Time2Stop: Adaptive and Explainable Human-AI Loop for Smartphone Overuse Intervention

Adiba Orzikulova
KAIST
Republic of Korea
adiorz@kaist.ac.kr

Han Xiao
Beijing University of Posts and Telecommunications
Beijing, Beijing, China
umihara@bupt.edu.cn

Zhipeng Li
Tsinghua University
Beijing, China
lizhipeng0603@gmail.com

Yukang Yan
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA
yukangy@andrew.cmu.edu

Yuntao Wang*
Tsinghua University
Beijing, Beijing, China
yuntaowang@tsinghua.edu.cn

Yuanchun Shi
Tsinghua University
Beijing, Beijing, China
shiyc@tsinghua.edu.cn

Marzyeh Ghassemi
Massachusetts Institute of Technology
Cambridge, MA, USA
mghassem@mit.edu

Sung-Ju Lee
KAIST
Republic of Korea
profsj@kaist.ac.kr

Anind K. Dey
University of Washington
Seattle, UW, USA
anind@uw.edu

Xuhai Xu†
Massachusetts Institute of Technology
Cambridge, MA, USA
orson.xuhai.xu@gmail.com

ABSTRACT

Despite a rich history of investigating smartphone overuse intervention techniques, AI-based just-in-time adaptive intervention (JITAI) methods for overuse reduction are lacking. We develop Time2Stop, an intelligent, adaptive, and explainable JITAI system that leverages machine learning to identify optimal intervention timings, introduces interventions with transparent AI explanations, and collects user feedback to establish a human-AI loop and adapt the intervention model over time. We conducted an 8-week field experiment (N=71) to evaluate the effectiveness of both the adaptation and explanation aspects of Time2Stop. Our results indicate that our adaptive models significantly outperform the baseline methods on intervention accuracy (>32.8% relatively) and receptivity (>8.0%). In addition, incorporating explanations further enhances the effectiveness by 53.8% and 11.4% on accuracy and receptivity, respectively. Moreover, Time2Stop significantly reduces overuse, decreasing app visit frequency by 7.0~8.9%. Our subjective data also echoed these quantitative measures. Participants preferred the adaptive interventions and rated the system highly on intervention time accuracy, effectiveness, and level of trust. We envision our work can inspire future research on JITAI systems with a human-AI loop to evolve with users.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in HCI**; *Interaction techniques.*

KEYWORDS

Just-in-time adaptive intervention, Smartphone overuse, Explainable AI, Human-in-the-loop

ACM Reference Format:
Adiba Orzikulova, Han Xiao, Zhipeng Li, Yukang Yan, Yuntao Wang, Yuanchun Shi, Marzyeh Ghassemi, Sung-Ju Lee, Anind K. Dey, and Xuhai Xu. 2024. Time2Stop: Adaptive and Explainable Human-AI Loop for Smartphone Overuse Intervention. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24), May 11–16, 2024, Honolulu, HI, USA.* ACM, New York, NY, USA, 20 pages. https://doi.org/10.1145/3613904.3642747

1 INTRODUCTION

The rapid advancement of technology has empowered the use of mobile devices to engage in almost every aspect of our lives. While bringing us convenience, smartphones also introduce numerous potential risks [16, 68]. Smartphone overuse is considered a major social problem as it adversely affects individuals' physical health (*e.g.*, headaches [11], chronic neck pain [94], sleep disturbance [46]); mental well-being (*e.g.*, anxiety and depression [5, 26], impaired cognitive abilities [92]); and social wellness (*e.g.*, distraction [57], family conflicts [79], degradation of academic and work performance [19]).

*Corresponding Author for Funding
†Corresponding Author

CHI '24, May 11–16, 2024, Honolulu, HI, USA

ACM ISBN 979-8-4007-0330-0/24/05
https://doi.org/10.1145/3613904.3642747

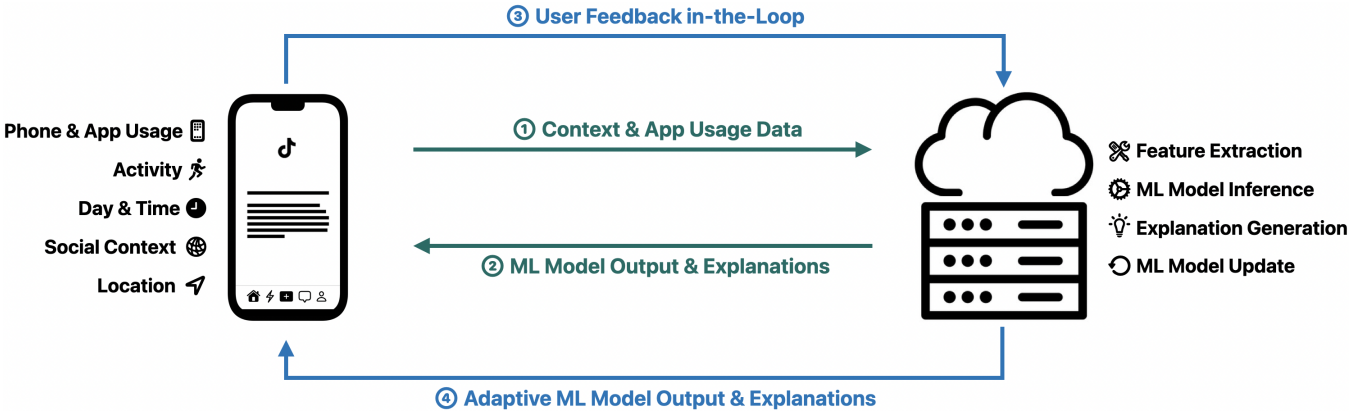

Figure 1: Time2Stop System Overview. The overall interaction flow consists of two loops. The first loop (green) includes: ① The mobile app continuously gathers contextual and app usage data (left) and transmits them to the cloud server. ② On the cloud server's end, feature extraction, ML model inference, and explanation generation occur (right). The ML model output and explanations are sent back to the user. The second loop (loop) includes: ③ In cases where the model predicts "overuse", an intervention would show up while allowing users to provide feedback. The feedback is then forwarded to the cloud server to update the ML model. ④ The updated ML model is subsequently employed to provide more personalized and adaptive interventions.

A plethora of research has been invested in designing and experimenting with various digital intervention tools and techniques to regulate smartphone overuse. These mechanisms promote digital well-being by informing users about their usage statistics [56, 62, 78, 82], restricting access to distracting apps [35, 38, 51, 90] or app functionalities [10, 54, 67]. While the proposed mechanisms were beneficial in enhancing self-awareness and reducing smartphone usage time, they primarily intervened based on simple criteria, *e.g.*, upon opening a specific app, at pre-determined intervals, or after achieving daily usage goals. However, due to the considerable variability of human behavior, interventions based on these basic criteria may not be optimal. For instance, users sometimes need to take a break, and blocking usage without considering such contexts could lead to sub-optimal designs and impact user experience. A system should offer intelligent interventions tailored to the user's preferences, app characteristics, dynamically changing context, and individual usage patterns.

In mobile health, Just-In-Time Adaptive Intervention (JITAI) was introduced as a promising technique that provides appropriate support at opportune times while dynamically adapting to users' internal and external context [63, 64]. Traditional JITAI-driven intervention systems incorporate a predefined set of rules (such as users' location) to determine the delivery time or content [25, 53, 73]. There has been initial research leveraging artificial intelligence (AI) and machine learning (ML) to deliver interventions, as AI can analyze large amounts of data and identify patterns that might not be captured through manual rule-setting [43, 61]. Furthermore, in a human-AI-loop setup, users can offer feedback to the AI, allowing the model to enhance its predictions and personalize interventions based on individual needs and behaviors.

However, very little work explores empowering AI-based JITAI with a human-in-the-loop setup [36, 49]. There is no prior work leveraging AI-based JITAI in the realm of smartphone overuse, not to mention the human-in-the-loop setup. Employing JITAI-based interventions for smartphone overuse is challenging as it requires a real-time ML pipeline (reacting in a few seconds when the user enters an app) and prompt adaptability to users' constantly evolving habits (updating the model on a daily basis).

Despite the accuracy of black-box AI models, they often face challenges of interpretability and transparency. This gave rise to the recent advance of Explainable AI (XAI) to help users comprehend AI systems' decisions, thereby fostering user trust and collaboration with AI [3, 72]. Recent intervention systems employed XAI to personalize education [31], manage stress [36], and set fitness goals [83]. However, there is no prior work integrating XAI into JITAI-based smartphone intervention. It can improve intervention delivery transparency, handle confusion caused by unexpected interventions, and cultivate users' trust through human-AI interaction.

No prior work has used AI-driven JITAI for smartphone overuse nor incorporated a human-in-the-loop setup. Moreover, integrating XAI into JITAI-based smartphone interventions remains unexplored. To address these gaps, we design and implement Time2Stop, **an intelligent, adaptive, and explainable smartphone overuse intervention system grounded in JITAI principles while taking user feedback in the loop**. Our system consists of four major parts (see Figure 1): (1) a smartphone-based sensing app to collect users' contexts and behavior, (2) a cloud-based ML pipeline that extracts behavior features, detects potential smartphone overuse behavior, and generates explanations, (3) an interface on local devices that introduces interventions when the ML model detects overuse behavior, provides intervention explanations, and collects user-provided feedback (*e.g.*, users' opinions on the accuracy of the intervention), and (4) a human-AI feedback loop that leverage users' reactions to update the ML model.

We conducted an eight-week field experiment (N=71) to deploy and evaluate the effectiveness of the two major characteristics of Time2Stop: (a) *Adaptive*, updating the model based on user feedback in the human-AI loop; (b) *Explainable*, providing feature explanations based on user behavior and model outcomes. Our findings demonstrate that interventions with the *adaptive* models significantly outperform both the basic (statistics-based) and the personalized ML (but non-adaptive) methods on smartphone overuse prediction accuracy (32.8∼55.5% relative advantage) and intervention receptivity (8.0∼29.0%). Moreover, incorporating *explanations* in interventions can further enhance the effectiveness (53.8∼97.5% on relative accuracy, 11.4∼39.6% on relative receptivity). From the perspective of smartphone usage behavior, our results indicate that app visit frequency was reduced significantly with the help of Time2Stop (7.0∼8.9%). We also observe an interesting nuance of explanations' effect on user behavior. Our qualitative results from the exit questionnaire and interview align with the quantitative findings, further supporting the advantage of Time2Stop. We discuss the mixed effects of explanations and the design considerations and ethical concerns of AI-based interventions.

