Exploring User Expectations of Proactive AI Systems

CHRISTIAN MEURISCH, Technical University of Darmstadt, Germany
CRISTINA A. MIHALE-WILSON, Goethe University Frankfurt, Germany
ADRIAN HAWLITSCHEK, Technical University of Darmstadt, Germany
FLORIAN GIGER, Technical University of Darmstadt, Germany
FLORIAN MÜLLER, Technical University of Darmstadt, Germany
OLIVER HINZ, Goethe University Frankfurt, Germany
MAX MÜHLHÄUSER, Technical University of Darmstadt, Germany

Recent advances in artificial intelligence (AI) enabled digital assistants to evolve towards proactive user support. However, expectations as to when and to what extent assistants should take the initiative are still unclear; discrepancies to the actual system behavior might negatively affect user acceptance. In this paper, we present an *in-the-wild* study for exploring user expectations of such user-supporting AI systems in terms of different proactivity levels and use cases. We collected 3,168 in-situ responses from 272 participants through a mixed method of automated user tracking and context-triggered surveying. Using a data-driven approach, we gain insights into initial expectations and how they depend on different human factors and contexts. Our insights can help to design AI systems with varying degree of proactivity and preset to meet individual expectations.

CCS Concepts: • **Human-centered computing → Empirical studies in HCI**; *Field studies*; User models.

Additional Key Words and Phrases: proactivity, personal assistants, artificial intelligence, privacy demands

ACM Reference Format:
Christian Meurisch, Cristina A. Mihale-Wilson, Adrian Hawlitschek, Florian Giger, Florian Müller, Oliver Hinz, and Max Mühlhäuser. 2020. Exploring User Expectations of Proactive AI Systems. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 4, 4, Article 146 (December 2020), 22 pages. https://doi.org/10.1145/3432193

1 INTRODUCTION

In recent years, digital assistants have made the transition from research prototypes to consumer products, entering our lives in the form of Apple Siri (2011) or Amazon Alexa (2015) [35]. From such voice-controlled systems, emerging assistants are increasingly evolving into AI-powered systems that base their actions on predictive models to anticipate user needs [68], proactively intervene [57], or complete tasks without users' explicit request [65]. However, latest studies [7, 12, 16, 19, 35] have revealed that user expectations of such AI

Authors' addresses: Christian Meurisch, Technical University of Darmstadt, Hochschulstr. 10, D-64295 Darmstadt, Germany, meurisch@tk.tu-darmstadt.de; Cristina A. Mihale-Wilson, Goethe University Frankfurt, Theodor-W.-Adorno-Platz 4, D-60323 Frankfurt am Main, Germany, mihale-wilson@wiwi.uni-frankfurt.de; Adrian Hawlitschek, Technical University of Darmstadt, Germany; Florian Giger, Technical University of Darmstadt, Germany, giger@tk.tu-darmstadt.de; Florian Müller, Technical University of Darmstadt, Germany, mueller@tk.tu-darmstadt.de; Oliver Hinz, Goethe University Frankfurt, Germany, hinz@wiwi.uni-frankfurt.de; Max Mühlhäuser, Technical University of Darmstadt, Germany, max@tk.tu-darmstadt.de.

2474-9567/2020/12-ART146 $15.00
https://doi.org/10.1145/3432193

systems are dramatically out of step with the actual system operation—leading to low user acceptance and low adoption of this technology [46].

Prior work on bridging such gaps has mainly focused on methods for shaping appropriate initial expectations to better meet the system operation [29, 30, 52]. By contrast, only a few works pursue the other way of automatically setting system operation at runtime to meet an individual's initial expectations for different use cases and what users want to use [60]. In particular, expectations as to when and to what extent assistants should take the initiative are not taken into account, as they are still unclear.

In the light of this deficit, we present an *in-the-wild* study to understand user expectations of such AI systems that can operate on different levels of proactivity. Compared to an online survey or a vignette study, an in-the-wild approach yields more reliable results: participants do not have to answer questions about fully-hypothetical situations but experience the real situation in which the questions fit.

More precisely, we have designed and developed *ProfileMe*—a mobile application implementing a new study method that situates the participants in the likely context of use to get more realistic answers about users' (initial) expectations towards proactive AI systems and to uncover human factors. Using ProfileMe, we collected a mix of objective and subjective data from 272 participants. Besides the in-situ acquired preferences of the participants with regard to the proactivity of AI systems, this data includes demographic data, personality tests, and daily summaries of user behavior. Based on this data, we examine and quantify how users want to use which kind of AI systems in specific physical and cognitive contexts. We revealed significant differences between users' expectations in various areas in which a user can be supported: participants are generally very open to proactive support; only the subarea of *mental health support*, where most users tend to prefer reactive support or no support at all, is the exception. We further found that users who tend to reject proactive AI systems are mainly afraid that their privacy could be violated or that such systems will be too intrusive.

All in all, the contribution of this paper is twofold. *Firstly*, we contribute to the current research on the design of AI systems by presenting – to the best of our knowledge – the first *in-the-wild* study that examines proactivity preferences of 272 participants for several use cases. *Secondly*, we gain new informative insights into user expectations of proactive AI systems and how they are influenced by different human factors.

2 TERMINOLOGY AND CLASSIFICATION OF AI SYSTEMS

The term 'AI' is used today ambiguously in various fields such as *human imitation*—in which the term was historically coined—, *machine learning*, or *intelligent infrastructures* to describe a web of computation, data, and physical entities that makes human environments more supportive [27]. In this paper, we use the term 'AI systems' to subsume all systems that base their actions on AI models to support a user. For systematic research, we now provide a classification of AI systems according to their degree of proactivity and their area of application.

2.1 Proactivity Levels

AI systems operate on the proactivity continuum ranging from zero to full initiative automation to support users [65]. We provide an adapted and refined categorization for AI systems along this continuum, more precisely, based on the following three classes (*reactive–proactive–autonomous*):

- *Reactive support.* Reactive AI systems only *respond to commands from users* in order to support them, currently representing the most widespread class of AI systems. Commands can be entered through different input modalities (e.g., voice, gesture). Probably, the most prominent representatives of this class are conversational agents[1], which have made the transition from research prototypes to consumer products in the current decade: Apple Siri (2011), Google Assistant (former: Google Now, 2012), Microsoft Cortana

[1]The design space of conversational agents and fine-grained terminological distinctions can be found in [24].

(2015), Amazon Alexa (2015), to name a few [35]. With such AI systems, users can trigger information searches or simple task executions (e.g., finding the fastest route) [17].

- *Proactive support.* From this class onwards, AI systems *proactively take actions without explicit user request.* We further divide this class into the following two levels: (1) AI systems *inform users what to pay attention to*, e.g., using notifications or alerts. For instance, location-aware services in this class can send updates to weather and traffic conditions, among other things, that may be relevant for users in their current situation [65]. (2) AI systems *make personalized recommendations*, i.e., they additionally interpret the outcome of the underlying AI models with respect to user needs [68] and, if necessary, suggest concrete actions to meet these needs [57]. For instance, fitness and health applications in this class automatically learn a user's physical activity to strategically suggest changes to this behavior (e.g., "take a break" or "walk for 30 minutes") for a healthier lifestyle [60] or meeting specific fitness goals [66]. In general, the actual actions must still be confirmed or taken by the users. To engage users in these actions, latest AI systems rely on predictive models to find opportune moments to notify [38, 59] and intervene users [57].
- *Autonomous Support.* AI systems in this class *act autonomously, making decisions and taking the actual actions on behalf of users without confirmation.* For instance, such AI systems enter our lives in the form of digital agents [74], (social) robots [58], or (un-)manned vehicles (e.g., drones [21]). Even though preliminary work and first prototypes exist, this class still holds many open challenges to achieve its breakthrough. In the coming decades, this class will increasingly become the focus of research, shaping our lives the most.

In our study, we use this proactivity classification of AI systems to investigate and quantify the factors that influence user expectations of such systems – the study design is described in more detail in Section 4.

2.2 Application Domains and Examples

Next, we identified three relevant—somewhat overlapping—domains/ areas of application that (i) cover most of the user's daily life, (ii) in which AI system provide direct support to users, and (iii) which are of current interest for research. If applicable, some domains are further divided into subareas. For each domain, we name representative application examples of AI systems that are either highly cited or widely known in the community.

2.2.1 Healthcare & Well-being. Well-being refers to "diverse and interconnected dimensions of physical, mental, and social well-being that extend beyond the traditional definition of health" [50]. Along this definition, we further categorize this domain into the following three subareas:

Physical health covers all 'visible' states and body conditions of an individual, "taking into consideration everything from the absence of disease to fitness level."[2] State-of-the-art AI systems can detect and analyze the user state with respect to specific factors, and intervene or make recommendations if either a worsening is imminent or an improvement is to be achieved – many of such works can be found in the related discipline of digital behavior change interventions [33]. For instance, `MyBehavior` automatically learns a user's physical activity and dietary habits to strategically suggest changes to this behavior for a healthier lifestyle [60]. Schmidt et al. present a digital coach that automatically identifies the user's strengths and weaknesses; on that basis, it creates appropriate training plans to motivate and help a user in achieving his fitness goals [66].

Mental health covers all 'invisible' states of an individual, including "subjective well-being, perceived self-efficacy, autonomy, competence, inter-generational dependence, and self-actualization of one's intellectual and emotional potential, among others" – according to the World Health Organization (WHO) [53]. This also includes coping with stress and unproductive work. In literature, there are several AI systems that support these factors. For instance, `InterruptMe` represents an AI system that automatically infers opportune moments for interruption [56]. Pielot et al. identify opportune moments where users are particularly open to engaging with

[2]https://www.eupati.eu/glossary/physical-health (retrieved 10/28/2020)

suggested content [59]. Other AI systems can detect and monitor stress [34], depressive states [13], emotions [62], to name a few.

Social well-being covers all social dimensions, including social acceptance, integration, and interaction with others [69]. For instance, `SociableSense` models the 'sociability' of users based on their co-location and interaction patterns; it provides users with real-time feedback to foster and improve social interactions [61]. Similarly, `BeWell` calculates well-being scores to raise a user's awareness, helping users to manage their overall well-being [32].

2.2.2 Activity Support. AI systems can support users in activities, or even take them over to achieve the desired result. We categorize this kind of support into the following two subareas [43]:

Support for physical activities covers all physical actions of an individual, e.g., movements, transportation, or cooking. For instance, Google Now supports users to organize their daily lives by suggesting when to leave places, finding the fastest routes and best transportation [57]. Other AI systems can also support specific professional activities, such as conducting experiments in a laboratory [67], or detecting nurse activities in a hospital [26].

