33, 7.81	0.78	0.11, 5.55	0.42	0.06, 2.84	2.72	0.45, 16.4
Gender										
(ref.) Female	—	—	—	—	—	—	—	—		
Male	0.92	0.41, 2.04	0.92	0.43, 1.97	0.91	0.35, 2.33	1.54	0.61, 3.91	1.74	0.73, 4.16
Population										
<200,000 inhabitants	—	—	—	—	—	—	—	—		
<500,000 inhabitants	0.59	0.17, 2.07	2.29	0.69, 7.63	2.19	0.50, 9.70	2.9	0.68, 12.3	3.49	0.89, 13.7
<1,000,000 inhabitants	0.82	0.23, 2.90	1.4	0.42, 4.69	1.32	0.30, 5.82	1.61	0.37, 6.91	**10.8*****	2.70, 43.0
>1,000,000 inhabitants	0.6	0.22, 1.64	0.56	0.22, 1.46	0.55	0.17, 1.79	0.93	0.29, 2.93	2.05	0.69, 6.08
Age*children										
age * No, but plan to have	0.9	0.78, 1.05	0.95	0.83, 1.10	0.88	0.74, 1.05	0.86	0.73, 1.02	1.07	0.91, 1.25
age * Yes	1.09	0.92, 1.30	1.07	0.91, 1.27	1.2	0.97, 1.47	1.11	0.91, 1.36	1.07	0.89, 1.29

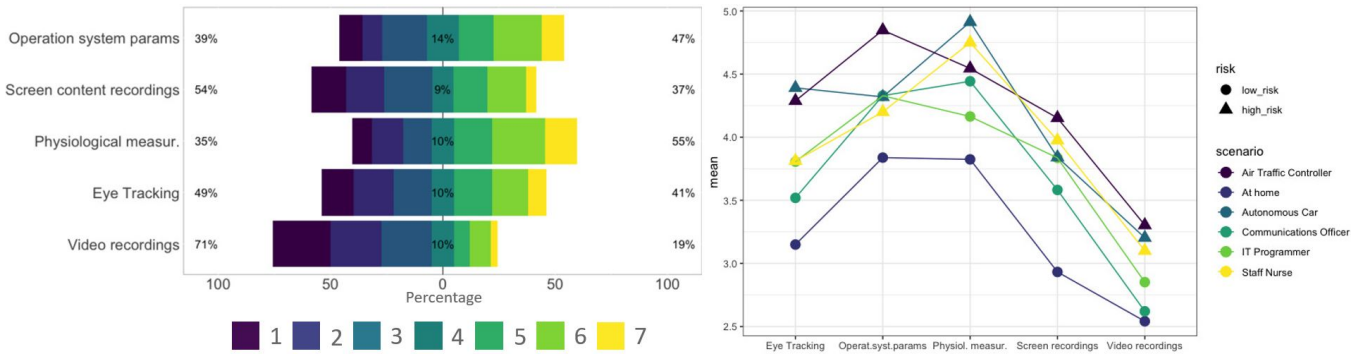

Figure 3: Data Collection Method Acceptance: "Please rate how comfortable you are with the system to use...". Distribution of answers on a 7-point Likert scale (left) and average participant's "path" through questions in each scenario (right). Each point is the mean per each question and scenario. Colours distinguish scenarios; symbols mark the level of risk in each scenario.

chances of getting a high comfort score on physiological measurements than those of low risk. The age variable revealed a reversed correlation with the score. With each year increase, holding other variables constant, the odds of the "comfort" score decrease by 10%.

Eye Tracking. Like in physiological measurements, for eye tracking, the Autonomous car scenario did not reveal significant differences in ORs for getting a high comfort score compared to the reference level (both are high-risk high-multitasking scenarios). The rest of the scenarios had significantly lower chances of getting a higher score. Ordinal regression analysis of the factors set confirmed the result. We estimated that the interaction of high-risk and high-multitasking levels, holding other variables constant, got a 171% higher chance of the eye tracking comfort score compared to low-risk and low-multitasking scenarios. We did not find significant OR differences for any recorded demographic factors.