The main contributions of our paper can be summarized as follows:

- We designed and implemented Time2Stop, an adaptive and explainable JITAI-grounded intervention system for smartphone overuse. Time2Stop performs real-time inference on smartphone overuse behavior, introduces just-in-time intelligent intervention with explanations, and evolves based on users' feedback.
- We conducted a longitudinal field experiment with micro-randomized trials to demonstrate the effectiveness of empowering interventions to be adaptive and explainable. Our results show that Time2Stop significantly outperforms baseline techniques.
- We share the lessons learned, and discuss the design considerations and ethical concerns when creating AI-based smartphone intervention systems with humans in the loop.

We envision that empowering AI-based JITAI with both human-in-the-loop and AI explanations can go beyond smartphone overuse. When focusing on another application, careful design of the human-AI-loop (*e.g.*, updating models with user feedback in our case) and the integration of an appropriate level of explanation (*e.g.*, highlighting feature types in our case) is necessary.

2 BACKGROUND

We first summarize existing research in smartphone overuse intervention techniques and just-in-time adaptive intervention (JITAI) methods. We also provide a brief overview of explainable AI (XAI), and emerging research in the intersection of XAI and JITAI intervention domains.

2.1 Smartphone Overuse and Intervention

Excessive smartphone usage has been connected to a variety of undesirable effects, such as distraction [57], anxiety and depression [26], neck pain [94], and disruptions in sleep patterns [46]. In response to the negative effects of smartphone overuse, there has been a wide range of commercial products and research solutions. For example, ScreenTime [78] on iOS and Digital Wellbeing [82] on Android are built-in tools designed to assist users in tracking app usage and setting usage limits. In addition, there are also various third-party apps for overuse intervention, such as Forest [70], Digital Detox [66], and StayFree [2].

Within the academic sphere, researchers have proposed a large array of works in the smartphone overuse intervention domain [10, 41, 52, 54, 62, 67]. These methods can be generally divided into two categories: (1) sending notifications or reminders [27, 39], and (2) blocking user access to apps or phones [35, 38, 90]. The first category aims to softly persuade users to limit digital consumption. For example, MyTime [27] informs users about their usage time and sends notifications upon reaching their time limit. NUGU [39] leveraged social effects by visualizing smartphone usage among social groups via a scoreboard. The second category is more restrictive, aiming to introduce a higher interaction cost and a gulf of instant gratification. For instance, LockNType [35] adds a typing task before users can access their apps to trigger System 2, *i.e.*, the reasoning and analytical system in the Dual Process Theory [30]. Building on top of this work, TypeOut [90] integrates the typing task with self-affirmation to effectively mitigate smartphone overuse.

Other than dividing interventions based on their restrictiveness, another line of work was devoted to building smartphone interventions at different granularities: device-level, app-level, and feature-level. A study by Roffarello *et al.* [62] found that intervening at the app-level is more effective compared to device-level, as the former can generate more precise and interpretable statistics for users. Orzikulova *et al.* [67] investigated app-level (*e.g.*, restricting Instagram and YouTube apps) and feature-level (*e.g.*, limiting the usage of app features such as viewing suggested feed on Instagram and watching shorts on YouTube) interventions on mobile social media apps. The results indicated that feature-level restrictive interventions were particularly effective in reducing the time spent on passive phone usage (*e.g.*, watching short videos).

While these intervention techniques are beneficial in enhancing user awareness and reducing phone use time, they rely on basic conditions (such as being triggered upon opening an app) or simple parameters specified by users (such as the daily usage limit). However, users' behavior is changing dynamically and these manual rules are often outdated. For an intelligent smartphone intervention system, it is essential to account for user's preferences, contexts, and smartphone usage patterns, so that it can achieve a good intervention performance continuously. To address this gap, Time2Stop implements real-time adaptability to accommodate users' evolving contexts and behavior.

2.2 Just-in-Time Adaptive Interventions

In the context of mobile health, JITAI is an emerging intervention design methodology that seeks to deliver tailored and timely support by dynamically adjusting to an individual's internal and contextual conditions [63, 64]. For a JITAI to be effective, intervention needs to be delivered when the user is both *vulnerable* and *receptive* [60, 64]. Vulnerability denotes a period during which individuals are more susceptible to experiencing negative health

consequences (*i.e.*, overusing smartphones in our case), whereas receptivity pertains to their ability to receive and process provided interventions (*i.e.*, accepting intervention and stopping using phones).

JITAI-driven systems may either be rule-based [25, 73] or AI-based [36, 43, 61]. Rule-based JITAI relies on predefined sets of rules and conditions to trigger interventions, typically established by domain experts. For example, Gustafson *et al.* [25] designed a JITAI system for alcohol consumption that will trigger intervention when users approach high-risk locations such as bars. Lukoff *et al.* [53] designed a proof-of-concept system with adaptable commitment interfaces for digital well-being.

In contrast, AI-based JITAI leverages large-scale user behavior data and trained AI/ML models to detect appropriate intervention timing and personalize interventions. For example, Saponaro *et al.* [74] developed two types of AI-based JITAI systems (population-based, personalized) to reduce users' sedentary behavior. Until recently, very few studies explored empowering AI-based JITAI systems with user-in-the-loop to involve user feedback or reactions [36, 49]. Mishra *et al.* [60] implemented an adaptive chatbot that updates the ML model based on users' receptivity to encourage physical activity. Rabbi *et al.* [71] and Liao *et al.* [49] integrated reinforcement learning algorithms into JITAI systems to adapt the model to each individual for more effective physical activity intervention.

To our best knowledge, there have been very few prior studies exploring JITAI-based intervention for smartphone overuse [53], not to mention the advanced version that leverages users' reactions in the human-AI loop. There is a set of technical challenges for such a system. First, the machine learning pipeline needs to respond within seconds when the user opens an app. Second, the model needs to promptly adapt to users' ever-shifting habits. Time2Stop aims to address these challenges by establishing a real-time human-AI loop system.

2.3 XAI and Interventions

Although black-box AI models excel in making complex predictions and handling intricate tasks, they often encounter challenges with interpretability and transparency, making it difficult for users to understand how the models arrive at specific decisions or predictions. This reflects the recent advance of explainable AI (XAI). By providing explanations for AI-driven decisions, XAI not only helps humans comprehend the rationale behind AI system outputs, but also instills a sense of trust and confidence in these systems [3, 72, 93]. Recent advances in XAI research have not only served AI/ML practitioners and data scientists, enabling them to engage in model debugging and model behavior inspection [33, 50], but have also extended to domain experts in diverse fields [14] and end-users [6]. In the context of smartphone overuse detection and intervention, by showing why the users need to stop using certain applications, explanations can help users understand AI decisions and their own device usage patterns. Moreover, XAI has the potential to activate System 2 within the Dual Process Theory [29], clarifying the reasoning behind interventions targeting smartphone overuse [80]. These explanations stimulate users' deliberate, analytical thinking (System 2) and introduce appropriate reliance on AI [80]. This active engagement prompts users to reflect on their usage habits and make conscious adjustments, potentially leading to changes in their smartphone usage behavior.

Despite the considerable attention given to the field of XAI, research at the intersection of XAI and JITAI remains limited. Woźniak *et al.* [83] observed that presenting users with both algorithm-derived fitness goals and a clear explanation for the recommendation could increase their trust towards the recommended goal. MindScope [36] is a stress management system providing different explanation levels. The results indicated that elaborate explanations helped users understand stress-related events, while categorical explanations allowed them to interpret stressors from their unique perspectives. In our case, Time2Stop integrates XAI into JITAI to enhance the intervention delivery transparency, effectively address users' confusion about unexpected smartphone interventions, and foster user trust through seamless human-AI interaction.

3 TIME2STOP DESIGN

We developed Time2Stop — an intelligent, adaptive, and explainable intervention system for smartphone overuse. Grounded in JITAI principles, the main contribution of our system is the integration of continuous user feedback to form a human-AI loop. The architecture of the Time2Stop system comprises two main building blocks: the ML pipeline to predict smartphone overuse (Section 3.1), and the intricate design of interventions to be shown to users when overuse is predicted (Section 3.2).

This section provides an overall introduction to these core system components and the overall intervention flow (Section 3.3).

3.1 Machine Learning for Smartphone Overuse Prediction

Constructing an ML-driven JITAI system for predicting smartphone overuse and triggering intervention needs careful design across four key aspects: (1) feature design, (2) label collection mechanism, (3) adaptive model updates, and (4) explanation generation. In this section, we offer an in-depth exploration of each aspect.

3.1.1 Feature Design. We design a set of five passive sensing feature categories [59, 87, 89] to capture smartphone overuse behavior:

(a) **Phone and App Usage**. Understanding smartphone overuse requires a thorough analysis of usage patterns. We investigate both the high-level phone usage and the low-level app usage pattern. For phone usage, we track screen interactions and battery status. Screen interactions encompass phone unlock frequency and duration, calculated from screen-on/off events. As for battery usage, we extract battery consumption rate, charge, and discharge durations. For app-related features, our approach includes statistical metrics (count, min, max, mean, standard deviation, sum) linked to app visit frequency and time spent. Additionally, fine-grained user interface interactions (e.g., scrolling, tapping) provide insights into smartphone overuse. We use UI-event-driven features, gathering data on quantities and proportions of events like scrolling, clicking, focusing, and window state changes. Moreover, we also include the count and diversity of notifications.

(b) **Activity**. Users' interaction with the environment presents another pivotal factor intertwined with smartphone overuse. Concerning attributes related to physical activity, we examine stationary and mobile durations. Furthermore, ambient light offers insights

into the user's specific location (such as the room) and may vary depending on the time of day. In the case of ambient light, we extract statistical lux-related features.