Support for digital activities covers all actions related to an individual's digital world. Probably the best-known example of this area is email spam filtering—concerning the processing of emails to organize them according to specified criteria [10]. `PrefMiner` is an example of how to manage mobile notifications so that users are not disturbed. [36]. Other AI systems help users, e.g., by rescheduling appointments or organizing tasks [74].

2.2.3 Ambient Intelligence. Last but not least, ambient intelligence is the support of users with and from the physical environment [49, 63, 73]. AI systems in this class can, for instance, automatically regulate the light and climate conditions based on user preferences [18] or support users in doing the laundry [15]. In smart car environments, such AI systems can support the driver in various situations—e.g., parking aid, collision avoidance, drowsiness detection—or by infotainment assistance [47].

3 RELATED WORK: EXPECTATION RESEARCH

User expectations *"represent an individual's prediction or anticipatory judgment about what they should or will receive through the performance of a product or service"* [12]. Such expectations have been explored in various fields including marketing [3], information systems [9], and recently in HCI research [46] with a focus on AI [29].

One prominent theory used in user expectation research are the *Expectation Confirmation Theory* (ECT) [51] and its extensions [9, 12]. ECT postulates that user satisfaction and acceptance of a system are directly related to the difference between initial user expectation and the perceived system performance. The latter represents not only the subjective assessment of the users but also their actual experience with the system [29]. In other words, a negative confirmation of users' initial expectations leads to lower user satisfaction, which in turn reduces user acceptance of the system. Based on this theoretical consideration, marketing studies have focused on product expectations and users' willingness to pay [48], while ubiquitous computing and HCI studies largely investigate the relationship between user experience and expectation [31, 46].

In the context of AI systems, latest HCI studies [7, 11, 12, 16, 19, 28, 35] have revealed that even users' expectations of reactive systems are dramatically out of step with the actual system operation. According to ECT, these discrepancies lead to low user acceptance and ultimately to the rejection of such AI systems [46]. Moving towards proactivity, these discrepancies may further increase the fear of losing control [71], which users already feel when using more autonomous systems [5].

Prior works aiming to lower discrepancies between user expectations and system operation have mainly focused on forming appropriate initial user expectations to better meet the system operation [25, 29, 30]. While some of these works are based on deceptive approaches, namely *priming* and *framing* [25], other approaches explore the expectations of crucial properties of AI-based systems. For instance, Kocielnik et al. propose techniques for shaping expectations of AI-powered technologies prior to use. These techniques range from simple indication

Table 1. Description of our *use cases* or scenarios in which AI systems can provide user support

Area	Sub-area	Use cases
Healthcare & Well-being *(12)*	*physical*	(#1) sleepiness, (#2) fitness plan, (#3) medication intake, (#4) physical complaints
	mental	(#5) exhaustion/stress, (#6) concentration, (#7) loneliness, (#8) boredom/addiction
	social	(#9) appointments, (#10) annual reminders, (#11) reservations, (#12) postponements
Activity Support *(6)*	*physical*	(#13) public transportation, (#14) travel planning, (#15) shopping
	digital	(#16) finance, (#17) emails, (#18) *do not disturb* handling
Ambient Intelligence *(5)*	–	(#19) lighting and climate control, (#20) housekeeping, (#21) laundry management, (#22) reorders, (#23) assignments

of system accuracy over example-based explanations to direct control of system performance [29]. In line with the latter, Alan et al. conclude that such systems should be designed in a way that allows users to adjust the degree of autonomy of the systems [1]; the authors further show that the user's decision, once made, on the degree of autonomy does not necessarily have to be constant over time [2]. Although these studies showed the effectiveness of expectation-shaping and control transfer techniques, they are each limited to a single use case, not exploring different ones. User expectations about when and to what extent AI systems should be proactive have not been considered at all.

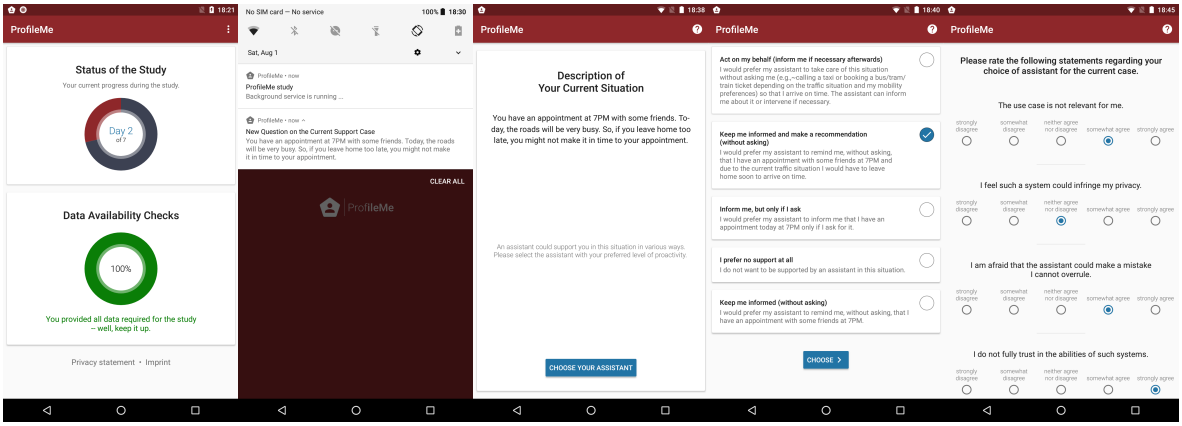
4 CAPTURING USER EXPECTATIONS OF PROACTIVE AI SYSTEMS

To investigate and quantify the factors that shape user expectations for proactive AI systems in different physical and cognitive situations, we decided to conduct an *in-the-wild* study. In the following, we describe and explain the study design, our data collection tool, and the study procedure. Finally, we give an insight into the recruitment process of the participants and the mechanisms for data protection compliance.

4.1 Study Design

User expectations can be very different and even unrealistic. For instance, users with little or no technical understanding may have the same expectations of a (voice-controlled) digital assistant as of a human assistant due to the embodiment and anthropomorphism of them [35]. On the other hand, user expectations can be quite difficult to express, especially when asking users without direct context. For instance, experienced managers can even express their initial expectations about when and to what extent their human assistants should take the initiative much better and more precisely when they are in the appropriate situation and feel the need for action. In order to cope with such circumstances, capturing user expectations of proactive AI services requires a more complex study design that cannot rely solely on online surveys and interviews.

Based on these considerations, we decided to conduct an in-situ study, as we expect that users can better assess their expectations of AI systems when they are in the physical situations in which the respective AI systems provide their support and in which the users are influenced by various aspects of the user context. To this end, we propose a new study method in the context of research on proactive AI systems: it comprises *automated user tracking* (to collect user data and identify specific situations of interest) and *context-triggered surveying* (to capture users' expectations while they are engaged in their daily activities). We further hypothesize that our method achieves more realistic answers than conventional online surveys, as the latter only describe both the hypothetical use case and the corresponding context. In contrast, in our method, the likely context of use (i.e., the situation in which the AI system provides its support) is experienced, while only the corresponding hypothetical use case is described there.



(a) Overview (b) In-situ Notification (c) UC Description (d) Proactivity Rating (e) Add'l. Questions

Fig. 1. Study tool: the *ProfileMe* application

Table 1 shows the 23 use cases or possible scenarios for user support by AI systems that are examined in our study. We derived all use cases from the existing literature and grouped them based on the three domains and their specific sub-areas of support identified in Section 2.2. Within our study and to reduce recall biases, which impact traditional self-reported measures [55], we then trigger questionnaires on these use cases in certain situations where an AI system can support the users in this way. More precisely, we developed a mobile application that monitors and records a user's context to identify specific situations and behavior (see Section 4.2). We augment this observational data with self-reported data from in-situ questionnaires that ask the users whether and to what extent AI systems should take the initiative to support them. To quantify this, we use the four proactivity levels introduced in Section 2.1—namely *reactive support*, *proactive support I*, *proactive support II*, and *autonomous support*—as well as an alternative choice if *no support* is desired.

In concrete terms, the participants first see a general description of the use case that fits their current situation. We then elicit the individual expectations of a user by letting him choose between different AI systems, each of which operates at a different level of the above-mentioned proactivity scale. We further aim to understand the user's decision by letting them provide us with more information related to the relevance and importance of the use case, their privacy concerns, trust levels, perceived loss of control, intrusiveness and self-determination. To additionally investigate the potential role of individual human factors and personality traits on user expectation of proactive AI systems, we are also gathering socio-demographics and personality traits from users.

4.2 Data Collection Tool & Study Procedure

We have developed an Android mobile application—namely *ProfileMe*—especially for this study to collect the necessary data (see Figure 1). This data is collected by the app in three ways: (1) a questionnaire in the setup phase of the study, (2) automated user tracking during the study, and (3) context-triggered surveying.

4.2.1 Questionnaires in the Setup Phase. In the setup phase of the study, the participants had to install the app from the Google Play store[3]. After installing the app on their Android mobile phones, the participants had to provide some *socio-demographic information* including their age, gender, country of residence, marital status, level of education, and employment status. Further, participants were asked to fill out the *44-item Big Five Inventory*

[3]The *ProfileMe* app was published in the Google Play store within a closed beta testing program.

(BFI) *personality* test [23], and report their *experience with digital assistants*—i.e., which assistants a user has already used and to what extent. It is important to note that we always use the term 'assistant' instead of 'AI system' when surveying the participants, as the former is already known to users and is, therefore, easier to communicate; the latter is more abstract and is used more from the technical perspective. The information gathered in the setup phase of the study forms the basis for the characterization of the participants.

4.2.2 Automated User Tracking. After the setup phase, the app directs the participants to the overview page (see Figure 1a), on which the user can see the progress of the study (how long the study will last) and data availability checks (whether all sensor data required for the study has been provided – otherwise, the user is guided how to improve this, e.g., by turning on the corresponding sensors.). In the background, the ProfileMe app also starts to continuously gather sensor data from the participants' mobile devices to (i) detect specific situations and trigger corresponding surveys, and (ii) create daily summaries of the participants' mobile and online behavior. The daily summaries are calculated locally at midnight; they include the participants' movements and mobility patterns (including their physical activity level, number of steps, places visited), and their smartphone and app interactions for that day. Additionally, the ProfileMe app captures the notification handling for our context-driven surveys. This tracking is based on the `Labels` and `Kraken.me` platforms – for more technical details see [45].