Video recordings. Video recordings got the lowest comfort scores according to descriptive statistics. With the help of CLMM, we found that the Communications Officer, IT Programmer, and At Home (all low-risk context) scenarios got significantly lower ORs compared to the reference level (Air Traffic Controller). The "At Home" scenario obtained OR=0.13 (p<.001), making it again the scenario with the lowest odds of getting the high comfort score. Factors set variation manifested the work/private environment to be a significant factor in predicting the score in video recordings. Therefore, work scenarios have 2.55 higher odds of getting a high score on the scale. The unexpected result showed the model with the demographic data. Respondents with a place of residence with a population between 500,000- 1,000,000 inhabitants have almost 11 times higher chances (p<.001) of rating video recordings with high scores of comfort.

4.3 General AMS Features

The distribution of answers is shown in Figure 4 (left). The emergency option for the close ones and "Reminding where I left my task" got the biggest number of scores above neutral (88% and 86%, respectively). Interestingly, notification delays (both for work and private life) got the smallest number (48% and 52%). Still, all features have scores above neutral. That said, the overall value of the presented features is meaningful and should be taken into consideration in future system development. The bar chart with means and their confidence intervals (conf. level = 0.95) illustrates the potential differences in each feature assessment (Figure 4, right). The bar chart gives the first intuition regarding the preferences for valuable features.

By applying ordinal regression to compare features' values, we found no significant differences between the reference level "Reminding where I left my task" and "Emergency option for boss or colleagues". "Emergency option for close ones" has a significantly higher OR (2.2, p<.001), indicating more than two times bigger chances of getting a high-value score. The rest of the features have significantly lower ORs than the reference one, with the lowest for notification delay for work-related tasks (OR=0.12, p<.001). Further conducted a pairwise comparison with Tukey correction confirmed the first assumptions from the mean comparison illustrated in the bar chart (Figure 4, right): no significant differences between notifications delay for work-related tasks and from private life, no significant differences between "Reminding where I left my task" and "Emergency option for boss/colleagues", no significant differences between arranging the order of the notifications and emergency option for boss/colleagues, and no significant differences between emergency option for closed ones and boss/colleagues. At the same time, both features, "Reminding where I left my task" and "Emergency option for close ones", have the biggest ORs to get a high-value score, while notifications delay per se showed the smallest odds relative to the other features. See Table 8 for summarized results.

Demographic factors as features' value predictors. Lastly, we found interesting results on how demographic characteristics predict the final value score of different features. Thus, for the "Notifications delay from private life" feature, it is almost five times (OR=4.76, 95% CI [1.29, 17.9], p<.05) bigger chance of having a higher score value if the respondent has children than if he/she does not and does not plan to have. In the emergency option for private life, there is a slight but significant difference in odds with the changes in age. With each increasing year, the chances of valuing this option in the higher response category decrease by 8% (OR=0.92, 95% CI

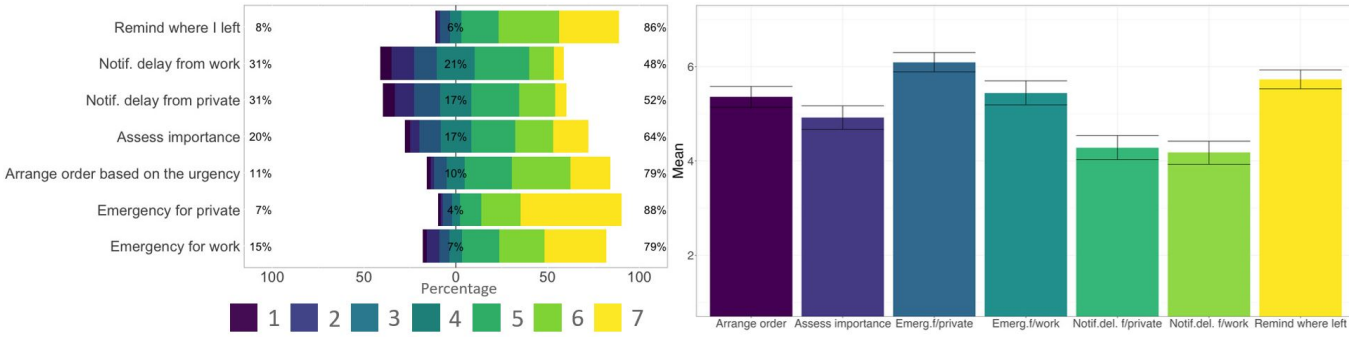

Figure 4: AMS Features: "Please rate the following features of the AMS based on their value for you." Left: Distribution of answers on a 7-point Likert scale. Right: bar chart with means per each feature and their confidence intervals (conf. level = 0.95)

Table 7: Descriptive Statistics for the AMS Helping Features.