(c) **Social Context**. Users might reduce smartphone interaction in specific social settings, such as when with friends or peers. We primarily focus on text message-derived features (*e.g.*, first message time, top contacts) and Bluetooth signals as a proxy of social contexts (*e.g.*, mean and standard deviations of scans, unique device counts). While we initially considered call-related features, we excluded them due to limited usage among participants.

(d) **Location**. Smartphone usage can also be tied to specific places. To capture this, we extract diverse location-based features, including location type, variance, entropy, time at the most-visited places, time at home, and duration of location stays. We also incorporate statistical WiFi data features involving metrics such as scans from visible and connected access points.

(e) **Time**. Previous work has highlighted different smartphone usage between times (*e.g.*, night *vs.* daytime) [1, 85]. Therefore, we further include an objective temporal feature to capture this.

3.1.2 Label Collection Mechanism. Time2Stop employs supervised learning for smartphone overuse prediction, which needs labeled data to train the model. We gather labels for two purposes: (i) to build the initial ML model and (ii) to update and adapt the ML model over time. Here we present the label collection mechanism for the first step, as shown in Figure 2 (Left). Further information about the data collection approach to update the ML model is provided in Section 3.1.3 and 3.2.

Our label collection mechanism harnesses Ecological Momentary Assessment (EMA) methodology [75] by presenting a prompt to ask users to report whether they are overusing their phones. We collect user responses and use them as labels to train our model. Labels are collected in-the-moment in real-time as users engage with their phones. This instantaneous label collection is crucial as it ensures timely, contextually relevant labels while users' app usage memory and experience remain fresh.

We designed three distinct in-the-moment label collection rules: entry-moment, leaving-moment, and during usage. Previous studies in smartphone notification management [28, 32, 69] showed the effectiveness of breakpoint-based notification delivery. Inspired by these works, we align optimal notification instances with task-switching moments, corresponding to accessing (i.e., entry-moment) and leaving a monitored app (i.e., leaving-moment prompts). In the smartphone overuse domain, empirical investigations revealed that once involved with a potentially addictive smartphone app, stopping usage becomes challenging, resulting in unforeseen excessive usage [34, 40]. To address these instances of overuse, we also incorporate label collection during usage, asking users to provide labels every 10 minutes during monitored app usage. Moreover, to alleviate the labeling burden, we implement a cool-down interval. This prevents additional label collection prompts within a cool-down period if a user recently provided a label. In addition, we also provide a post-hoc labeling process, where users can go back to check any missing annotations of their app usage history. [1]

3.1.3 Adaptive Model Updates. We leave the specific ML model choice for the result section (Section 6) as it needs empirical evaluation. Here, we focus on the model updates first. To tailor the ML model to each individual and accommodate their evolving behaviors, dynamic context, and smartphone usage patterns, we regularly update the ML model. We collect user feedback on the intervention as new labels. Together with the corresponding contextual data, these data can be used to update the ML model. It is noteworthy that the model update is personalized, *i.e.*, we train and update a model for each user.

Updating ML models involves deciding *when* to update the model and *how* to update the model. For the when part, while it's possible to instantly update the model with each new data point, frequent model updates will incur expensive computational costs, causing delays in intervention delivery. We conduct ML updates daily from 12 AM to 1 AM to retain the real-time aspect of just-in-time interventions. This process, taking 2 hours on average, ensures that the adapted models are available by the morning of each intervention day.

How to update the model is another crucial design consideration. A simple approach would involve re-training the model with equal weights for all data samples, treating historical and current user behavior equally. However, this method is sub-optimal as it fails to account for changing user dynamics. In our approach, we adopt decay-based sample weight assignment, *i.e.*, recent data will have relatively higher weights. Specifically, we adopted a linear decreasing assignment from 1.0 to a minimum cap of 0.5. Based on empirical testing with pilot experiments, we assign the weight of the most recent day as 1.0, and it decreases linearly every half-week until it reaches 0.5. This can help the re-trained model adapt to evolving conditions and smartphone usage behaviors. By emphasizing recent observations, the model becomes more adaptive, effectively capturing current trends while gradually reducing the impact of outdated information. We discuss other model update methods, such as reinforcement learning, as future work in Section 7.

3.1.4 Model Explanation. We provide explanations derived from model predictions to enhance users' understanding of the ML-based intervention system's decisions and foster trust and collaboration with AI [3, 72]. These explanations are generated based on the top features contributing to an "overuse" prediction. We designed two explanation detail levels: high and low. A high-level explanation represents the feature category. As for the low-level explanation, a straightforward option is to use the actual feature name. However, our internal testing found that it introduced unnecessary details and cognitive load. Therefore, we simplify and abstract the raw feature name into a feature description (see Appendix A). For example, consider the location feature "time spent at the second most frequent location". The high-level explanation is "location", and the low-level explanation is "time at frequent locations". By default, users will see the high-level explanations and can access more detailed, low-level explanations if they are interested.

[1] We recognize that such a label collection mechanism may intervene and affect phone usage behavior. Therefore, in Section 5, we intentionally inserted a break week between the label collection and intervention deployment to reduce the impact.

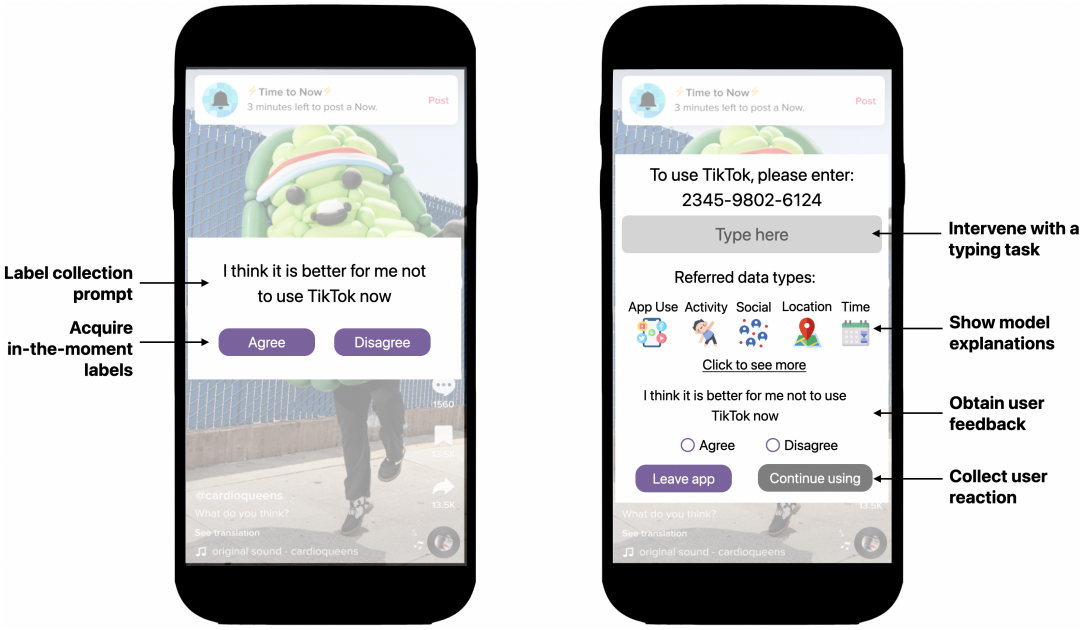

Figure 2: In-the-Moment Labeling and Intervention Interfaces. (Left) In-The-Moment Label Collection Interface; (Right) Time2Stop Intervention Interface. It encompasses four key components from top to bottom: (1) typing-based intervention task, (2) ML model explanations highlighting feature categories aligned with the model's output, (3) collection of user feedback – this is an optional question that users can choose to respond or ignore, and (4) user actions.

3.2 Intervention Design

The JITAI-based intervention system aims to provide *accurate and timely support* while *accommodating shifts in user context and conditions*, as discussed in Nahum *et al.*'s work [65]. Following these principles, we develop an intervention mechanism based on a typing task (offering the right support level). These interventions are triggered by an intelligent ML model detecting instances of "overuse" (optimal timing). Concurrently, user feedback is collected to enhance adaptation to individual user conditions and context (accommodation). This feedback loop subsequently drives updates to the ML model. Meanwhile, we also provide explanations derived from the model predictions.

3.2.1 Intervention Mechanism. The majority of prior work in the smartphone intervention domain provides interventions by either sending notifications/reminders [27, 40] or employing app access restrictions [35, 38]. However, notification-based interventions can be easily circumvented, while excessive restriction may agitate users and lead to counterproductive outcomes. We followed the previous work to balance intervention efficacy and usability and leveraged a typing-based intervention mechanism [35, 90] as interaction friction. Users are asked to input specific digits before they can proceed to use a monitored app. The digits are randomly generated within each intervention instance. Users can exit the application anytime and return to the home screen. Prior work [35] suggested that a typing task with a medium workload (10 - 20 digits) was effective and usable. Considering this, we designed a typing task comprising 12 digits. Note that the specific intervention mechanism is not the main focus of our paper, and we envision our adaptive and explainable system can be integrated with other mechanisms easily.

3.2.2 Intervention Timing. Our ML model decides whether to intervene based on users' current context and app usage behavior. Users first select the apps for which they want to receive the intervention (*i.e.*, monitored apps), and the intervention will only focus on these apps. When the model predicts "overuse", an intervention interface appears, as shown in Figure 2 (Right). Moreover, another design choice before triggering intervention involves determining the frequency of feature extraction and model prediction. Previous Just-in-Time (JIT)-based smartphone overuse techniques triggered interventions when users launched a monitored app [35, 51], or when the duration of a target app usage reached a predefined threshold [27, 67]. Our design takes both the launching moment and the usage period into account. We opt to initiate the feature extraction and model prediction process both upon app launch and periodically while the target app is in use. We empirically set the prediction interval as 5 minutes based on our pilot study. We further defined a 10-minute cool-down period after triggering an intervention to avoid a disrupted user experience.