4.2.3 Context-triggered Surveying. Aiming to capture the situational preferences of participants for proactive AI support, the ProfileMe application was developed to identify their context and then trigger appropriate surveys. To this end, we developed five groups of context triggers based on *time*, *location*, *activity*, *app usage*, and *wake-up events*—relying as far as possible on the established (energy-efficient) methods of the Android OS and otherwise using state-of-the-art methods from literature.

In detail, the app runs a background service to monitor and process the requested sensor data. For instance, the app uses the OS-provided scheduling features for the straightforward *time* triggers.

For the more complex *location* triggers, the app must first identify the meaningful places of a user, namely *home* and *work*. To this end, the app mostly relies on the place detection algorithm by [37], as described below. The app filters inaccurate location values (horizontal accuracy > 50 meters) obtained from Android's fused location provider[4] to ensure a better quality of the identified places. Based on the user's speed calculated from two successive location values, the location readings are grouped into those where the user was in motion and those where the user was not. The former are marked as 'en route'; the latter are used for the clustering approach to identifying the user's (stationary) places. The resulting clusters are labeled according to [70]: *home* is the cluster in which a user spends most of the night and early morning hours; *work* is the cluster in which a user spends most of the day hours; any other cluster is labeled as 'other' – for further (general) place type identification, we used Google's Place Types[5]. It is important to note that the app has a short cold start phase of two working days only for these trigger criteria (and gradually becomes more reliable over time). Once these places are identified, the app uses the OS-provided geofencing features[6] to receive entry and exit events for these locations.

For the *activity* triggers, the app relies on the OS-provided activity recognition feature[7], which high-frequently outputs a list of probable activities, each with a confidence value. Inspired by [77], we use a Markov model to smooth these time series and obtain higher-quality information about the user's actual (physical) activity.

For the *app usage* triggers, ProfileMe receives events from the OS whenever a change in app usage (e.g., launching new apps, re-opening an already used app, and killing old apps) occurs. By comparing two consecutive events, we can further determine the usage per application [76].

[4]https://developers.google.com/location-context/fused-location-provider (retrieved 10/28/2020)
[5]https://developers.google.com/places/android-sdk/overview (retrieved 10/28/2020)
[6]https://developers.google.com/location-context/geofencing (retrieved 10/28/2020)
[7]https://developers.google.com/location-context/activity-recognition (retrieved 10/28/2020)

Table 2. Description of the developed *context triggers*, which can be combined depending on the target use case

Group	Possible trigger criteria
Time	Timestamp, time range, time of day, on weekdays/weekend, day of the week
Location	Entry/exit event for a geo-fence, detected places of the user (home, work, others), en route
Activity	Still, walking, running, on bicycle, in vehicle, sleeping
App usage	Opening/leaving event for specific apps or categories
Wake-up event	First smartphone use after sleep activity

Last but not least, for modeling the *wake-up* trigger, we use the first smartphone use after a sleep activity. To recognize the latter, we decided to use a simplified approach, as the individual sleep stages are not relevant for our study [75]: we label a lasting *still* activity over several (> 4) night hours at home as 'sleeping' – outliers (less than one minute), e.g., when users briefly look at their smartphone in the night, are filtered.

Finally, it is important to note that the app processes all data *locally* in an energy-efficient way; the latter is achieved by the fact that most of the processing operations are event-driven or scheduled, and thus, no continuous (resource-intensive) processing is necessary.

Table 2 gives an overview of these context triggers, which can be combined (rule-based) as needed to cover all use cases of our study. For instance, to trigger an in-situ questionnaire for the *social health support* use case *"You have an appointment at 7 p.m. with some friends. Today, the roads will be very busy. So, if you leave home too late, you might not make it in time to your appointment."* (see Table 1, #9), we combine the time (*6 p.m. - 6:45 p.m.*) and location (*home*) triggers. It is important to state that these use cases are hypothetical (see Section 4.1); they are only grounded in the context in which they are very likely to occur and which is covered by the triggers above.

As soon as a use case was triggered by the app, participants were notified and given the opportunity to either answer the use case related questions immediately or postpone them (see Figure 1b). The use case specific survey is displayed only when the context-triggered notification was accepted. On the contrary, if the time was unfavorable and the participant dismissed the notification sent by the app, the survey was postponed to a later point in time. This mechanism was introduced for safety reasons to avoid possible damage or injury that could be caused by responding to our survey in inappropriate (risky) situations.

In the case of not viewed notifications (i.e., the participant has not accepted the app notification), a timeout removes such notifications. We set the timeout to typically 10 minutes, as it is becoming increasingly unlikely that the participants are still in the desired situation. Notifications that have been dismissed or removed by timeouts are rescheduled when the participants are back in the corresponding situation. This rescheduling loop is repeated until the participant responds to all our notifications and the corresponding surveys. However, the ProfileMe application triggered a maximum of 8 notifications per day (excluding the notifications deleted by timeouts). This approach is based on the intention not to overload or disturb the participants by asking them to answer too many surveys in a single day. Similarly, the app did not trigger notifications when participants were asleep.

Finally, it is noteworthy that any use cases that the participants could not answer in-situ (e.g., because the use case was simply not triggered by the participant's daily behavior) could be answered by the participant at the end of the study.

In the case that ProfileMe triggered a use case, and the participant accepted the notification for this, the corresponding in-situ survey is started. The in-situ questionnaire entails information about the use case and questions, which are presented to the participants in three consecutive stages: (1) *Use case (UC) description*—in the first stage, the participants see the explanation of the use case scenario in which an assistant could support the participants in their current situation (see Figure 1c). (2) *Assessing the degree of proactivity*—the participants are then presented with different randomly-ordered proactivity options representing *no support*, *reactive support*,

Table 3. Overview of the questions for better understanding the participants' proactivity choices – we use the term 'assistant' as a substitute for 'AI system' to better communicate the more technical concepts of AI systems to the participants.

ID	Additional Questions (I)
Q1	The use case is not relevant for me.
Q2	I feel such a system could infringe my privacy.
Q3	I am afraid that the assistant could make a mistake I cannot overrule.
Q4	I do not fully trust in the abilities of such systems.
Q5	An assistant which has more functionalities would be too intrusive.
Q6	The use case is too important to let an assistant decide for me.
Q7	I love organizing, deciding and doing everything myself.

(a) Questions triggered if participants selected any of the options *except* for the *autonomous support* option

ID	Additional Questions (II)
Q1	I trust the abilities of the assistant.
Q2	I feel the assistant makes my world more predictable.
Q3	The assistant makes it easier for me to structure/organize my daily life.
Q4	The assistant makes me feel secure and does not violate my privacy.
Q5	The assistant is very helpful to prevent stress in my daily life.
Q6	The assistant is very convenient.
Q7	I feel that the assistant can prevent problems.

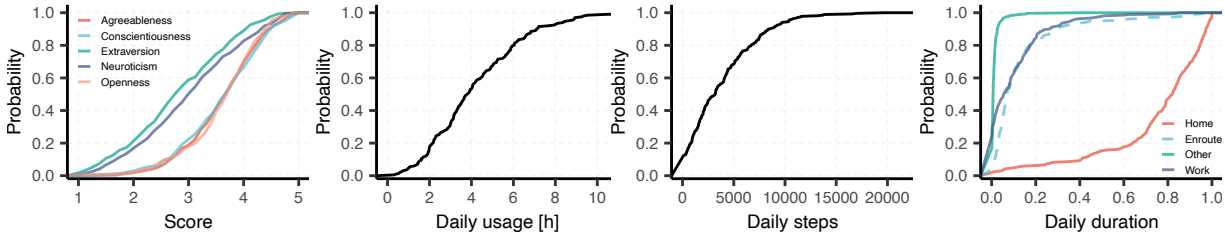
(b) Questions triggered if participants opted for the *autonomous support* option (i.e., the highest proactivity level)

proactive support I, *proactive support II*, and *autonomous support* – the random order is required for our plausibility checks. In this stage, the participants were asked to make their choice for one of the mentioned support options (see Figure 1d). (3) *Questions for understanding the decision*—during the last stage, participants were asked to rate one of two sets (see Figure 1e), each comprising seven different statements (see Table 3). The "additional questions II" (see Table 3b) was shown only if participants opted for *autonomous support* (i.e., the highest level of proactivity). Otherwise, participants were asked to rate the statements comprising "additional questions I" (see Table 3a). For this purpose, we use five-item Likert scales ranging from *strongly disagree* to *strongly agree*. Each set of statements allow us to better understand why participants selected a certain level of proactivity in stage (2).

4.3 Recruitment of Participants

The study participants were recruited via Prolific[8]—a crowd working platform for social science experiments [54]. Although our study is not limited to a specific group of people, we used the pre-screening feature of Prolific to filter participants by language (English), operation system versions of Android (5.0 or higher), and driver's license (including owning a car). The pre-screening of participants was necessary to meet restrictions imposed by the app (e.g., operating system) or the requested use cases (e.g., the driver's license filter is necessary, as at least one of our use cases targets drivers). Monetary incentives motivated the participants to take part in our study. Participants who completed the study were rewarded with $8 each. In total, the app was installed by 458 participants. However, many participants stopped answering the questionnaires either during the setup

[8]https://prolific.ac (retrieved 10/28/2020)



(a) Personality (b) Smartphone usage (c) Movement activity (d) Mobility pattern

Fig. 2. Characterization of the participants by CDFs

phase or after a few days. In the end, a total of 276 participants completed the study, resulting in a drop-out rate of 39.7% – a typical drop-out rate, considering the length and effort required to complete our in-situ study [8].

4.4 Privacy Compliance

To ensure privacy compliance, the ProfileMe application implements mechanisms compliant to the EU General Data Protection Regulation (GDPR). According to this regulation and in particular to the *right to be informed*, participants had to explicitly give their consent (i) to participate in the study, (ii) to have their data collected by the ProfileMe app, and (iii) to use the data for research purposes. In addition, according to the *right to restrict processing*, participants had to give explicit permission for each type of data collected, as required by the Android operating system. During the study, all raw data was stored locally. Only at the end of the study, after renewed explicit consent of the participants, the aggregated daily summaries (e.g., app usage, mobility patterns) and the answers to the in-situ questionnaires were uploaded to our server. The study participants are identifiable only by their Prolific ID—i.e., a 24-digit combination of letters and numbers, which does not allow any inference as to the real identity of the participant. Based on their *right of access and data portability*, participants were able to download all data collected by the app (in CSV-format) at any time. Finally, all participants could have stopped the study at any time and requested the deletion of their data (*right to object and erase*).

5 DATASET AND USER CHARACTERIZATION

The data was collected during one week in November 2019. In total, 276 participants have successfully completed the study. To ensure high data quality, we first apply coarse plausibility checks based on completion times and response patterns. During these checks, four participants were removed from our dataset. In the following, we report all results based on our final dataset comprising *272 participants* with complete and plausible data.