Valued Features	M (SD)	Med (IQR)	Mode
Reminding where I left my task	5.73 (1.29)	6 (2)	6
Notif. delay for work-related tasks	4.18 (1.58)	4 (2)	5
Notif. delay from private life	4.28 (1.64)	5 (3)	5
Assessing the importance of the message	4.92 (1.6)	5 (2)	5
Arranging the order of the notif. based on the urgency/importance	5.36 (1.41)	6 (1)	6
Emergency option for your closed ones	6.09 (1.34)	7 (1)	7
Emergency option for your boss/ colleagues	5.44 (1.65)	6 (2)	7

Table 8: Ordinal Regression with CLMM for or Data Collection Methods Acceptance: "Please rate the following features of the AMS based on their value for you."

Questions		OR	95% CI
(ref.) Reminding where I left my task		—	—
Notif. delay for work-related tasks		**0.12*****	0.08, 0.18
Notif. delay from private life		**0.14*****	0.09, 0.22
Assessing the importance of the message		**0.32*****	0.21, 0.49
Arranging the order based on the importance		**0.54****	0.36, 0.82
Emergency option for your closed ones		**2.2*****	1.40, 3.46
Emergency option for your boss/colleagues		0.71	0.46, 1.10
Contrast (EM means pairwise comparisons)		Coef.	SE
Reminding where I left my task	Notif. delay for work-related tasks	**2.130*****	0.222
Reminding where I left my task	Notif. delay from private life	**1.961*****	0.221
Reminding where I left my task	Assessing the importance of the message	**1.130*****	0.217
Reminding where I left my task	Arranging the order based on the importance	0.614'	0.214
Reminding where I left my task	Emergency option for your closed ones	**-0.788***	0.231
Reminding where I left my task	Emergency option for your boss/colleagues	0.339	0.339
Notif. delay for work-related tasks	Notif. delay from private life	-0.169	0.206
Notif. delay for work-related tasks	Assessing the importance of the message	**-1.000*****	0.21
Notif. delay for work-related tasks	Arranging the order based on the importance	**-1.516*****	0.213
Notif. delay for work-related tasks	Emergency option for your closed ones	**-2.919*****	0.239
Notif. delay for work-related tasks	Emergency option for your boss/colleagues	**-1.792*****	0.224
Notif. delay from private life	Assessing the importance of the message	**-0.831****	0.211
Notif. delay from private life	Arranging the order based on the importance	**-1.347*****	0.213
Notif. delay from private life	Emergency option for your closed ones	**-2.750*****	0.239
Notif. delay from private life	Emergency option for your boss/colleagues	**-1.623*****	0.224
Assessing the importance of the message	Arranging the order based on the importance	-0.516	0.211
Assessing the importance of the message	Emergency option for your closed ones	**-0.918*****	0.233
Assessing the importance of the message	Emergency option for your boss/colleagues	**-0.791****	0.22
Arranging the order based on the importance	Emergency option for your closed ones	**-1.402*****	0.229
Arranging the order based on the importance	Emergency option for your boss/colleagues	-0.275	0.219
Emergency option for your closed ones	Emergency option for your boss/colleagues	1.127	0.236

[0.85, 0.98], p<.05). There were no other significant findings for demographic factors among all other features.

5 DISCUSSION

We have examined various aspects of designing an intelligent Attention Management System with a specific focus on delaying notifications based on the incoming messages, the user's cognitive load, and external context, obtained through data tracking methods. However, before implementing such a system, numerous (human) factors need to be considered. Our analysis revealed pivotal insights that inform the design of effective AMSs. Considering delays, there are different comfortable minimum/maximum thresholds for them. Critical and urgent notifications require considerably shorter delay times in both work and private settings compared to less time-sensitive notifications. Intriguingly, the likelihood of individuals desiring timely delivery of work-related notifications surpasses that of important notifications pertaining to private matters. Furthermore, the gap between the urgency of work-related critical notifications and routine private notifications is more pronounced than the reverse scenario. In other words, individuals are more inclined to tolerate slightly longer delays for important private messages than they are for non-critical work-related messages. This highlights the modern work-oriented tendency toward a nearly immediate response to critical work-related issues **RQ1b**. Beyond notification content, the context in which individuals engage with the AMS also plays a role. High-risk work environments correlate with an increased likelihood of tolerating notification delays, even for critical work notifications. Similarly, the work context alone increases the chances of accepting delays for important private messages by more than three times. Individual traits also play a role in appropriate delay times **RQ1a**. Notably, males have significantly lower chances of accepting longer delays for critical notifications in both work and private contexts. Additionally, parental status and aspirations influence delay tolerance for routine private messages; with increasing age, parents and individuals desiring parenthood are less likely to tolerate prolonged delays than those who do not plan to have children. We could try to explain it by the friends- and family-oriented nature of the former groups, while individuals with "child-free" views may have different priorities, for instance, careers where disturbance of routine messages may be less desirable **RQ1**.