3.2.3 User Feedback to Update Model. To adapt to dynamic shifts in user context and app usage behavior, we update each individual's ML model regularly. This entails obtaining fresh labels during the intervention period. One straightforward approach is employing the same label collection mechanism for constructing the initial model (see Section 3.1.2). However, this would considerably hinder system usability. Users would have to contend with both labeling prompts and intervention notifications. We integrated user labeling

within the intervention interface (Figure 2 Right) to address this issue.

When the intervention pops up, users are encouraged to provide feedback with a simple click to indicate whether they are overusing the phone. We design the labeling prompt with simplicity while ensuring it provides guidance to identify instances of smartphone overuse. The phrasing is deliberately structured to avoid potentially eliciting negative feelings regarding users' behavior. Rather than posing a direct query about smartphone or app overuse, users are prompted to indicate their agreement or disagreement with the statement: "I think I shouldn't use `AppName` now." In cases of agreement, the data point is categorized as "overuse", which can be used to reinforce the ML model; conversely, in instances of disagreement, it is classified as "not overuse", which can serve as a correction to the model. Once we receive feedback, we utilize them as new labels to update the ML model, following the design we introduced in Section 3.1.3. Note that this is an optional question, and users are not forced to respond. This process can capture the false positive cases, *i.e.*, an intervention pops up when users are not overusing their phones. Moreover, users can also leverage the post-hoc labeling to provide feedback on false negative cases, *i.e.*, an intervention does not pop up when they are overusing their phones.

3.2.4 Model Explanations. As detailed in Section 3.1.4, our explanation framework generates explanations at two levels: high and low. Previous XAI-based JITAI work in stress management by Kim *et al.* [36] showed that although most users favored more detailed explanations, such low-level explanations could potentially undermine the system's trustworthiness. Based on these findings, we decided to highlight the categories of essential features, such as "location", "activity", "app usage" (see Figure 2 Right). The interface only presents the top three crucial feature categories for the ML model inference and hides other categories to avoid confusion. We use their high-level explanations as icons in Figure 2. Furthermore, we provide low-level feature descriptions for users seeking deeper insights by clicking the "Click to see more" button.

3.3 Intervention Flow

Combining the ML and intervention design in Section 3.1 and 3.2, the intervention flow of Time2Stop is visualized in Figure 1. There are two loops within the flow: (1) the inner loop (green) is dedicated to the ML model inference process, and (2) the outer loop (blue) manages the ML model update process.

In the inner loop, ML model inference is performed through a sequence of steps. ① Contextual and app usage data are initially collected by the mobile app (depicted on the left) and transmitted to the cloud server. ② Here, the cloud server pre-processes raw data, extracts features, performs inference, obtains prediction output, and generates corresponding explanations (depicted on the right of Figure 1). The output of the model's prediction and explanations are then relayed to the user. In cases where the model predicts "overuse", the intervention interface (as illustrated in Figure 2) will pop up.

Conversely, the outer loop takes charge of the ML model update through user feedback and model enhancement cycles. ③ When the interface appears, users can provide feedback indicating whether they are overusing their phones. ④ This feedback is then transmitted to the cloud server, where new labels and features are re-trained. The updated ML model is then employed to generate more tailored and adaptive predictions, which are conveyed back to the user.

4 SYSTEM IMPLEMENTATION

Based on the system design in Section 3, we then introduce the implementation details of Time2Stop. We instantiated Time2Stop on Android OS (end-user side) and a server (cloud side), as shown in Figure 3. We conducted a one-week pilot field study with four authors of this paper to debug and finalize the system implementation, which includes the sensing platform (Section 4.1), the intervention interface (Section 4.2), and the ML pipeline (Section 4.3).

4.1 Context Sensing

To obtain the data we mentioned in Section 3.1.1, we leverage AWARE, an open-source passive sensing platform designed for behavioral data collection [15]. Our data collection includes multiple sensor streams: location, Bluetooth, Wi-Fi, network, light, screen activity, activity recognition, and communication (including SMS and calls). We further build our custom app usage tracker with Android's `AccessibilityService` API [12] that adeptly identifies the start and end of app sessions, dynamically monitors time allocation and visit frequencies, captures notifications from monitored applications, and records fine-grained user interactions with monitored apps, including scrolling, clicking, focusing, and window state changes.

4.2 Intervention Interface

We have introduced the interface design in Section 3.2, making the Android implementation straightforward. Moreover, the interface is implemented as an `AlertDialogue`, which becomes an overlay on top of the monitored app. When users enter the displayed random digits correctly into the input form and click the "Continue using" button, the overlay window is dismissed, and users are allowed to use the app. Conversely, upon clicking the "Leave app" button, the phone programmatically returns to the home screen. Users' reactions to the intervention will not impact the content in the app.

4.3 Machine Learning Pipeline

Our ML pipeline consists of three parts: (1) model inference, (2) model update, and (3) explanation generation. The technical details of these components are described in Figure 3. The end-user side is a mobile app running on Android OS, and the cloud side consists of a web app (Flask), a back-end (Redis and Celery), and a database (MySQL). The upper figure describes the model inference, and the lower sub-figure describes the model update.

4.3.1 Model Inference. The pipeline contains seven steps. ① The Android client posts the contextual data to the cloud web server with the frequency described in Section 3.2.2. ② A Flask-based web app manages a task queue that handles the arrived tasks. Once the task arrives at the cloud, the web app enqueues the task and dequeues in a first-in-first-out (FIFO) manner. ③ The back-end processes the raw data by imputing missing values, normalizing the raw values, and extracting features. ④ It also performs the inference

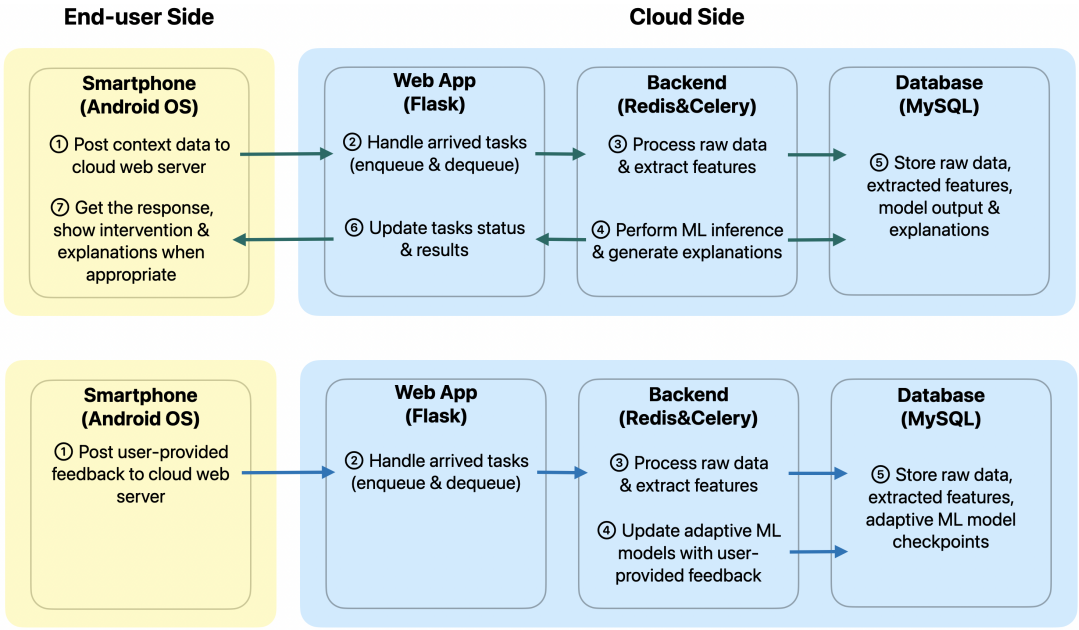

Figure 3: Overview of System Implementation. (Top): Model Inference. (Bottom): Model Update Leveraging User Feedback.

using the ML model to obtain overuse prediction. Explanations are generated using the SHAP method (see Section 4.3.3) [55]. ⑤ Raw data, extracted features, model outputs and generated explanations are then stored in the cloud database. ⑥ Once model outputs and explanations are ready, the web app updates the task status and the results so that the client can pick it up. ⑦ The Android client sends a request to obtain the results. In our pilot study, most of the responses arrived within 3 seconds. If the user is still using the monitored app when the results arrive, it checks the model output. If the prediction is "overuse", the Android client will pop up the intervention, together with the feature explanations. Otherwise, no intervention will show up.

4.3.2 Model Update. As introduced in Section 3.1.3 and 3.2.3, Time2Stop updates the model on a daily basis. This pipeline includes five steps. ① User-provided feedback is stored in the mobile app and sent to the cloud server. The next two steps of task handling (②) and feature extraction (③) are similar to the ones in model inference. ④ Next, the adaptive ML model is re-trained, using the user-provided feedback as new labels. We adopt the weight assignment introduced in Section 3.1.3 during the re-training. ⑤ Lastly, all data, extracted features, and the new model checkpoint are stored in the database.

4.3.3 Explanation Generation. To interpret the model predictions, we measure feature importance with SHapley Additive exPlanations (SHAP) [55], an XAI method that computes the impact of each feature on prediction outcomes. We rank the features based on their importance and obtain the corresponding high-level and medium-level explanations introduced in Section 3.1.4 and 3.2.4. These explanations are sent to the user along with the model outcome during the model inference.

Table 1: Multiple Intervention Types with Characteristics. The last row represents our complete Time2Stop system with ML-powered adaptive and explainable JITAI.

Intervention Type	Characteristics		
	ML-based	Adaptive	Explainable
Control	✗	✗	✗
Personalized	✓	✗	✗
Adaptive-wo-Exp	✓	✓	✗
Adaptive-w-Exp (*i.e.*, Time2Stop)	✓	✓	✓

5 FIELD EXPERIMENT

To investigate how AI-powered intelligent and explainable JITAI can affect smartphone overuse in real-life scenarios, we conducted an 8-week field experiment using Time2Stop. Our study aims to evaluate both the adaptive aspect and explainable aspect of Time2Stop, which requires careful experiment design (Section 5.1). We then introduce our field experiment procedure (Section 5.2) and participants (Section 5.3).

5.1 Experiment Design

5.1.1 How to Evaluate Adaptive and Explainable Interventions? To assess the efficacy of the adaptive and explainable components, we devised four distinct intervention types, each taking one step more advanced than the previous method (see Table 1).