The final dataset is almost balanced between the genders, with 138 females and 134 males. All participants are aged between 18 and 64 ($\mu = 35.0, \sigma = 9.8$). Although we did not limit the study to country-specific target groups, participants originate from 13 European countries (incl. GB, PT, DE, ES) or North America (US, CA). The three most frequent countries are GB (46.7%), US (29.4%), and PT (8.1%). This distribution is probably due to the fact that the Prolific platform is based in GB and to its degree of popularity in these countries. Furthermore, the majority of participants are married (39.0%), single (37.8%), or in a partnership (23.2%). Participants have either no kids (61.8%), one (14.0), two (15.8%), or more children (8.4%). Moreover, our participants' education ranges from individuals with no degree (2.2%) over participants with high school (14.3%), a professional degree (4.1%) to college (pursuing BSc: 23.2%; BSc: 37.5%; MSc: 15.8%) or a doctorate (2.9%). In terms of employment status, some participants report to have no job (11.8%), others are full-time students (6.3%), have part-time jobs (15.8%), or are retired (0.7%). Even so, most of the participants have full-time jobs (65.4%).

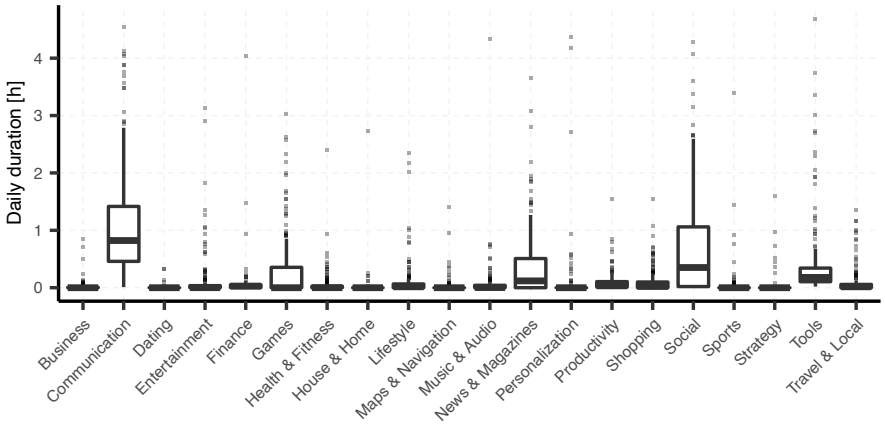


Fig. 3. App usage of the participants for top 20 categories (*in alphabetical order*)

As mentioned previously, our participants also completed a *Big Five* personality test [23]. Figure 2a shows the results of the *Big Five Personality Traits Test*: extraversion ($\mu = 2.85, \sigma = 0.93$), agreeableness ($\mu = 3.66, \sigma = 0.68$), conscientiousness ($\mu = 3.65, \sigma = 0.75$), neuroticism ($\mu = 3.06, \sigma = 0.95$), and openness ($\mu = 3.64, \sigma = 0.69$). These results show that the majority of our participants are influenced more by their agreeableness, conscientiousness and openness than by their neuroticism and extraversion. In terms of participants' previous experience with popular digital assistants, notably, around a quarter of the participants (70) stated that they had never used an assistant. In contrast, the majority of the participants uses at least one assistant from time to time (56.3%), often (14.7%), or very frequently (3.3%). The participants further indicated that they had already used the following assistants: Google Assistant (160), Amazon Alexa (73), Microsoft Cortana (25), or Apple Siri (12). Google Assistant's supremacy is no surprise, as our study required participants with Android mobile phones.

We further collected daily summaries from the participants' sensor data to better characterize them. Figure 2b shows the (relatively high) average daily smartphone usage ($\mu = 4.23\,h, \sigma = 2.33\,h$). More specifically, Figure 3 breaks down the usage per app category. Similar to [64, 72, 76], we used the app categories retrieved from Google Play, as these are sufficient for our purpose. We can see that our participants mostly use *communication* (e.g., WhatsApp) and *social* (e.g., Facebook) apps. To further characterize the users, such future studies should also analyze the app/device usage sessions as in [4, 22]. Moreover, our participants move relatively little in terms of steps ($\mu = 3865.3, \sigma = 3489.2$) (see Fig. 2c). The mobility pattern shown in Figure 2d further reveals that our participants are, on average, very often – three quarters of the day ($\mu = 75.0\%, \sigma = 24.5\%$) – at home, only leaving the house to work ($\mu = 10.3\%, \sigma = 14.0\%$), while staying in other places is quite rare ($\mu = 1.4\%, \sigma = 0.3\%$). Generally speaking, the statistics show that the majority of our participants live a sedentary lifestyle from home or at work, and social life mostly takes place online with a tendency towards smartphone addiction—a status quo reflecting some of the problems humans face in the age of digitization, in which digital personal assistants or user-supporting AI systems often have their focus [57].

For our analysis, we are primarily using the in-situ questionnaire responses. Accordingly, we analyze users' reactivity to notifications linked with these questionnaires. If the participants accepted a notification, the median reaction time was $83.5\,s$ ($\mu = 331.3\,s, \sigma = 475.2\,s$). For dismissed notifications, the median reaction time was $155.7\,s$ ($\mu = 450.9\,s, \sigma = 637.9\,s$). This indicates that most of the in-situ questionnaires were seen and answered sufficiently fast that users were most likely still in the specific situation.

Before the participants could finish the study, they had to answer the remaining questions that were not answered in situ (similar to an online survey). We collected a total of 6,256 responses, of which 3,168 (50.6%) were answered in situ. We are aware that our dataset has potential limitations stemming from the inherent nature of an *in-the-wild* study [39] (see the *study limitations* in Section 8.2). The dataset remains unbalanced: it is not possible to trigger each use case specific survey for each user in situ, as some users never came into the required situations during our study. Even if users come into these situations, in practice, there is a probability that they will miss or dismiss the notification. Nevertheless, we did not exclude any users who have successfully completed the study, as all users have responded to in-situ surveys ($\mu = 11.7$, $\sigma = 3.9$, $x_{min} = 6$, $x_{max} = 23$); only their number of in-situ responses differs. The following exploratory analysis is conducted along the in-situ responses, grouped per use case ($\mu = 137.7$, $\sigma = 80.5$, $x_{min} = 21$, $x_{max} = 272$).

6 UNDERSTANDING THE EXPECTED PROACTIVITY

In this section, we explore the influence of various human factors and application areas on the expected level of proactivity, using the collected dataset from Section 5.

6.1 The Measurement Effect of In-Situ Surveying

We first examine the effect of the in-situ design of our *in-the-wild* study. To quantify this measurement effect, we encode the proactivity levels with the following ordinal scale: *no support=0*; *reactive support=1*; *proactive support I=2*; *proactive support II=3*; and *autonomous support=4*. We then apply a non-parametric Kruskal–Wallis test to test the null hypothesis stating that "there is no significant difference between the means of the two methods, namely our proposed method and conventional online surveys". The *in-situ* reported proactivity levels are used for the former, while the reported proactivity levels *at the end of the study* are used for the latter. According to our results, $\chi^2(1) = 14.3, p < .001$, there exists a statistically significant difference between both study methods on capturing participants' expectations. More specifically, our results show that the median proactivity levels of the in-situ responses are slightly higher than those of the responses given at the end of the study (or in online surveys). We, therefore, decided to use only the 'recall bias-reduced' responses and report our analysis based on the 3,168 *in-situ* questionnaire responses.

6.2 The Role of the Use Case and Area of Support

To investigate the role of the use case type on the users' expected proactivity level, we visualized the respective data in Figure 4. The graph shows the corresponding participants' in-situ responses on the proactivity scale. The *yellow ochre shades* on the left indicate no proactive support (i.e., no or only reactive support) while *turquoise shades* on the right show proactive support expectations. In over 78% of all use cases (18 out of 23), participants opted to receive proactive support through an AI system. Especially for the use cases #9/#10, which concern appointments and recurring events in the social context, the majority of participants (90%) wish to be proactively supported by an AI system. Specifically, the participants wish for fully autonomous support for use cases such as (#18) *do not disturb/availability* handling and (#19) smart home control, i.e., light and climate control – these use cases relate to digital activity support and ambient intelligence respectively. In contrast, many participants want no or reactive support in the (#15) *shopping* use case and in most of the *mental health* related use cases (#5-8).

Across all use cases, the most frequent choice is *proactive support II* ($\mu = 32.5\%, \sigma = 12.5\%$), followed by *proactive support I* ($\mu = 19.7\%, \sigma = 9.0\%$) and *reactive support* ($\mu = 20.1\%, \sigma = 10.5\%$). However, in some cases ($\mu = 15.5\%, \sigma = 12.8\%$) participants opt not to be supported at all – neither reactive nor proactive. Interestingly, the least preferred choice of participants is *autonomous support* ($\mu = 12.2\%, \sigma = 9.6\%$).

To further examine whether the participants' preferences for proactivity differ in the different areas of support, we grouped the use cases according to Table 1. We can see that the type of area has a clear impact on the

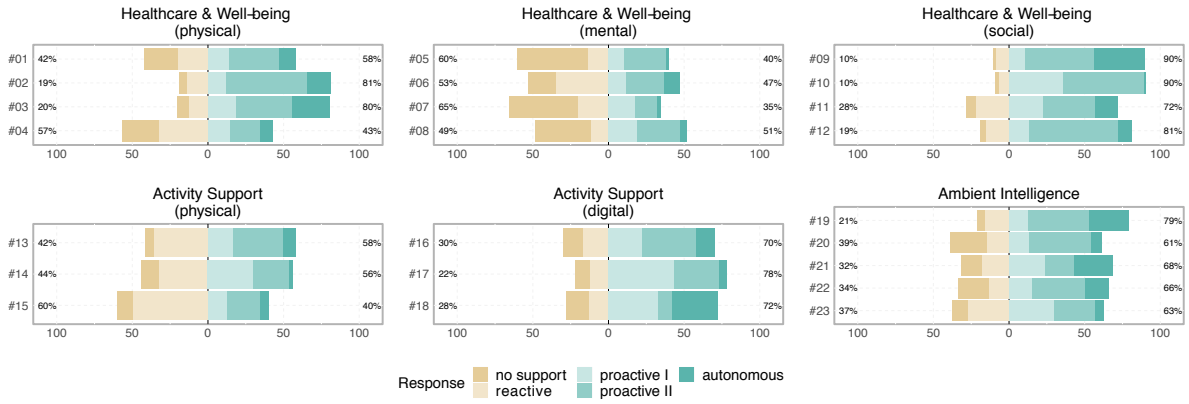


Fig. 4. The in-situ responses to expected proactivity for the different use cases on the proactivity scale

choice of the proactivity level. In particular, the preferred proactivity level for the area of *mental health support* ($\mu = 1.40, \sigma = 0.94$) is the lowest of all areas – the average preferred level corresponds to the reactive support. In the four areas of *physical health support* ($\mu = 2.01, \sigma = 0.88$), *physical activity support* ($\mu = 1.99, \sigma = 0.78$), *ambient intelligence* ($\mu = 2.24, \sigma = 0.97$), and *digital activity support* ($\mu = 2.28, \sigma = 0.79$), most of the participants rather prefer a proactive support in terms of notifications and alerts. Only in the area of *social well-being support* ($\mu = 2.66, \sigma = 0.94$), well over three quarters of all users prefer proactive support; the average preferred level is the highest of all areas and corresponds to a proactive support with personalized recommendation.