To understand the suitable delay range, it is important for AMSs to know the external context as well as the urgency of the notification. More contextual data enhance prediction precision. Our findings in the "Data Collection Methods" section impart subtle nuances of the interaction of the variables. Users are mostly comfortable being tracked through operation system parameters and physiological measurements, while screen content recordings and eye tracking evoke slightly uncomfortable feelings, with video recordings at the bottom of the ranking with the most popular opinion that this method is completely uncomfortable. Nonetheless, contextual nuances can increase method acceptance. Thus, respondents have almost three times higher chances of tolerating screen content recordings and 2.55 video recordings when collected in a work context. There is also a 171% higher chance of tolerating eye-tracking surveillance if it is needed to navigate in high-risk, high-multitasking situations. While the home environment emerges as the least comfortable for applying any data collection methods **RQ2**. Regarding valuable AMS features, our findings indicate positive ratings for most presented features and should be considered in the development. The highest-rated features include an emergency option for close contacts and hints about where the task was left. However, the assessment of the notification delay function is only slightly above neutral, indicating that its value could be intertwined with additional features **RQ3**.

In summary, our study provides insights crucial for the development of an AMS centered on notification delays. We emphasize the significance of considering the context of the future application and the problems the AMS should help navigate. Smartly choosing accompanying features in the AMS design, such as appropriate adjustable delay times and toleratable data collection methods, anticipates enhancing user experience and future acceptance of the technology.

5.1 Limitations & Future Work

Although our study provides valuable findings about people's opinions on the characteristics of AMS, some limitations deserve recognition. The study declared as a vignette method has a form different from the classical one (combination of factors) and has elements of a story used in qualitative methods studies, which gives much more freedom in interpretations and possible hidden variables. The study's findings are contingent on the specific scenarios and contexts examined. Broader generalizations should be made with caution, considering potential variations in user preferences and requirements across diverse real-world situations. Also, we focused on a particular combination of demographic factors for our analysis, and other factors (such as users' occupations or professional backgrounds related to the scenarios) should be addressed in future studies. Finally, the vignette method per se has its limitations. The study primarily focuses on user preferences and perceptions without directly testing AMS in a dynamic context. Further research could delve into close-to-real user experiences (see [28] for "do not disturb challenge" in the real office). Additionally, a comprehensive qualitative exploration of human opinion - especially considering human personality traits and attitudes - would be of great value.

5.2 Summary and Recommendations

Based on the findings elucidated in our study, we present practical recommendations that can guide the development of effective and user-friendly Attention Management Systems centered around notification delays:

- **Customizable Delay Thresholds**: Recognize the variability in comfortable delay times for different types of notifications and contexts. Our results indicate that many users accept longer delays for private messages than for job-related ones, especially in the working environment. High-risk working notifications should be timed with particular caution, and people tend to be highly wary of postponing working notifications at high-paced jobs. Allow users to customize delay settings according to their specific preferences.
- **Account for Individual Traits** when determining delay times, features, and data collection methods. Take into account demographic factors such as gender, family status, age,

and other parameters. That is, our study revealed that men are less tolerant of notification delays, especially if messages are related to work. Additionally, with increasing age, people tend to accept longer delays in private messages.