(1) ***Control.*** This was a baseline method. It intervened with users simply based on probability (*e.g.*, a user might receive intervention when launching an app and every five minutes in 30% of the cases). The individual probability of the intervention was derived from the

user-provided labels during the first phase of the experiment (the modeling phase, see Section 5.2).

(2) ***Personalized***. This method added the ML component on top of *Control*, using the data collected during the modeling phase. To ensure a *Personalized* model aligned with each user's behavioral patterns while leveraging the rich data from other users, greater emphasis was placed on the user's own data by assigning it higher weights than the data collected from others. Through empirical tests, the weight for self-data was set at 1.0, while others' data received a weight of 0.1. The personalized model remained static and unchanged throughout the intervention period.

(3) ***Adaptive-wo-Exp***. This method further added the adapting component on top of *Personalized*. The model underwent a similar training procedure as *Personalized* at first. It also involved daily model re-training and updates, using continuous user feedback and corrections in response to intervention prompts, as we introduced in Section 3.1.3.

(4) ***Adaptive-w-Exp***. Finally, this method added the explanation component on top of *Adaptive-wo-Exp* and completed the whole Time2Stop system. The model of *Adaptive-w-Exp* was identical to that of *Adaptive-wo-Exp*. The only distinction was that *Adaptive-w-Exp* provided ML output explanations in the intervention interface, as introduced in Section 3.1.4 and shown in Figure 2.

Note that both *Adaptive-wo-Exp* and *Adaptive-w-Exp* fell into the category of adaptive models. Other than *Adaptive-w-Exp* that displayed explanations, the interface of the other three types was exactly the same to reduce bias.

5.1.2 Micro-Randomized Trials. Considering the sample size to compare four groups, we adopted a within-subject design. Specifically, we employed Micro-Randomized Trials, which is an experimental design technique optimized for JITAI-grounded intervention within the mHealth domain [37]. Instead of having users go through different experiment groups one by one, this method proposes to randomize the groups with smaller units (*e.g.*, daily or each intervention), so that the effect of potential confounding variables can be reduced.

In our case, we altered the intervention type among the four types on a daily basis, and each participant experienced only one type of intervention every day. In order to minimize the order effect, we employed the Latin Square design (n=4) [13] to diversify the intervention altering order. During our study onboarding sessions, we briefly introduced the four intervention types to users, but they were not informed of the specific order or dates for the four intervention types during the field study. This was also designed to reduce cognitive bias.

5.1.3 Evaluation metrics. We focused on four quantitative metrics to evaluate the performance of Time2Stop and the other three intervention types. The first two were about the model performance: (1) intervention accuracy, (2) intervention receptivity. The other two focused on its impact on users' phone usage patterns: (3) app usage duration and (4) app visit frequency.

Specifically, intervention accuracy represented the proportion of interventions that were marked as "correct" by users among the total number of intervention pop-ups. Note that this is a subjective algorithm measure instead of an objective measure, as there is no way to obtain an objective ground truth of overuse. Intervention receptivity, on the other hand, referred to users' reaction after encountering interventions, which included stopping usage (*e.g.*, returning to the home screen, triggering a screen-off event by locking the phone) or continuing usage. Instances where users quit the app were considered receptive interventions, while instances of continued usage were designated as non-receptive interventions. The other two metrics of app usage duration and visit frequency were calculated from the collected app usage log.

For qualitative metrics, we revealed the exact dates for each intervention type to users at the end of the study. We highlighted the latest four days to help users recall their experience with the four different techniques, as they had the most fresh memory. Then, we distributed a final questionnaire, asking them to rank the four types based on their preferences, as well as their perceived accuracy, effectiveness, and level of trust in different intervention types. Moreover, we conducted semi-structured exit interviews with participants to collect their feedback and intervention preferences. Our interview started with questions: "What do you think of the four intervention techniques? What's the reason behind your preference ranking? What do you think of the explanations coming with interventions?" For participants with low intervention accuracy and receptivity, we also asked about their thoughts and reactions towards intervention. We then followed the participants' lead and followed up with more detailed questions. The interviews were recorded, and three researchers followed the procedure of thematic analysis [7] to independently analyze and code the data. Then, they met, discussed, and iterated the coding until convergence.

5.2 Procedure

Our field experiment consisted of eight weeks, as shown in Figure 4. After the orientation and onboarding session, our field deployment experiment followed a sequence of four phases: (1) an initial modeling phase involving label collection lasting for two weeks, (2) a one-week break phase, (3) a subsequent week dedicated to baseline data collection without any intervention, and (4) a final four-week intervention phase with the design of micro-randomized trials.

In the modeling phase, we passively collected contextual and app usage data along with user-provided labels, using the label collection mechanism (see Section 3.1.2). These data points were used to calculate the individual probability (for *Control*) and train the initial ML models (for *Personalized*, *Adaptive-wo-Exp*, and *Adaptive-w-Exp*). To mitigate carry-over effects inherent in label collection, we incorporated a designated break phase. Then, we proceeded with the baseline week, during which we gathered baseline app usage data (usage duration and visit frequency of monitored apps) when there was no intervention. This data would serve as a comparative benchmark against diverse intervention types. Finally, during the intervention stage, interventions were introduced to users. User feedback and the accompanying behavioral data were collected (see Section 3.2.3) to update the *Adaptive-wo-Exp* and *Adaptive-w-Exp* models. Since users could still provide feedback during the *Control* and *Personalized*, these data were also collected to update the adaptive models.

At the end of the intervention phase, the intervention order was presented to participants. They then filled out the questionnaire

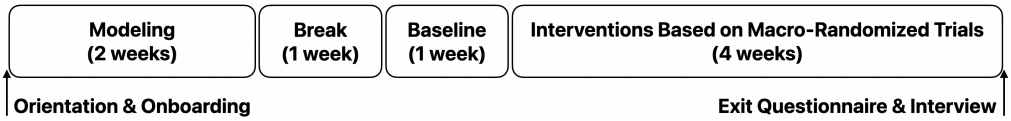

Figure 4: Field Experiment Flowchart

and completed the exit interview. They were compensated up to $50 based on their study compliance.

5.3 Participants

We posted a call for participation on large university community forums, together with a survey including basic demographics and a Smartphone Addiction Scale (SAS, score ranging from 33 to 198) [44]. We selected participants who used an Android smartphone as their primary phone and had a high SAS score (>120). 176 participants met the criteria. 127 of them attended the onboarding session. Among these participants, 49 discontinued their participation during the field study. Out of the 49 discontinued participants, 20 chose to exit the study citing personal reasons, 17 encountered software and hardware issues, 8 experienced compatibility concerns, 3 raised privacy issues, and 1 attributed their departure to battery concerns. Seven participants whose sensor or usage data only covered three or fewer intervention types were also eliminated from the analysis. In total, 71 participants (48 females, and 23 males, aged 21.8 ± 2.3, from 18 to 27) completed the whole study and provided high-quality data. Our analysis results were based on these participants.

6 RESULTS

Throughout our field experiment, we collected 497,458 minutes of usage data for 149 monitored apps (17 ± 5 apps per person) from 207,898 app sessions. App categories of entertainment, social media, and shopping emerged as the most frequently selected app categories. In total, we collected 75,670 ground truth labels during the modeling phase. 60.5%, 24.5%, and 14.9% of them were collected at the entry, using, and exit stages. During the intervention phase, we captured 47,939 intervention encounters, among which we collected 39,188 additional labels from user feedback. These data were used for our quantitative analysis. We also investigated the qualitative data from questionnaires and interviews.

To build the optimal initial AI-based intervention models, we first compared multiple ML models using the data from the modeling phase (Section 6.1). After checking the intervention frequency among different intervention methods (Section 6.2), we then evaluated the adaptiveness and explanation aspects of Time2Stop from multiple metrics, including accuracy and receptivity (Section 6.3), app usage duration and visit frequency (Section 6.4), as well as participants' perceived effectiveness of different intervention types (Section 6.5). Overall, our findings showed the consistent advantage of the adaptive component (*Adaptive-w-Exp*/*Adaptive-wo-Exp vs. Personalized*/*Control*). We also observed interesting effects of explanations (*Adaptive-wo-Exp vs. Adaptive-w-Exp*) on app usage behavior and user experience.

6.1 ML Model Comparison

Using the data gathered during the modeling phase, we compared a wide range of off-the-shelf ML models, including Naive Bayes (NB), Logistic Regression (LR), Support Vector Machines (SVM), Decision Trees (DT), Random Forest (RF), and K-Nearest Neighbors (KNN). To account for real-world temporal changes in user behavior and simulate actual model deployment, we used the first week for training and the subsequent week for testing.

The collected data was imbalanced (37.8% overuse, 62.2% non-overuse). Other than calculating individual probabilities for the *Control* intervention type (42.4 ± 24.4%), we experimented with SMOTE-based under-sampling and up-sampling methods for model training [9]. We also tuned hyperparameters on promising models with grid search. Our results indicated that RF (number of estimators: 100, max depth: 10, min samples split: 5), coupled with the up-sampling method, had the best performance across all models, with an F1 score of 66.7%. Other models had worse results: NB (55.3%), LR (59.0%), SVM (59.0%), DT (59.6%), KNN (62.6%). We use this RF model as the static ML model for *Personalized*, as well as the initial model for *Adaptive-wo-Exp* and *Adaptive-w-Exp*.

We also performed a feature importance analysis across all users' models. Our analysis revealed consistency in vital features among participants: the most common important features were related to phone usage (unlock duration), location (total travel distance, moving to static ratio), and temporal feature (*e.g.*, whether night time).

6.2 Intervention Frequency

Prior to the comparison of intervention effectiveness, we first compare the frequency of intervention in our field experiment. Our Friedman test across four intervention types showed that the number of daily interventions was significantly different ($\chi^2 = 16.60$, $p < 0.001$). Our post-hoc pairwise comparison (Wilcoxon signed-rank test with Holm-Bonferroni correction) indicated differences between *Control* and all the rest threes (ps < 0.01), but not for other pairs. This means that the personalized and adaptive models sent fewer interventions to users. As shown in the rest of this section, they were more effective with less intervention frequency.