All in all, our results indicate that users are less open to proactive support when it comes to supporting their mental problems. Although state-of-the-art AI systems could detect mental issues and initiate countermeasures, our participants prefer to receive only reactive support or no support at all in such situations. In contrast, users are more open or wish proactive support in areas that affect their physical health (e.g., #2–fitness plan), ambience (e.g., #19–smart home control), or digital surroundings (e.g., #17–email management). Notably, users are most open to proactive support for social interactions (e.g., #9–arranging appointments, or #10–annual reminders).

6.3 The Impact of Socio-demographics and Personality Traits

We now investigate the impact of different socio-demographic factors on the expected proactivity level. To this end, we use different graphical exploratory approaches to gather preliminary information that may help to suggest hypotheses in the second step. As there were only a few noticeable differences for countries of origin, material status, school degree, and employment status, we will only briefly discuss the largest differences.

Regarding countries of origin, participants from *Canada* ($\mu = 1.73, \sigma = 0.48$) prefer lower proactivity levels across most of the use cases than participants from any other country; the *German* participants ($\mu = 2.08, \sigma = 0.86$) have the greatest variance in their given answers. Regarding material status, we can only say that *separated* participants ($\mu = 2.72, \sigma = 0.58$) are most open to proactivity and even to autonomous AI systems; *singles* ($\mu = 2.06, \sigma = 0.89$) have the highest variance in their responses. Regarding school degree, participants with *doctorate degree* ($\mu = 2.38, \sigma = 0.72$) are most open to proactivity, while participants with a *professional degree* ($\mu = 2.15, \sigma = 1.06$) stand out for the high variance in their responses. Regarding employment status, the only thing that stands out is that participants with a *full-time job* ($\mu = 2.49, \sigma = 0.58$) prefer higher proactive support levels than all others; those with *no job* ($\mu = 2.16, \sigma = 0.56$) prefer lower proactive support levels than all others – both are consistent for all use cases and in particular for the area of social well-being support.

Table 4. Results of the multinomial logistic regression with *proactivity levels* as dependent variables, and *socio-demographics* and *personality traits* as independent variables.

	No Support		Reactive Support		Proactive Support I		Autonomous Support	
Variable	*RRR*	*p*	*RRR*	*p*	*RRR*	*p*	*RRR*	*p*
Socio-demographics								
gender	1.196	0.250	1.055	0.718	1.126	0.354	0.941	0.699
age	1.017*	0.045	0.100	0.989	0.996	0.576	1.001	0.864
no. of children	0.977	0.780	1.014	0.868	1.047	0.513	1.173*	0.049
Personality traits								
extraversion	0.874	0.163	0.832*	0.045	1.035	0.663	0.939	0.516
agreeableness	1.029	0.826	1.081	0.529	1.086	0.444	0.788	0.065
conscientiousness	1.355*	0.015	1.045	0.702	1.005	0.957	1.162	0.225
neuroticism	1.093	0.396	1.117	0.274	1.122	0.186	1.087	0.432
openness	0.748*	0.016	1.235	0.083	0.959	0.678	1.016	0.896

N=1,865
Base outcome: *proactive support II* ($\mu = 32.5\%, \sigma = 12.5\%$)

Next, we investigate the role of users' personality traits on the *reported* expected proactivity level. Previous studies on reactive systems [20] or notifications [39] showed that the personality might have an impact on an individuals' choice behavior. To better understand this behavior (also in terms of causality), we conducted a multinomial logistic regression analysis. In addition to the five personality traits, we also consider the remaining (selected) socio-demographic traits, namely age, gender, and the number of children [14].

Table 4 shows the results of the regression analysis with the estimated coefficients (as relative risk ratios – abbreviated as *RRR*) and *proactive support II* as the base outcome (i.e., overall, the most frequently chosen option). As our results show, elderly participants with low openness levels and higher conscientiousness levels are significantly more likely to choose the *no support* option. In contrast, participants with higher levels of openness are more likely to opt for an AI system with *proactivity support II*. As an example, an increase in participants' age by one unit (i.e., one year) leads to an $1.7\,\%$ increase in the likelihood ($p = .045$) of preferring *no support* at all, over an AI system that can intervene and issue personalized recommendations (*proactivity support II*). Analogously, an increase by 10 years would lead to a 17% higher probability that participants choose no support over proactive support, and so on. In addition, regarding *reactive support*, it is notable that participants with lower levels of extraversion will more likely ($p = .045$) to choose a proactive AI system (*proactivity support II)* over a reactive one (i.e., an increase of extraversion by one unit results in a 16.8% decrease of the likelihood to choose a reactive system). We further found no significant factors influencing the participants' choice for *proactive support I* in comparison to the base outcome. Finally, the number of children increases the likelihood of a participant wishing for the highest proactive support (i.e., *fully autonomous support*). Specifically, each additional child increases the chances of the parent choosing the fully autonomous support by 17.3%.

All in all, the results have face validity and can be well-explained by existing theories and considerations from the realm of psychology. For instance, our estimations reveal that participants with lower levels of openness are more likely to opt for *no support* than their peers. Considering the Big Five personality traits framework [23], this result is plausible as openness refers to the excitement and curiosity for new things.

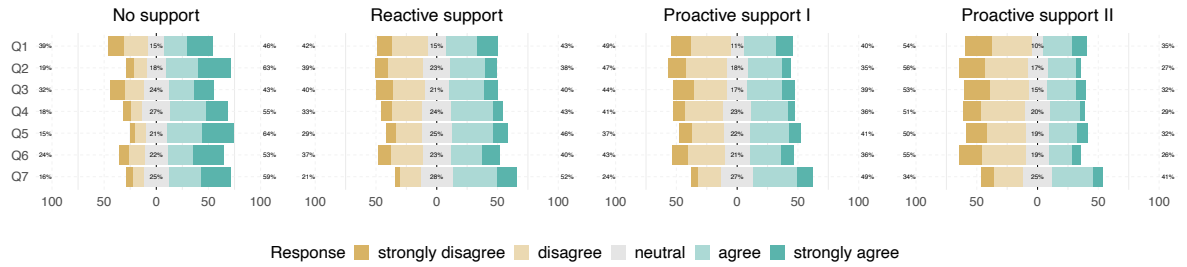


Fig. 5. The responses to the additional questions (see Table 3a) asked immediately after the user's decision for a proactivity level lower than autonomous support.

6.4 Does User Pattern Matter?

Finally, we examine the impact of aspects that are characterized by users' behavior patterns on the expected proactivity level. For this purpose, we compute Pearson correlations of the *total*, *per area* and *use case-specific* proactivity levels for 34 different user patterns, including mobility and movement patterns as well as interaction patterns with the mobile device. We found several *weak* correlations for participants' mobility and app usage patterns. For instance, the use of apps in the *lifestyle* category (e.g., MoodTracker, OpenHab) is positively correlated with proactivity expectations of AI systems in the *social well-being* ($r = .382,\ p = .009$) and *digital activity support* areas ($r = .432,\ p < .001$). Regarding mobility, the average and maximal residence times at work ($r = .429,\ p = .002$) are positively correlated with the use case #5 (exhaustion/stress). This is plausible considering that overworked/stressed participants want a more proactive AI system to mentally support them. Another notable example is that the residence time at home is negatively correlated ($r = -.241,\ p = .045$) and the residence time at work is positively correlated ($r = .372,\ p = .025$) with the use case #8 (boredom). Moreover, we found *no* significant correlation of the participants' movement patterns and general smartphone use with participants proactivity choices.

7 UNDERSTANDING THE REASONS FOR THE DECISIONS

In this section, we analyze the reasons for the decisions users have made regarding the expected level of proactivity. After the participants had selected their preferred option from the five possible AI support options, they had to rate various explanatory statements (see Table 3) on a five-point Likert scale. Depending on the proactivity levels, a different set of statements is displayed. With the first set of statements (see Table 3a, "Additional Questions I"), we pursue to find out why users did not choose a higher proactivity level when participants chose one of the four support options that are lower than *autonomous support*. Conversely, the second set of statements (see Table 3b, "Additional Questions II") aims to find out why the participants have chosen the highest level of proactivity.

Figure 5 shows the ratings of the participants for the first set of explanatory statements. These ratings reveal that between a third and a half of the participants who expected either *reactive support*, *proactive support I or II* rated the in-situ presented use cases as not particularly relevant to their daily lives (Q1), and therefore could not imagine a higher level of proactivity for them. Interestingly, participants who opted for the *no support* option (Q2) are concerned that AI systems with a higher degree of proactivity may violate their privacy (63%); (Q4) report that they have trust issues in the abilities of AI systems (55%); (Q5) feel that a more proactive system may be too intrusive (64%); or (Q7) report that they are satisfied with the organization of their daily lives alone (59%).

In contrast, participants who opted for the *autonomous support* option (i.e., the highest proactivity level) have a more positive attitude towards AI systems and the support they could provide. More specifically, 90% of the participants who chose an autonomous AI system stated that they trust in the abilities of such systems (Q1).

82% of the participants feel that such AI systems can make their lives more predictable (Q2). Over 90% of these participants feel that autonomous AI systems would make their life more organized (Q3) and could help prevent stress (Q5), as it is also very comfortable (Q6). 89% of the participants feel that AI systems could under certain circumstances even prevent the occurrence of further problems (Q7). Although these participants have a positive attitude towards autonomous systems, only 57% feel safe and that such systems do not violate their privacy (Q4).