- **Data Collection Methods**: In private environments, focus on less intrusive methods like operation system parameters and aim for a compromise between privacy and system performance. Also, physiological measurements from devices such as smartwatches are widely accepted. Excessive surveillance (for example, including video feeds of the screen contents or the user interacting with the system) should (depending on the particular context) be limited to high-risk and/or working environments.
- **Feature Emphasis**: Prioritize features that resonate most with users. Recognize that notification delay per se is perceived as more valuable when bundled with additional features that enhance task management and user control. For example, develop emergency options for close contacts to bring back control to the user, and aim at developing useful resumption cues to ease resuming a before-interrupted task.
- **User-Centric Design**: Align the AMS design with the context and challenges users encounter. Emphasise features that address specific user needs, ensuring the system not only adapts to users' mental load but also assists them in managing interruptions effectively. For example, the feature "reminding where I left my task" was highly valued in our feature evaluation. Conduct thorough usability testing to evaluate user interactions with the AMS. Gather user feedback and insights to continuously refine the system and ensure its effectiveness in real-world scenarios.

ACKNOWLEDGMENTS

This project is supported by the Austrian Science Fund (FWF) under grant Nr.P35976-N

REFERENCES

[1] Piotr D. Adamczyk and Brian P. Bailey. 2004. If Not Now, when?: The Effects of Interruption at Different Moments Within Task Execution. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Vienna, Austria) *(CHI '04)*. ACM, New York, NY, USA, 271–278. https://doi.org/10.1145/985692.985727

[2] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. 2019. Guidelines for Human-AI Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3290605.3300233

[3] Christoph Anderson, Isabel Hübener, Ann-Kathrin Seipp, Sandra Ohly, Klaus David, and Veljko Pejovic. 2018. A Survey of Attention Management Systems in Ubiquitous Computing Environments. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 2, Article 58 (jul 2018), 27 pages. https://doi.org/10.1145/3214261

[4] Christiane Atzmüller and Peter M. Steiner. 2010. Experimental Vignette Studies in Survey Research. *Methodology* 6, 3 (2010), 128–138. https://doi.org/10.1027/1614-2241/a000014 arXiv:https://doi.org/10.1027/1614-2241/a000014

[5] Thom Baguley, Grace Dunham, and Oonagh Steer. 2022. Statistical modelling of vignette data in psychology. *British Journal of Psychology* 113, 4 (2022), 1143–1163.

[6] Brian P. Bailey and Joseph A. Konstan. 2006. On the need for attention-aware systems: Measuring effects of interruption on task performance, error rate, and affective state. *Computers in Human Behavior* 22, 4 (2006), 685 – 708. https://doi.org/10.1016/j.chb.2005.12.009 Attention aware systems.

[7] Peter Bogunovich and Dario Salvucci. 2011. The effects of time constraints on user behavior for deferrable interruptions. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 3123–3126.

[8] JP Borst and NA Taatgen. 2007. The costs of multitasking in threaded cognition. , 133–138 pages.

[9] Jelmer P Borst, Niels A Taatgen, and Hedderik Van Rijn. 2010. The problem state: a cognitive bottleneck in multitasking. *Journal of Experimental Psychology: Learning, memory, and cognition* 36, 2 (2010), 363.

[10] 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*. 2971–2980.

[11] Caroline Bradbury-Jones, Julie Taylor, and OR Herber. 2014. Vignette development and administration: a framework for protecting research participants. *International Journal of Social Research Methodology* 17, 4 (2014), 427–440.

[12] David M Cades, Deborah A Boehm Davis, J Gregory Trafton, and Christopher A Monk. 2007. Does the difficulty of an interruption affect our ability to resume?. In *Proceedings of the human factors and ergonomics society annual meeting*, Vol. 51. SAGE Publications Sage CA: Los Angeles, CA, 234–238.

[13] R. H. B. Christensen. 2022. ordinal—Regression Models for Ordinal Data. R package version 2022.11-16. https://CRAN.R-project.org/package=ordinal.

[14] Jon D Elhai, Jason C Levine, Robert D Dvorak, and Brian J Hall. 2016. Fear of missing out, need for touch, anxiety and depression are related to problematic smartphone use. *Computers in Human Behavior* 63 (2016), 509–516.

[15] Spencer C Evans, Michael C Roberts, Jared W Keeley, Jennifer B Blossom, Christina M Amaro, Andrea M Garcia, Cathleen Odar Stough, Kimberly S Canter, Rebeca Robles, and Geoffrey M Reed. 2015. Vignette methodologies for studying clinicians' decision-making: Validity, utility, and application in ICD-11 field studies. *International journal of clinical and health psychology* 15, 2 (2015), 160–170.