6.3 Intervention Accuracy and Receptivity

In this section, we investigate the effectiveness of adaption (Section 6.3.1) and explanation (Section 6.3.2) through the perspective of intervention accuracy and receptivity. We also measure the performance dynamics over time (Section 6.3.3). Since individual behaviors varied greatly across participants, we used *Control* as the benchmark and normalized accuracy and receptivity metrics for each participant accordingly. A value higher than 1.0 means better

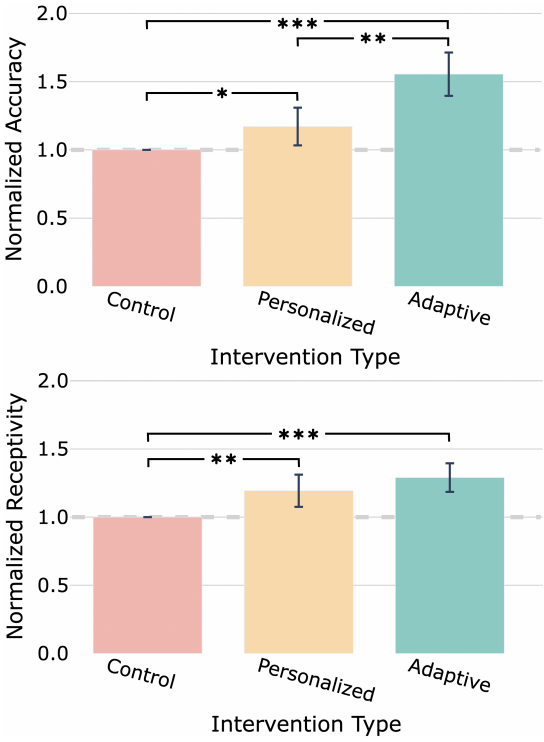

Figure 5: Intervention Accuracy (Top) and Receptivity (Bottom) Comparison across Three Intervention Types. Error bar indicates standard deviation. The same below. The two adaptive versions (with and without explanation) are merged into `Adaptive` to highlight better that adaptive ML-based methods had higher intervention accuracy and receptivity.

performance and a value lower than 1.0 indicates worse performance.

6.3.1 Effectiveness of Adaptation. **Our results indicated that adaptive methods achieved significantly higher intervention accuracy and receptivity.** To evaluate the effectiveness of our intelligent intervention types (static or adaptive ML models), we first merged *Adaptive-wo-Exp* and *Adaptive-w-Exp* into a type called *Adaptive* to highlight the adaptation property better. Figure 5 (Left) shows the comparison across the three types *Control*, *Personalized* (Δ=17.1% over *Control*), and *Adaptive* (Δ=55.5%). We fitted a Generalized Linear Mixed Model (GLMM) on intervention accuracy, with the Gamma family based on a Kolmogorov–Smirnov distribution test[2]. We set intervention type as the main effect and participant ID as the random effect. Our results showed that the intervention type had a significant effect ($\chi^2(2)$=24.52, p<0.001). Post-hoc analysis with Holm-Bonferroni correction further indicated that both the static *Personalized* model (p<0.05) and the *Adaptive* models (p<0.001) had significantly higher intervention accuracy compared to the *Control* baseline. Furthermore, the *Adaptive* model further significantly outperformed the static *Personalized* model (p<0.01, Δ=32.8%). These results not only revealed the effectiveness of the ML component (*Personalized vs. Control*), but also more importantly, indicated the effectiveness of the adaptation part (*Adaptive vs. Personalized*).

While accuracy refers to explicit user subject feedback on interventions, receptivity describes their actual behavior (*i.e.*, continue using the app or quitting it). Hence, receptivity metrics enable us to measure how different interventions affect participants' actual behavior. We ran another GLMM on the intervention receptivity with the same setup as the accuracy test. Similarly, the results also indicate the significance of intervention type on receptivity ($\chi^2(2)$=18.44, p<0.001), as shown in Figure 5 (Right). The post-hoc pairwise results indicated that participants were more receptive when using the *Personalized* (p=0.005, Δ=19.4%) and *Adaptive* (p<0.001, Δ=29.0%) intervention types compared to the *Control*. These observations on the receptivity metric were consistent with those in the accuracy metric.

6.3.2 Effectiveness of Explanations. **Our results suggested that adding explanations significantly enhanced intervention accuracy and receptivity.** To investigate the impact of explanations, we divided the *Adaptive* type back to the original two groups *Adaptive-wo-Exp* and *Adaptive-w-Exp*. The comparison results of the four intervention types are shown in Figure 6 (Left). We ran another GLMM on accuracy, with the four intervention types as the main effect and participant ID as the random effect. The results showed significance of intervention types ($\chi^2(3)$=35.70, p<0.001), and the post-hoc analysis suggested that *Adaptive-w-Exp* (*i.e.*, our complete Time2Stop system) interventions exhibited the highest accuracy by outperforming *Control* (p<0.001, Δ=97.5%), *Personalized* (p<0.01, Δ=66.9%), and even *Adaptive-wo-Exp* (p<0.05, Δ=53.8%). This evidence suggested the effectiveness of explanations: By explaining why they might be overusing smartphones, Time2Stop could help participants better realize and recognize their overuse behavior than the cases without explanations.

Similar to accuracy, our GLMM on receptivity also showed significance ($\chi^2(3)$=25.57, p<0.001). *Adaptive-w-Exp* also achieved the highest receptivity, as shown in Figure 6 (Right), with strong significance over *Control* (p<0.001, Δ=39.6%), as well as marginal significance over *Personalized* (p=0.06, Δ=18.9%) and *Adaptive-wo-Exp* (p=0.07, Δ=11.4%). Combining the results of both intervention accuracy and receptivity, we found that Time2Stop could not only help participants recognize their overuse behavior (higher accuracy), but also help them stop using an app in the moment (higher receptivity). This finding suggests the effectiveness of Time2Stop by delivering interventions when the users were receptive [61, 64].

6.3.3 Effectiveness over Time. **Both adaptive models had increasing intervention performance over time.** *Adaptive-wo-Exp* and *Adaptive-w-Exp* approaches both regularly updated the ML model nightly. We also evaluated their intervention performance as the field study progressed. As intervention receptivity provides a more objective reflection on user behavior, we analyzed receptivity dynamics over time, as presented in Figure 7. The Y-axis represents normalized receptivity, while the X-axis denotes the progress of the intervention phase. Since we used micro-randomized trials, we took four days as an intervention block, constituting a complete cycle of four distinct intervention types. Block 2 coincided with a national

[2]Unless noted otherwise, we repeated the same procedure for the rest of the GLMM models.

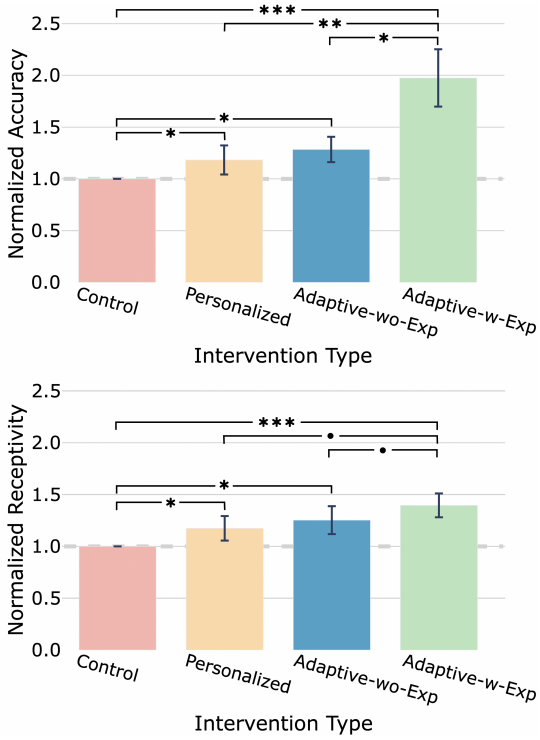

Figure 6: Intervention Accuracy (Top) and Receptivity (Bottom) Comparison across Four Intervention Types. The two versions of `Adaptive` are divided (*Adaptive-w-Exp* and *Adaptive-wo-Exp*) to better highlight that adding explanations can further enhance the performance of interventions.

holiday period, during which participants were on a break and did not attend classes. We observed a significant drop in receptivity from Block 1 to 2 (see the left of Figure 7. Thus, we focused our analysis after Block 3.

We observed an increasing trend in receptivity for the two adaptive intervention types, with *Adaptive-w-Exp* having the most positive slope (r=0.16, Δ=63.6%), followed by *Adaptive-wo-Exp* (r=0.06, Δ=19.1%). These results indicated that our adaptive models could gradually improve over time and that explanations could continuously enhance the intervention's effectiveness. Moreover, *Personalized*'s receptivity was consistently higher than the *Control* baseline across all blocks. However, *Personalized* showed a slight decreasing trend (r=-0.03), while *Control* showed a slight increasing trend (r=0.05). This result may indicate that participants got used to the static ML-based intervention and had less receptivity over time. Our interview data revealed the potential reason behind this interesting finding. We will present more results in Section 6.5.1.

6.4 App Usage Behavior

In addition to intervention accuracy and receptivity, app usage behavior patterns were also important metrics to objectively measure the impact of interventions. We analyzed the app usage logs to investigate the changes in participants' app usage frequency (Section 6.4.1) and duration (Section 6.4.2) between the baseline week and the intervention phase. Similar to Section 6.3, we also normalized our data against the baseline week data to reduce the bias introduced by individual differences.