In summary, a brief comparison between the groups of participants who opted for the AI systems with *no support*, *reactive support*, *proactive support I*, or *proactive support II* reveals that the choice of participants can be partly explained by their attitudes and beliefs towards the AI system. It is remarkable that participants who have chosen a higher proactivity level are less concerned that an AI system might violate their privacy (Q2); the same is true for trust in the abilities of AI systems (Q4) and the unobtrusiveness of such AI systems (Q5). These findings indicate that if users are given a high level of confidence in the capabilities of AI systems and the protection of privacy, they will be more open to higher levels of proactivity.

8 DISCUSSION

In this section, we discuss our findings in terms of their *implications* and describe the *limitations* of our study that should be considered when interpreting the results.

8.1 Implications

8.1.1 Openness towards Proactive AI systems. All in all, our findings reveal that the participants are very open to proactive AI systems, as the most common expectation relates to the level of *proactive support II*; i.e., AI systems that base their actions on predictive models to intervene in the lives of users and trigger personalized recommendations. We also found that participants' preferences regarding the proactivity of AI systems vary not only based on personality and socio-demographic characteristics (e.g., age, number of children), but also the use case itself (e.g., *do not disturb* handling) or more generally in the areas of support (e.g., mental health support). Our findings further suggest that participants' privacy concerns and trust issues in the capabilities of AI systems influence participants' preferences in terms of the level of proactivity. These findings have several implications for the design and development of proactive AI systems.

Firstly, AI systems must be versatile and able to adapt their level of proactivity according to the support area and user characteristics. *Secondly*, AI systems should be intelligible at all times so that users can understand the proactive or autonomous actions of such systems and build trust in their capabilities. *Thirdly*, based on these results, designers of such AI systems should design and implement AI systems that meet the different levels of privacy affinity of their participants. Especially at higher levels of proactivity and intelligence, when such AI systems have to provide their models with increasingly sensitive user data, AI systems can be designed to process this data only locally. For instance, general AI models can be personalized to users in a subsequent personalization step [42, 44] running in a trusted execution environment to protect both user data and the provider's proprietary AI algorithms/models [40, 41].

8.1.2 Area-adapted Proactivity to Lower the Entry Barrier. We found that user expectations widely vary between the different application domains or areas of user-centered support. In some areas, the majority of users wants to stay in control or do not want to receive proactive support. Especially in the area of *mental health support*, surprisingly many users do not want any support at all (not even reactive support) – in this case, researchers may need to rethink such applications completely. For the rest, AI systems should respect the initial preferences of a user to avoid direct rejection of these technologies. Our findings can help to meet these initial expectations, which may be the 'entry ticket', even if the degree of proactivity is less than that with which the system can work. As the user gradually builds confidence in the AI system and sees its benefits (*onboarding*), the system can explore a higher level of proactivity – with the user's consent, of course.

8.1.3 Building a User-dependent Proactivity Model. We have observed that the expected level of proactivity on which an AI system should work is influenced by parts of the user's socio-demographic and personality traits. For certain use cases and areas of support, the mobility and usage behavior of users can also be an indicator of their expectations. These show the potential for building a user-dependent proactivity model taking into account socio-demographics, user patterns, and personality traits, which can already be easily and partially automatically captured. Such estimation models can then be generalized to a group of users who share the same human factors. The resulting community models can help to overcome the inherent cold start problem and to better and more quickly determine the degree of proactivity of an AI system that meets the different individual expectations.

8.2 Study Limitations

Although our decision to conduct an *in-the-wild* study allows us to gain novel insights, which might not be observable in lab environments with limited number of participants or through online surveys, most limitations of this study stem from its in-situ nature and design.

8.2.1 Recruitment and User Base. Firstly, the tradeoffs made in recruitment might have affected the results. Specifically, we targeted a sample (user base) that adequately represents the general population – and is not limited to computer scientists/ computer science students or participants experienced with this kind of technology.

Next to the challenge of sample size, the type of our study (in-the-wild study with in-situ surveying) and its used data collection method (24/7 collection of sensor data) posed challenges in recruiting participants as well. The former requires a higher degree of time commitment from participants (to answer questions in several small sessions at specific times or situations); the latter requires the willingness of participants to provide privacy-sensitive data for research purposes.

To address these recruitment challenges, we have made the following trade-offs. In-house promotion would probably have largely targeted regional participants only and would have caused considerable effort – achieving a large sample size would probably not have been feasible in the short term (taking into account the study requirements). Therefore, we decided to use a professional recruitment platform to achieve a decent number of participants – but at the expense of monetary incentives and a high refusal or dropout rate. Such incentives are potentially more likely to motivate people to participate in such a 'time-consuming' study, install the study tool on their personal device, and provide their data (consistent with the privacy paradox [6]). On the downside, getting responses from people who only want to get the reward and are really not that interested could hurt the quality of data. To counteract this, we explicitly pointed out our plausibility and quality checks to participants before the study is considered as 'successfully completed' and they receive the reward. To prevent participants from having privacy concerns, the study tool relies on GDPR-compliant and local data processing; participants only have to provide us with aggregated data (daily summaries) for research purposes (one-time upload at the end of the study, but only after confirmation by the user). This, in turn, limits our analysis options, as we do not have raw user data. Nevertheless, this recruitment method provides the best cost-effectiveness tradeoff for our study and research purposes.

Secondly, this recruitment method, on the other hand, might lead to a (self-selection) bias in the types of participants. Given that the study tool is already an application that collects sensor data from users and acts 'proactively' (automatically triggering in-situ surveys), participants might have been biased towards accepting a similar level of proactivity. An argument for this could be that if they had been more concerned about automatic actions and privacy, they might not have installed the study tool or participated in the study at all. In this case, this might lead to more positive results or even overall impressions than is the case in reality – it is therefore all the more important that AI systems implement the onboarding mechanisms proposed above so that they are not immediately rejected. An argument against this is that the otherwise rather conservative participants nevertheless took part in the study as well, either out of mere interest or because it is 'only' a user study with a

short (manageable) time frame and no autonomous actions (with possible unwanted/unexpected consequences for them). In this case, the results would reflect the user expectations from both types of participants.

Finally, the recruitment platform (Prolific), which is based in the UK, and the language of the study/app (English) might have biased our user base. While participants are from 15 unique countries of Europe and North America, more than three-quarters of them (76.1%) come from two English-speaking countries, namely the UK and the USA. Apart from that, the only restriction we made on recruitment was to allow only Android (5.0+) smartphone users who also have a driver's license and own a car (to be able to trigger corresponding in-situ surveys). Even so, we recruited 272 participants who are more diverse than the participants in typical controlled studies.

8.2.2 Study Design and In-situ Responses. Firstly, our algorithm for automatically detecting relevant situations (to trigger the in-situ questionnaires) is based only on the to-date available context triggers on user's mobile device. Accordingly, our study represents an approximation of the situations (hypothetical use cases) in which future AI-based systems will support users through their underlying prediction models.

Another limitation of this study is the fact that our dataset is imbalanced in terms of in-situ responses. Due to the nature of the study, some users never experienced the situations required to trigger the use cases prompted in the study. We also decided to trigger each in-situ survey for a particular use case only once so that participants do not find the study annoying if they receive too many notifications – a tradeoff we had to make in order *not* to compromise data quality and reliability of the participants' responses. This, in turn, limits our analysis options, as it is not possible to investigate whether participants consistently respond to the same use cases.

Finally, some of the additional questions (e.g., *"I feel such a system could infringe my privacy"*) to participants could have biased their responses – e.g., the sample question strongly implies threats to the participant's privacy. Therefore, such a potential bias should be considered in particular for the answers to Q2+3 of the Additional Questions I in Figure 5 and the answers to Q4 of the Additional Questions II (see Section 7). In follow-up studies, such questions should be formulated more neutrally in order to reduce such a potential bias to the participants.

9 CONCLUSION

In this paper, we explored user expectations of emerging AI systems that can operate on different proactivity levels to provide user support. The contribution of this paper is twofold. *Firstly*, we designed and conducted an *in-the-wild* study, surveying 272 participants for several use cases in situ. *Secondly*, based on the observational and self-reported data collected in this study, we gained insights when and to what extent AI systems should take the initiative, and investigated the effect of different human factors and contexts influencing these user expectations. The findings of our study may help to design future AI-powered systems and preset their proactivity level to meet an individual's expectation.

ACKNOWLEDGMENTS

This *collaborative* research work has been co-funded by (1) the German Federal Ministry of Education and Research and the Hessian Ministry of Higher Education, Research, Science and the Arts within their joint support of the National Research Center for Applied Cybersecurity ATHENE, by (2) the German Federal Ministry for Economic Affairs and Energy (BMWi) as part of the ENTOURAGE project [01MD16009F], and by (3) the German Federal Ministry of Education and Research (BMBF) Software Campus project "TheNextSmartHome" [01|S17050].

REFERENCES

[1] Alper T Alan, Enrico Costanza, Joel Fischer, Sarvapali D. Ramchurn, Tom Rodden, and Nicholas R. Jennings. 2014. A Field Study of Human-Agent Interaction for Electricity Tariff Switching. In *Proceedings of the 2014 International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS'14)*. IFAAMAS, 965–972. https://eprints.soton.ac.uk/360820/

[2] Alper T Alan, Enrico Costanza, Sarvapali D Ramchurn, Joel Fischer, Tom Rodden, and Nicholas R Jennings. 2016. Tariff Agent: Interacting with a Future Smart Energy System at Home. *ACM Transactions on Computer-Human Interaction* 23, 4 (2016), 1–28. https://doi.org/10.1145/2943770

[3] Rolph E Anderson. 1973. Consumer Dissatisfaction: The Effect of Disconfirmed Expectancy on Perceived Product Performance. *Journal of Marketing Research* 10, 1 (1973), 38–44. https://doi.org/10.2307/3149407

[4] Nikola Banovic, Christina Brant, Jennifer Mankoff, and Anind Dey. 2014. ProactiveTasks: The Short of Mobile Device Use Sessions. In *Proceedings of the 16th International Conference on Human-computer Interaction with Mobile Devices and Services (MobileHCI'14)*. 243–252. https://doi.org/10.1145/2628363.2628380

[5] Louise Barkhuus and Anind Dey. 2003. Is Context-aware Computing Taking Control Away from the User? Three Levels of Interactivity Examined. In *International Conference on Ubiquitous Computing (UbiComp'03)*. Springer, 149–156. https://doi.org/10.1007/978-3-540-39653-6_12

[6] Susanne Barth and Menno DT De Jong. 2017. The Privacy Paradox–Investigating Discrepancies Between Expressed Privacy Concerns and Actual Online Behavior–A Systematic Literature Review. *Telematics and informatics* 34, 7 (2017), 1038–1058. https://doi.org/10.1016/j.tele.2017.04.013

[7] Frank Bentley, Chris Luvogt, Max Silverman, Rushani Wirasinghe, Brooke White, and Danielle Lottrjdge. 2018. Understanding the Long-Term Use of Smart Speaker Assistants. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 3 (2018), 91:1–91:24. https://doi.org/10.1145/3264901

[8] Niels Van Berkel, Denzil Ferreira, and Vassilis Kostakos. 2017. The Experience Sampling Method on Mobile Devices. *Comput. Surveys* 50, 6 (2017), 93:1–93:40. https://doi.org/10.1145/3123988

[9] Anol Bhattacherjee. 2001. Understanding Information Systems Continuance: An Expectation-confirmation Model. *MIS Quarterly* 25, 3 (2001), 351–370. https://doi.org/10.2307/3250921

[10] Enrico Blanzieri and Anton Bryl. 2008. A Survey of Learning-based Techniques of Email Spam Filtering. *Artificial Intelligence Review* 29, 1 (2008), 63–92. https://doi.org/10.1007/s10462-009-9109-6

[11] Dario Bonino and Fulvio Corno. 2011. What Would You Ask to Your Home If It Were Intelligent? Exploring User Expectations about Next-generation Homes. *Journal of Ambient Intelligence and Smart Environments* 3, 2 (2011), 111–126. https://doi.org/10.3233/AIS-2011-0099

[12] Thomas M Brill. 2018. *Siri, Alexa, and Other Digital Assistants: A Study of Customer Satisfaction With Artificial Intelligence Applications*. Ph.D. Dissertation. University of Dallas.