[16] Janet Finch. 1987. The vignette technique in survey research. *Sociology* 21, 1 (1987), 105–114.

[17] Joel E. Fischer, Chris Greenhalgh, and Steve Benford. 2011. Investigating Episodes of Mobile Phone Activity as Indicators of Opportune Moments to Deliver Notifications. In *Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services* (Stockholm, Sweden) *(MobileHCI '11)*. Association for Computing Machinery, New York, NY, USA, 181–190. https://doi.org/10.1145/2037373.2037402

[18] Helen M Hodgetts and Dylan M Jones. 2006. Interruption of the Tower of London task: support for a goal-activation approach. *Journal of Experimental Psychology: General* 135, 1 (2006), 103.

[19] Eric Horvitz. 1999. Principles of Mixed-initiative User Interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Pittsburgh, Pennsylvania, USA) *(CHI '99)*. ACM, New York, NY, USA, 159–166. https://doi.org/10.1145/302979.303030

[20] ECMCE Horvitz. 2001. Notification, disruption, and memory: Effects of messaging interruptions on memory and performance. In *Human-Computer Interaction: INTERACT*, Vol. 1. 263.

[21] Shamsi T. Iqbal and Brian P. Bailey. 2005. Investigating the Effectiveness of Mental Workload As a Predictor of Opportune Moments for Interruption. In *CHI '05 Extended Abstracts on Human Factors in Computing Systems* (Portland, OR, USA) *(CHI EA '05)*. ACM, New York, NY, USA, 1489–1492. https://doi.org/10.1145/1056808.1056948

[22] Shamsi T. Iqbal and Eric Horvitz. 2010. Notifications and Awareness: A Field Study of Alert Usage and Preferences. In *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work* (Savannah, Georgia, USA) *(CSCW '10)*. Association for Computing Machinery, New York, NY, USA, 27–30. https://doi.org/10.1145/1718918.1718926

[23] 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*. 107–110.

[24] Anouk Mols and Jason Pridmore. 2021. Always available via WhatsApp: Mapping everyday boundary work practices and privacy negotiations. *Mobile Media & Communication* 9, 3 (2021), 422–440. https://doi.org/10.1177/2050157920970582 arXiv:https://doi.org/10.1177/2050157920970582

[25] Alexander Neff and Philipp Wintersberger. 2022. An Experience Sampling Study to Evaluate Why Users Dismiss Smartphone Notifications. In *Adjunct Publication of the 24th International Conference on Human-Computer Interaction with Mobile Devices and Services* (Vancouver, BC, Canada) *(MobileHCI '22)*. Association for Computing Machinery, New York, NY, USA, Article 18, 5 pages. https://doi.org/10.1145/3528575.3551446

[26] Tadashi Okoshi, Julian Ramos, Hiroki Nozaki, Jin Nakazawa, Anind K. Dey, and Hideyuki Tokuda. 2015. Reducing Users' Perceived Mental Effort Due to Interruptive Notifications in Multi-Device Mobile Environments. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (Osaka, Japan) *(UbiComp '15)*. Association for Computing Machinery, New York, NY, USA, 475–486. https://doi.org/10.1145/2750858.2807517

[27] Antti Oulasvirta and Pertti Saariluoma. 2006. Surviving task interruptions: Investigating the implications of long-term working memory theory. *International Journal of Human-Computer Studies* 64, 10 (Oct. 2006), 941–961. https://doi.org/10.1016/j.ijhcs.2006.04.006

[28] Martin Pielot and Luz Rello. 2015. The Do Not Disturb Challenge: A Day Without Notifications. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI EA '15)*. Association for Computing Machinery, New York, NY, USA, 1761–1766. https://doi.org/10.1145/2702613.2732704

[29] Martin Pielot, Amalia Vradi, and Souneil Park. 2018. 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* (Barcelona, Spain) *(MobileHCI '18)*. Association for Computing Machinery, New York, NY, USA, Article 3, 11 pages. https://doi.org/10.1145/3229434.3229445

[30] Benjamin Poppinga, Wilko Heuten, and Susanne Boll. 2014. Sensor-Based Identification of Opportune Moments for Triggering Notifications. *IEEE Pervasive Computing* 13, 1 (2014), 22–29. https://doi.org/10.1109/MPRV.2014.15

[31] Claudia Roda and Julie Thomas. 2006. Attention aware systems: Theories, applications, and research agenda. *Computers in Human Behavior* 22, 4 (2006), 557–587.