6.4.1 Change of App Visit Frequency. **The two adaptive methods achieved a significant or marginally significant reduction in visit frequency compared to the base week. However, showing explanations was not as helpful.** Figure 8 (Left) compares the normalized visit frequency of the four intervention types. The average daily visit frequency to monitored apps during the baseline collection period was 94.97 times (SD=52.57). Our results indicated that the visit frequency was reduced for all intervention types: *Control* (93.0%), *Personalized* (92.2%), *Adaptive-w-Exp* (91.7%), and *Adaptive-wo-Exp* (89.8%). We ran a GLMM on the visit frequency, with the intervention type as the main effect and participant ID as the random effect, which showed significance ($\chi^2(4)$=13.85, p<0.01). Post-hoc results with Holm-Bonferroni correction showed that the visit frequency during the days of *Adaptive-wo-Exp* intervention was significantly lower than the baseline week (p<0.01), and that the frequency of *Adaptive-w-Exp* show marginal significance (p=0.07<0.1).

This showed the advantage of the two adaptive methods over the *Personalized* and *Control* methods. However, although the direct comparison between *Adaptive-wo-Exp* and *Adaptive-w-Exp* was not significant, we observed an interesting reversed effect of explanations: In Sections 6.3 and 6.4, explanations could help to improve the intervention accuracy and receptivity; However, when looking into the app visit frequency, adaptive intervention without explanations had better performance.

6.4.2 Change of App Usage Duration. We also observed similar trends for app usage duration. The average time spent on monitored apps during the baseline week was 214.00 minutes (SD=103.57). The usage duration was reduced for all intervention types: *Control* (93.3%), *Personalized* (93.1%), *Adaptive-w-Exp* (91.7%), and *Adaptive-wo-Exp* (89.9%). We still observe the similar advantage of *Adaptive-wo-Exp* over *Adaptive-w-Exp*, but the GLMM on usage duration did not show significance ($\chi^2(4)$=2.62, p=0.62). With *Adaptive-w-Exp*, although participants recognized and stopped more immediate overuse behavior, their overall usage patterns did not change much as *Adaptive-wo-Exp*. We discuss these findings more in Section 6.5 and Section 7.1.

6.5 Subjective Measure

In addition to the intervention accuracy, receptivity, and app usage behavior results, participants' survey responses and comments during exit interviews also provided interesting insights.

6.5.1 Clear Advantage of Adaptive Intervention Methods. Overall, participants had a clear preference for *Adaptive-w-Exp* (*i.e.*, Time2Stop) and *Adaptive-wo-Exp*, followed by *Personalized*, and then *Control*. The left of Figure 9 presents participants' ranking results among the four intervention techniques. *Adaptive-w-Exp* received the most NO.1 ranking (45% of participants), and *Adaptive-wo-Exp* came as the second (43%). This observation was confirmed by a non-parametric Friedman test on ranking numbers that showed strong significance ($\chi^2(3)$=88.01, p<0.001). Our post-hoc pairwise

comparison (Wilcoxon signed-rank test with Holm-Bonferroni correction) indicated significance among all pairs (ps<0.001) except *Adaptive-w-Exp vs. Adaptive-wo-Exp* (p=0.45).

Meanwhile, participants' ratings on the time accuracy, intervention effectiveness, and level of trust were consistent with the ranking results, as shown in the right of Figure 9. We ran three individual Friedman tests on the three metrics. All of them indicated significance (ps<0.001). The post-hoc analysis showed that almost all pair comparisons were significant (for *Adaptive-w-Exp vs. Adaptive-wo-Exp*: $p_{effectiveness}$<0.01, p_{trust}<0.05, all others ps<0.001). The only exception was *Adaptive-w-Exp vs. Adaptive-wo-Exp* on time accuracy (p=0.15).

Our interview data also triangulated these quantitative findings. Many participants felt the difference when comparing *Personalized* and *Control*. "*The random version [Control] didn't make sense, and the timing was strange some days. I think the personalized ML version [Personalized] was consistent with my annotations a few weeks ago.*" (P14) A similar distinction was also observed when comparing the two adaptive versions and *Personalized*. "*I can feel that the adaptive version [Adaptive-wo-Exp] has been learning about my behavior. At the later stage of the study, some days more interventions would pop up if I overused more.*" (P34) "*The version with explanations [Adaptive-w-Exp] is clearly adaptive. The intervention timing became more comfortable after I used it for a while.*" (P55) It is noteworthy that the interface of *Adaptive-wo-Exp*, *Personalized*, and *Control* were the same, and participants only learned the exact dates for intervention methods after the study finished. So, their feeling of differences was mainly based on their experience of the intervention timing.

These findings are in line with the results in the previous section about the advantage of *Personalized* over *Control*, and more importantly, the advantage of *Adaptive-w-Exp* and *Adaptive-wo-Exp* over other two methods.

Moreover, we also noticed that there was a small proportion of users ranking *Control* as the top 1 type (Figure 9 left). Participants commented that this technique was "*surprising/unexpected*". This was in contrast to the *Personalized* method. "*Later in the study, I could somehow expect when it [Personalized] would show up. But that method [Control] is hard to predict. So sometimes it is refreshing.*" (P15) This is also supported by previous work [42], which could explain the increasing trend of *Control*'s receptivity and the decreasing trend of *Personalized* over time in Section 6.3.3 and Figure 7.

6.5.2 Trade-off between with vs. without Explanation. We also had interesting observations that could explain the difference between intervention receptivity (where *Adaptive-w-Exp* had the best performance, as shown in Figure 6 and 7) and app usage behavior (where *Adaptive-wo-Exp* was the best, Figure 8).

Our survey results suggested that *Adaptive-w-Exp* and *Adaptive-wo-Exp* had similar performance. We also found diversity in preference ranking of *Adaptive-w-Exp*: Although *Adaptive-w-Exp* received 45% of the NO.1 voting (compared to a similar 43% for *Adaptive-wo-Exp*), it also received 19% of the NO.3 voting (compared to a much lower 8% for *Adaptive-wo-Exp*). While most participants liked *Adaptive-w-Exp*, a certain proportion of participants found it less preferable.

We dug deep into this difference during our exit interviews. On the one hand, participants who preferred *Adaptive-w-Exp* found explanations could trigger more self-awareness: "*Seeing the explanations could help me to better self-reflect, which often made me stop using my phone.*" (P26) "*Those explanations pushed me to think more about the reason behind my phone usage.*" (P29) "*Explanations helped me to trust the system better.*" (P24) These results indicated that showing explanations could better trigger System 2 (the reasoning and analytical system) with reasoning and self-analysis and improve users' trust in the intervention. These could explain the significantly better effectiveness and level of trust in *Adaptive-w-Exp* (yet the effective sizes were limited $r_{effectiveness}$=0.20, r_{trust}=0.16). On the other hand, participants who did not like *Adaptive-w-Exp* found explanations overly broad and sometimes confusing. "*Sometimes, the explanations felt accurate. But they were very broad so I am not sure.*" (P34) Some participants found explanations unnecessary. "*I was aware of my phone overuse, so I didn't need explanations.*" (P59)

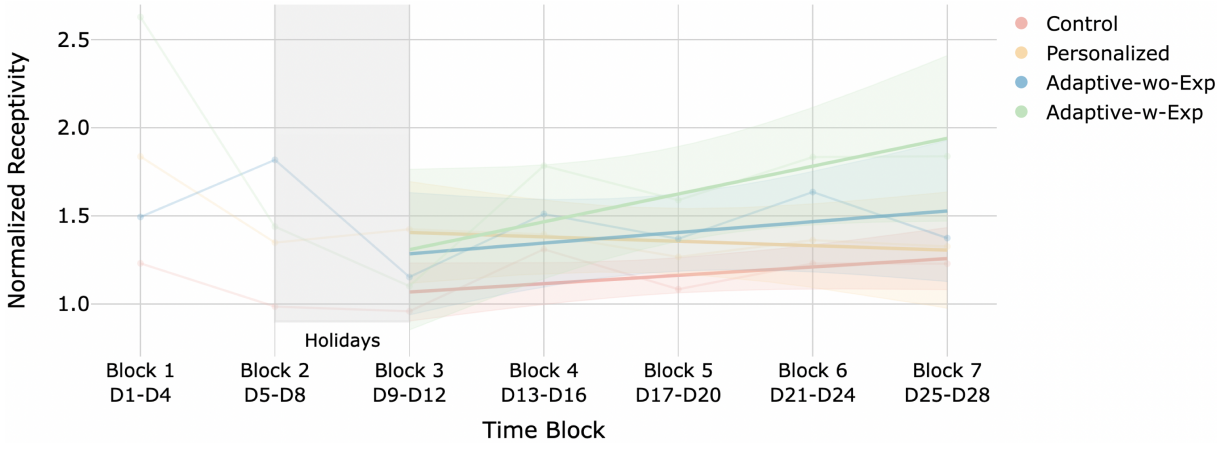

Figure 7: Intervention Performance Over Time. Both adaptive models had an increasing trend, followed by the *Control* group, while *Personalized* method showed a decreasing trend.

These results suggest a more detailed and fine-grained explanation could be helpful for smartphone overuse intervention.

These diverse user reactions toward intervention explanations could explain the mixed results when comparing *Adaptive-w-Exp* and *Adaptive-wo-Exp*. We will have more discussion about this in Section 7.

6.6 Summary of Results

Our 8-week field experiment showed that AI-powered JITAI interventions effectively reduce smartphone overuse. Our two *Adaptive* models provided significantly more accurate interventions compared to *Control* (55.5%) and *Personalized* (32.8%) groups. This trend was consistent for intervention receptivity: participants were significantly more receptive to the two *Adaptive* models compared to *Personalized* (8.0%) and *Control* (29.0%) models. Furthermore, the intervention accuracy and receptivity were further enhanced with explanations. *Adaptive-w-Exp*, *i.e.*, our complete system Time2Stop, could significantly better help users to recognize their overuse (high accuracy) than *Adaptive-wo-Exp* (53.8%), *Personalized* (66.9%), and *Control* (97.5%) methods. Similarly, explanations helped users to be more receptive to interventions and quit using apps. *Adaptive-w-Exp* was more receptive than *Adaptive-wo-Exp* (11.4%), *Personalized* (18.9%), and *Control* (39.6%). We also discovered that the receptivity of adaptive models improved throughout the intervention period, showing the potential of benefiting from long-term deployment with adaptive ML models.