[13] Luca Canzian and Mirco Musolesi. 2015. Trajectories of Depression: Unobtrusive Monitoring of Depressive States by Means of Smartphone Mobility Traces Analysis. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp'15)*. ACM, 1293–1304. https://doi.org/10.1145/2750858.2805845

[14] Gokul Chittaranjan, Jan Blom, and Daniel Gatica-Perez. 2013. Mining Large-scale Smartphone Data for Personality Studies. *Personal and Ubiquitous Computing* 17, 3 (2013), 433–450. https://doi.org/10.1007/s00779-011-0490-1

[15] Enrico Costanza, Joel E. Fischer, James A. Colley, Tom Rodden, Sarvapali D. Ramchurn, and Nicholas R. Jennings. 2014. Doing the Laundry with Agents: A Field Trial of a Future Smart Energy System in the Home. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'14)*. ACM, 813–822. https://doi.org/10.1145/2556288.2557167

[16] Benjamin R Cowan, Nadia Pantidi, David Coyle, Kellie Morrissey, Peter Clarke, Sara Al-Shehri, David Earley, and Natasha Bandeira. 2017. What Can I Help You With?: Infrequent Users' Experiences of Intelligent Personal Assistants. In *Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services (MobileHCI'17)*. ACM, 43:1–43:12. https://doi.org/10.1145/3098279.3098539

[17] Allan de Barcelos Silva, Marcio Miguel Gomes, Cristiano André da Costa, Rodrigo da Rosa Righi, Jorge Luis Victoria Barbosa, Gustavo Pessin, Geert De Doncker, and Gustavo Federizzi. 2020. Intelligent Personal Assistants: A Systematic Literature Review. *Expert Systems with Applications* (2020), 113193. https://doi.org/10.1016/j.eswa.2020.113193

[18] George Demiris and Brian K Hensel. 2008. Technologies for an Aging Society: A Systematic Review of "Smart Home" Applications. *Yearbook of medical informatics* 17, 01 (2008), 33–40. https://doi.org/10.1055/s-0038-1638580

[19] Mateusz Dubiel, Martin Halvey, and Leif Azzopardi. 2018. A Survey Investigating Usage of Virtual Personal Assistants. arXiv:1807.04606

[20] Patrick Ehrenbrink, Seif Osman, and Sebastian Möller. 2017. Google Now Is for the Extraverted, Cortana for the Introverted: Investigating the Influence of Personality on IPA Preference. In *Proceedings of the 29th Australian Conference on Computer-Human Interaction (OzCHI'17)*. ACM, 257–265. https://doi.org/10.1145/3152771.3152799

[21] Milan Erdelj, Enrico Natalizio, Kaushik R Chowdhury, and Ian F Akyildiz. 2017. Help from the Sky: Leveraging Uavs for Disaster Management. *IEEE Pervasive Computing* 16, 1 (2017), 24–32. https://doi.org/10.1109/MPRV.2017.11

[22] Denzil Ferreira, Jorge Goncalves, Vassilis Kostakos, Louise Barkhuus, and Anind K Dey. 2014. Contextual Experience Sampling of Mobile Application Micro-usage. In *Proceedings of the 16th International Conference on Human-computer Interaction with Mobile Devices and Services (MobileHCI'14)*. 91–100. https://doi.org/10.1145/2628363.2628367

[23] Lewis R Goldberg. 1992. The Development of Markers for the Big-five Factor Structure. *Psychological Assessment* 4, 1 (1992), 26–42. https://doi.org/10.1037/1040-3590.4.1.26

[24] Jonathan Grudin and Richard Jacques. 2019. Chatbots, Humbots, and the Quest for Artificial General Intelligence. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI'19)*. ACM, 209:1–209:11. https://doi.org/10.1145/3290605.3300439

[25] Jan Hartmann, Antonella De Angeli, and Alistair Sutcliffe. 2008. Framing the User Experience: Information Biases on Website Quality Judgement. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'08)*. ACM, 855–864. https://doi.org/10.1145/1357054.1357190

[26] Sozo Inoue, Naonori Ueda, Yasunobu Nohara, and Naoki Nakashima. 2016. Recognizing and Understanding Nursing Activities for a Whole Day with a Big Dataset. *Journal of Information Processing* 24, 6 (2016), 853–866. https://doi.org/10.2197/ipsjjip.24.853

[27] Michael Jordan. 2019. Artificial Intelligence–The Revolution Hasn't Happened Yet. *Harvard Data Science Review* (2019). https://doi.org/10.1162/99608f92.f06c6e61

[28] Julia Kiseleva, Kyle Williams, Jiepu Jiang, Ahmed Hassan Awadallah, Aidan C. Crook, Imed Zitouni, and Tasos Anastasakos. 2016. Understanding User Satisfaction with Intelligent Assistants. In *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval (CHIIR'16)*. ACM, 121–130. https://doi.org/10.1145/2854946.2854961

[29] Rafal Kocielnik, Saleema Amershi, and Paul N Bennett. 2019. Will You Accept an Imperfect AI?: Exploring Designs for Adjusting End-user Expectations of AI Systems. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI'19)*. ACM, 411:1–411:14. https://doi.org/10.1145/3290605.3300641

[30] Sari Kujala, Ruth Mugge, and Talya Miron-Shatz. 2017. The Role of Expectations in Service Evaluation: A Longitudinal Study of a Proximity Mobile Payment Service. *International Journal of Human-Computer Studies* 98 (2017), 51–61. https://doi.org/10.1016/j.ijhcs.2016.09.011

[31] Minae Kwon, Malte F Jung, and Ross A Knepper. 2016. Human Expectations of Social Robots. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI'16)*. IEEE, 463–464. https://doi.org/10.1109/HRI.2016.7451807

[32] Nicholas D Lane, Mu Lin, Mashfiqui Mohammod, Xiaochao Yang, Hong Lu, Giuseppe Cardone, Shahid Ali, Afsaneh Doryab, Ethan Berke, Andrew T Campbell, et al. 2014. BeWell: Sensing Sleep, Physical Activities and Social Interactions to Promote Wellbeing. *Mobile Networks and Applications* 19, 3 (2014), 345–359. https://doi.org/10.1007/s11036-013-0484-5

[33] Neal Lathia, Veljko Pejovic, Kiran K Rachuri, Cecilia Mascolo, Mirco Musolesi, and Peter J Rentfrow. 2013. Smartphones for Large-Scale Behavior Change Interventions. *IEEE Pervasive Computing* 12, 3 (2013), 66–73. https://doi.org/10.1109/MPRV.2013.56

[34] Hong Lu, Denise Frauendorfer, Mashfiqui Rabbi, Marianne Schmid Mast, Gokul T Chittaranjan, Andrew T Campbell, Daniel Gatica-Perez, and Tanzeem Choudhury. 2012. StressSense: Detecting Stress in Unconstrained Acoustic Environments Using Smartphones. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing (UbiComp'12)*. ACM, 351–360. https://doi.org/10.1145/2370216.2370270

[35] Ewa Luger and Abigail Sellen. 2016. Like Having a Really Bad PA: The Gulf between User Expectation and Experience of Conversational Agents. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI'16)*. ACM, 5286–5297. https://doi.org/10.1145/2858036.2858288

[36] Abhinav Mehrotra, Robert Hendley, and Mirco Musolesi. 2016. PrefMiner: Mining User's Preferences for Intelligent Mobile Notification Management. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp'16)*. ACM, 1223–1234. https://doi.org/10.1145/2971648.2971747

[37] Abhinav Mehrotra, Sandrine R Müller, Gabriella M Harari, Samuel D Gosling, Cecilia Mascolo, Mirco Musolesi, and Peter J Rentfrow. 2017. Understanding the Role of Places and Activities on Mobile Phone Interaction and Usage Patterns. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 3 (2017), 84:1–84:22. https://doi.org/10.1145/3131901

[38] Abhinav Mehrotra and Mirco Musolesi. 2017. Intelligent Notification Systems: A Survey of the State of the Art and Research Challenges. arXiv:1711.10171

[39] Abhinav Mehrotra, Veljko Pejovic, Jo Vermeulen, Robert Hendley, and Mirco Musolesi. 2016. My Phone and Me: Understanding People's Receptivity to Mobile Notifications. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI'16)*. ACM, 1021–1032. https://doi.org/10.1145/2858036.2858566

[40] Christian Meurisch, Bekir Bayrak, Florian Giger, and Max Mühlhäuser. 2020. PDSProxy: Trusted IoT Proxies for Confidential Ad-hoc Personalization of AI Services. In *2020 29th International Conference on Computer Communications and Networks (ICCCN'20)*. IEEE, 1–2. https://doi.org/10.1109/ICCCN49398.2020.9209655

[41] Christian Meurisch, Bekir Bayrak, and Max Muhlhauser. 2019. EdgeBox: Confidential Ad-hoc Personalization of Nearby IoT Applications. In *2019 IEEE Global Communications Conference (GLOBECOM'19)*. IEEE, 1–6. https://doi.org/10.1109/GLOBECOM38437.2019.9013520