[32] Dario D. Salvucci and Peter Bogunovich. 2010. Multitasking and Monotasking: The Effects of Mental Workload on Deferred Task Interruptions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Atlanta, Georgia, USA) *(CHI '10)*. Association for Computing Machinery, New York, NY, USA, 85–88. https://doi.org/10.1145/1753326.1753340

[33] Dario D Salvucci and Niels A Taatgen. 2008. Threaded cognition: an integrated theory of concurrent multitasking. *Psychological review* 115, 1 (2008), 101.

[34] Dario D. Salvucci, Niels A. Taatgen, and Jelmer P. Borst. 2009. Toward a Unified Theory of the Multitasking Continuum: From Concurrent Performance to Task Switching, Interruption, and Resumption. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Boston, MA, USA) *(CHI '09)*. Association for Computing Machinery, New York, NY, USA, 1819–1828. https://doi.org/10.1145/1518701.1518981

[35] Christina Schneegass and Fiona Draxler. 2021. Designing Task Resumption Cues for Interruptions in Mobile Learning Scenarios. In *Technology-Augmented Perception and Cognition*, Evangelos Niforatos and Tilman Dingler (Eds.). Springer International Publishing, Cham, 125–181. https://doi.org/10.1007/978-3-030-30457-7_5 Series Title: Human–Computer Interaction Series.

[36] Hawal Shamon and Carl Berning. 2019. Attention check items and instructions in online surveys with incentivized and non-incentivized samples: Boon or bane for data quality? , 55–77 pages.

[37] Cheri Speier, Joseph S Valacich, and Iris Vessey. 1999. The influence of task interruption on individual decision making: An information overload perspective. *Decision sciences* 30, 2 (1999), 337–360.

[38] Namrata Srivastava, Rajiv Jain, Jennifer Healey, Zoya Bylinskii, and Tilman Dingler. 2021. Mitigating the Effects of Reading Interruptions by Providing Reviews and Previews. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI EA '21)*. Association for Computing Machinery, New York, NY, USA, Article 229, 6 pages. https://doi.org/10.1145/3411763.3451610

[39] Fabio Sticca and Sonja Perren. 2013. Is cyberbullying worse than traditional bullying? Examining the differential roles of medium, publicity, and anonymity for the perceived severity of bullying. *Journal of youth and adolescence* 42 (2013), 739–750.

[40] Roel Vertegaal et al. 2003. Attentive user interfaces. *Commun. ACM* 46, 3 (2003), 30–33.

[41] Peter Vorderer, Dorothée Hefner, Leonard Reinecke, and Christoph Klimmt. 2017. Permanently online, permanently connected: Living and communicating in a POPC world.

[42] Mark Weiser. 1991. The Computer for the 21 st Century. *Scientific american* 265, 3 (1991), 94–105.

[43] Tom Wilks. 2004. The use of vignettes in qualitative research into social work values. *Qualitative social work* 3, 1 (2004), 78–87.

[44] Philipp Wintersberger, Clemens Schartmüller, and Andreas Riener. 2019. Attentive user interfaces to improve multitasking and take-over performance in automated driving: the auto-net of things. *International Journal of Mobile Human Computer Interaction (IJMHCI)* 11, 3 (2019), 40–58.

[45] Glenn Wylie and Alan Allport. 2000. Task switching and the measurement of "switch costs". *Psychological research* 63, 3 (2000), 212–233.

[46] Manuela Züger, Christopher Corley, André N. Meyer, Boyang Li, Thomas Fritz, David Shepherd, Vinay Augustine, Patrick Francis, Nicholas Kraft, and Will Snipes. 2017. Reducing Interruptions at Work: A Large-Scale Field Study of FlowLight. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 61–72. https://doi.org/10.1145/3025453.3025662

[47] Manuela Züger and Thomas Fritz. 2018. Sensing and Supporting Software Developers' Focus. In *Proceedings of the 26th Conference on Program Comprehension* (Gothenburg, Sweden) *(ICPC '18)*. Association for Computing Machinery, New York, NY, USA, 2–6. https://doi.org/10.1145/3196321.3196323

[48] 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* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3173574.3174165