[42] Christian Meurisch, Bekir Bayrak, and Max Mühlhäuser. 2020. Privacy-Preserving AI Services Through Data Decentralization. In *Proceedings of The Web Conference 2020 (WWW'20)*. ACM, 190–200. https://doi.org/10.1145/3366423.3380106

[43] Christian Meurisch, Maria-Dorina Ionescu, Benedikt Schmidt, and Max Mühlhäuser. 2017. Reference Model of Next-Generation Digital Personal Assistant: Integrating Proactive Behavior. In *Adjunct Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2017 ACM International Symposium on Wearable Computers (UbiComp/ISWC'17 Adjunct)*. ACM, 149–152. https://doi.org/10.1145/3123024.3123145

[44] Christian Meurisch, Sebastian Kauschke, Tim Grube, Bekir Bayrak, and Max Mühlhäuser. 2019. {P}Net: Privacy-preserving Personalization of AI-based Models by Anonymous Inter-person Similarity Networks. In *Proceedings of the 16th EAI International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services (MobiQuitous'19)*. ACM, 60–69. https://doi.org/10.1145/3360774.3360819

[45] Christian Meurisch, Benedikt Schmidt, Michael Scholz, Immanuel Schweizer, and Max Mühlhäuser. 2015. Labels: Quantified Self App for Human Activity Sensing. In *Adjunct Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2015 ACM International Symposium on Wearable Computers (UbiComp/ISWC'15 Adjunct)*. ACM, 1413–1422. https://doi.org/10.1145/2800835.2801612

[46] Jaroslav Michalco, Jakob Grue Simonsen, and Kasper Hornbæk. 2015. An Exploration of the Relation Between Expectations and User Experience. *International Journal of Human-Computer Interaction* 31, 9 (2015), 603–617. https://doi.org/10.1080/10447318.2015.1065696

[47] A Cristina Mihale-Wilson, Jan Zibuschka, and Oliver Hinz. 2019. User Preferences and Willingness to Pay for In-vehicle Assistance. *Electronic Markets* 29, 1 (2019), 37–53. https://doi.org/10.1007/s12525-019-00330-5

[48] Cristina Mihale-Wilson, Jan Zibuschka, and Oliver Hinz. 2017. About User Preferences and Willingness to Pay for a Secure and Privacy Protective Ubiquitous Personal Assistant. In *Proceedings of the 25th European Conference on Information Systems (ECIS'17)*. AIS Electronic Library, 32–47.

[49] Max Mühlhäuser, Christian Meurisch, Michael Stein, Jörg Daubert, Julius Von Willich, Jan Riemann, and Lin Wang. 2020. Street Lamps as a Platform. *Commun. ACM* 63, 6 (2020), 75–83. https://doi.org/10.1145/3376900

[50] Huseyin Naci and John PA Ioannidis. 2015. Evaluation of Wellness Determinants and Interventions by Citizen Scientists. *JAMA* 314, 2 (2015), 121–122. https://doi.org/10.1001/jama.2015.6160

[51] Richard L Oliver. 1980. A Cognitive Model of the Antecedents and Consequences of Satisfaction Decisions. *Journal of marketing research* 17, 4 (1980), 460–469. https://doi.org/10.1177/002224378001700405

[52] Richard W Olshavsky and John A Miller. 1972. Consumer Expectations, Product Performance, and Perceived Product Quality. *Journal of Marketing Research* 9, 1 (1972), 19–21. https://doi.org/10.1177/002224377200900105

[53] World Health Organization. 2001. *The World Health Report 2001: Mental Health: New Understanding, New Hope*. World Health Organization.

[54] Stefan Palan and Christian Schitter. 2018. Prolific.ac—A Subject Pool for Online Experiments. *Journal of Behavioral and Experimental Finance* 17 (2018), 22–27. https://doi.org/10.1016/j.jbef.2017.12.004

[55] Gaurav Paruthi, Shriti Raj, Seungjoo Baek, Chuyao Wang, Chuan-che Huang, Yung-Ju Chang, and Mark W Newman. 2018. Heed: Exploring the Design of Situated Self-Reporting Devices. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 3 (2018), 132:1–132:21. https://doi.org/10.1145/3264942

[56] Veljko Pejovic and Mirco Musolesi. 2014. InterruptMe: Designing Intelligent Prompting Mechanisms for Pervasive Applications. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp'14)*. ACM, 897–908. https://doi.org/10.1145/2632048.2632062

[57] Veljko Pejovic and Mirco Musolesi. 2015. Anticipatory Mobile Computing: A Survey of the State of the Art and Research Challenges. *Comput. Surveys* 47, 3 (2015), 47:1–47:29. https://doi.org/10.1145/2693843

[58] Zhenhui Peng, Yunhwan Kwon, Jiaan Lu, Ziming Wu, and Xiaojuan Ma. 2019. Design and Evaluation of Service Robot's Proactivity in Decision-Making Support Process. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI'19)*. ACM, 98:1–98:13. https://doi.org/10.1145/3290605.3300328

[59] Martin Pielot, Bruno Cardoso, Kleomenis Katevas, Joan Serrà, Aleksandar Matic, and Nuria Oliver. 2017. Beyond Interruptibility: Predicting Opportune Moments to Engage Mobile Phone Users. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 3 (2017), 91:1–91:25. https://doi.org/10.1145/3130956

[60] Mashfiqui Rabbi, Min Hane Aung, Mi Zhang, and Tanzeem Choudhury. 2015. MyBehavior: Automatic Personalized Health Feedback from User Behaviors and Preferences Using Smartphones. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp'15)*. ACM, 707–718. https://doi.org/10.1145/2750858.2805840

[61] Kiran K Rachuri, Cecilia Mascolo, Mirco Musolesi, and Peter J Rentfrow. 2011. SociableSense: Exploring the Trade-Offs of Adaptive Sampling and Computation Offloading for Social Sensing. In *Proceedings of the 17th Annual International Conference on Mobile Computing and Networking (MobiCom'11)*. ACM, 73–84. https://doi.org/10.1145/2030613.2030623

[62] Kiran K Rachuri, Mirco Musolesi, Cecilia Mascolo, Peter J Rentfrow, Chris Longworth, and Andrius Aucinas. 2010. EmotionSense: A Mobile Phones Based Adaptive Platform for Experimental Social Psychology Research. In *Proceedings of the 12th ACM International Conference on Ubiquitous Computing (UbiComp'10)*. ACM, 281–290. https://doi.org/10.1145/1864349.1864393

[63] Fariba Sadri. 2011. Ambient Intelligence: A Survey. *Comput. Surveys* 43, 4, Article 36 (2011), 36:1–36:66 pages. https://doi.org/10.1145/1978802.1978815

[64] Alireza Sahami Shirazi, Niels Henze, Tilman Dingler, Martin Pielot, Dominik Weber, and Albrecht Schmidt. 2014. Large-scale Assessment of Mobile Notifications. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'14)*. ACM, 3055–3064. https://doi.org/10.1145/2556288.2557189

[65] Ruhi Sarikaya. 2017. The Technology Behind Personal Digital Assistants: An Overview of the System Architecture and Key Components. *IEEE Signal Processing Magazine* 34, 1 (2017), 67–81. https://doi.org/10.1109/MSP.2016.2617341

[66] Benedikt Schmidt, Sebastian Benchea, Rüdiger Eichin, and Christian Meurisch. 2015. Fitness Tracker or Digital Personal Coach: How to Personalize Training. In *Adjunct Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2015 ACM International Symposium on Wearable Computers (UbiComp/ISWC'15 Adjunct)*. ACM, 1063–1067.

https://doi.org/10.1145/2800835.2800961

[67] Philipp M Scholl, Matthias Wille, and Kristof Van Laerhoven. 2015. Wearables in the Wet Lab: A Laboratory System for Capturing and Guiding Experiments. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp'15)*. ACM, 589–599. https://doi.org/10.1145/2750858.2807547

[68] Yu Sun, Nicholas Jing Yuan, Yingzi Wang, Xing Xie, Kieran McDonald, and Rui Zhang. 2016. Contextual Intent Tracking for Personal Assistants. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'16)*. ACM, 273–282. https://doi.org/10.1145/2939672.2939676

[69] Daniel Teghe and Kathryn Rendell. 2005. Social Well-being: A Literature Review. *School of Social Work and Welfare Studies, CQU: Rockhampton* (2005). https://doi.org/10.13140/RG.2.2.28891.26406

[70] Fani Tsapeli and Mirco Musolesi. 2015. Investigating Causality in Human Behavior from Smartphone Sensor Data: A Quasi-experimental Approach. *EPJ Data Science* 4, 1 (2015), 24. https://doi.org/10.1140/epjds/s13688-015-0061-1

[71] Iis P Tussyadiah and Dan Wang. 2016. Tourists' Attitudes Toward Proactive Smartphone Systems. *Journal of Travel Research* 55, 4 (2016), 493–508. https://doi.org/10.1177/0047287514563168

[72] Niels van Berkel, Chu Luo, Theodoros Anagnostopoulos, Denzil Ferreira, Jorge Goncalves, Simo Hosio, and Vassilis Kostakos. 2016. A Systematic Assessment of Smartphone Usage Gaps. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI'16)*. ACM, 4711–4721. https://doi.org/10.1145/2858036.2858348

[73] Mark Weiser. 1991. The Computer for the 21st Century. *Scientific American* 265, 3 (1991), 94–105. http://www.jstor.org/stable/24938718

[74] Neil Yorke-Smith, Shahin Saadati, Karen L Myers, and David N Morley. 2012. The Design of a Proactive Personal Agent for Task Management. *International Journal on Artificial Intelligence Tools* 21, 01 (2012), 1250004. https://doi.org/10.1142/S0218213012500042

[75] Bing Zhai, Ignacio Perez-Pozuelo, Emma AD Clifton, Joao Palotti, and Yu Guan. 2020. Making Sense of Sleep: Multimodal Sleep Stage Classification in a Large, Diverse Population Using Movement and Cardiac Sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 2 (2020), 1–33. https://doi.org/10.1145/3397325

[76] Sha Zhao, Julian Ramos, Jianrong Tao, Ziwen Jiang, Shijian Li, Zhaohui Wu, Gang Pan, and Anind K Dey. 2016. Discovering Different Kinds of Smartphone Users Through Their Application Usage Behaviors. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp'16)*. ACM, 498–509. https://doi.org/10.1145/2971648.2971696

[77] Mingyang Zhong, Jiahui Wen, Peizhao Hu, and Jadwiga Indulska. 2017. Advancing Android Activity Recognition Service with Markov Smoother: Practical Solutions. *Pervasive and Mobile Computing* 38 (2017), 60–76. https://doi.org/10.1016/j.pmcj.2016.09.003