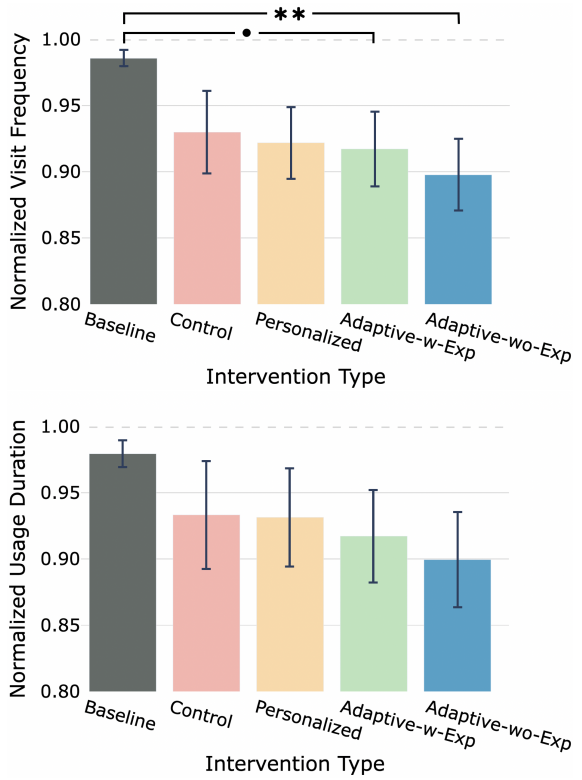

Figure 8: App usage visit frequency (Top) and usage duration (Bottom). The two adaptive methods reduced the most app visit frequency and usage duration. Interestingly, in contrast to Figure 5-7, showing explanations did not augment the performance from the perspective of app usage behavior.

Regarding the actual smartphone usage behavior, all intervention types helped users reduce usage compared to the baseline week. We observe a significant reduction in app visit frequency for *Adaptive-wo-Exp* (8.9%) and a marginally significant reduction for *Adaptive-w-Exp* (7.0%) Analysis of subjective responses also aligned with our quantitative findings. Most participants ranked *Adaptive-w-Exp* and *Adaptive-wo-Exp* as their preferred options. Moreover, time accuracy, effectiveness, and trust were consistent with the results by showing the superiority of the two adaptive models. Interestingly, we observed an unexpected mixed effect of explanations. The intervention accuracy and receptivity results indicated the advantage of explanations, while the app usage behavior suggested the opposite. Our qualitative results revealed that some users appreciated explanations for higher transparency and trustworthiness. On the other hand, other participants found explanations sometimes redundant or overly broad. We discuss this interesting observation in the next section.

7 DISCUSSION

We designed and developed a novel ML-based explainable JITAI system Time2Stop for smartphone overuse intervention. To systematically evaluate the effectiveness of making the system adaptive and explainable, we conducted a micro-randomized study to deploy and measure four different intervention types. Each type added one more component on top of the previous version: ML-based intelligence (*Personalized vs. Control*), adaptivity (*Adaptive-wo-Exp vs. Personalized*), and explainability (*Adaptive-w-Exp vs. Adaptive-wo-Exp*). Our results demonstrate each component can improve the performance of the intervention system to some extent, with an interesting observation of the mixed effect of explanations. Combining these components, Time2Stop provides a trustworthy and effective intervention with accurate timing while adapting to individuals' behaviors. In this section, we discuss the potential reasons behind the explanations' effect (Sec. 7.1), the design considerations and takeaways from our field experiment (Sec. 7.2), the ethical concerns accompanying AI-based JITAI systems (Sec. 7.3), as well as the limitations in our work (Sec. 7.4).

7.1 The Mixed Effect of Explanations

In our field experiment, the advantages of *Personalized* over *Control* and *Adaptive-wo-Exp* over *Personalized* are stable across different metrics. However, the comparison between *Adaptive-wo-Exp* and *Adaptive-w-Exp* shows diverse results. In Figures 6 and 7, the advantage of *Adaptive-w-Exp* is clear, while in Figure 8 we observe the advantage of *Adaptive-w-Exp* instead. These results indicate that during the experiment, participants tended to agree with the intervention timing (higher accuracy) and leave the target apps (higher receptivity) when interventions provided explanations. However, such "successful intervention" did not have a lasting effect. Participants still visited and spent more time in target apps. Although our explanations successfully triggered their System 2 and led to usage pauses, some participants did not effectively internalize the

explanation content and were "pushed" by explainable interventions without deep reflection. Our interview results also support this. Although most participants found explanations helpful for self-reflection, some found explanations confusing and overly broad.

This illustrates the need for more advanced explanation generation techniques in future deployment. Now that we have built adaptive ML models, future explanations should be dynamic, personalized, and adaptive to users. Our interview results reveal that individuals have different preferences in the level of detail. Thus, our system must adapt explanations to fulfill users' specific needs [77]. Meanwhile, recent research suggests a few promising directions, such as explanation selection (to ensure preference alignment) [45] and verifiability (to verify the correctness of AI outputs) [17]. Future work can be explored along with these directions to enhance the effectiveness of explanations further.

7.2 Design Consideration of Intelligent JITAI with Human-in-the-Loop

We made a range of design decisions in our deployment. We reflect on important considerations and share the lessons from our study.

Alternative User-in-the-Loop Labels. In our study, we designed a simple single-click feedback button to collect user feedback and establish the human-AI loop (see Figure 2). We then used such feedback as new labels to tune our ML model. This design has pros and cons. On the one hand, it retrieves users' real-time reactions explicitly so that the model adapts toward users' subjective experience and preference, providing transparency and user agency to some extent. On the other hand, it requires extra effort from users and can miss data when users do not provide feedback. Moreover, this approach does not consider potential bias (compulsively engaged users providing incorrect labels). Another alternative is to leverage users' reactions towards the intervention as implicit feedback labels (*e.g.*, leaving the app could be marked as being receptive to the intervention). This method is also adopted by some previous work in AI-based JITAI systems [49, 71]. It reduces user effort and adapts the model based on their actual behaviors. However, such implicit labels can be affected by noisy behavior, *i.e.*, users could leave/stay in the app for external reasons other than the intervention. Additionally, a hybrid method utilizing both explicit (user-provided) and implicit (behavioral reactions) labels could be implemented. This hybrid model could be refined by measuring the consistency between implicit and explicit labels to assign varying weights to samples, facilitating the enhancement or updating of the model. Moreover, involving health experts in the human-AI loop could provide a valuable solution. Collaborating with experts allows for a more nuanced and balanced definition of overuse, incorporating both user perspectives and health-related guidelines. This collaboration ensures a more objective and informed approach toward setting criteria that align with both user behavior and health considerations, thereby enhancing the accuracy and reliability of the model's updates and interventions. Researchers, designers, and developers must carefully inspect the specific use cases and choose between explicit and implicit feedback or a combination of both.

Real-time *vs.* Reflective Feedback. Other than collecting user feedback *in situ* (*i.e.*, when using target apps), we also explored another method to ask users to label their data at the end of the day. This post-hoc labeling offers users more time to reflect on their behavior. However, recalling earlier smartphone usage cases can be challenging, especially for quick usage such as habitual phone checks. We introduced this method in our experiment for participants who wanted to make up for missing labels. However, our results indicated that they barely used this method (around 3%), thus we did not include them in our analysis. This was mainly because our label collection and feedback design was simple enough. A reflective feedback mechanism could be a promising solution for other behavior intervention studies involving a more complex label collection process.

Prediction Model Update Methods and Frequency. We did not explore more advanced models in Time2Stop, such as deep learning or recent large language models (LLMs) [58, 91], as the model itself is not the main focus of our paper. Time2Stop employs re-training with recency-based weight assignment for model updates. Although this method is robust, other advanced methods, such as reinforcement learning, can be explored in future work.

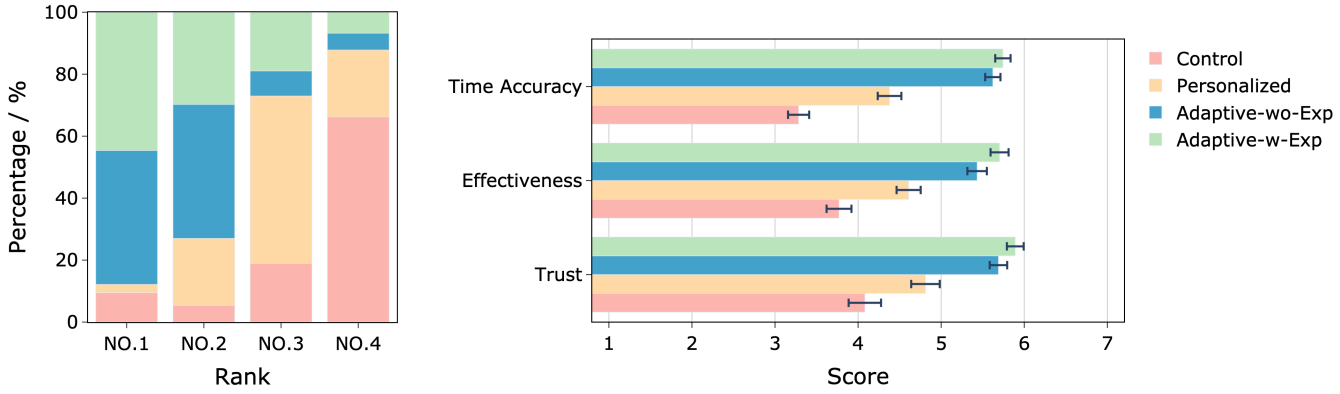

Figure 9: Survey Results Summary. (Left) User Preference Rankings among The Four Intervention Methods. (Right) User Ratings on Intervention Time Accuracy, Perceived Effectiveness, and Level of Trust to Different Methods. The two adaptive methods received the highest user subjective preference and ratings.

Moreover, the update frequency of the prediction model is crucial for the system's adaptability. We updated the model daily to balance our current design's performance and computational costs. But there can be other options. A high frequency of updates (*e.g.*, hourly or even after each interaction) can allow the system to rapidly adapt to users' changing behavior and provide more timely and relevant interventions. However, this comes with higher computational costs and the risk of overfitting to temporary changes in user behavior. Conversely, a lower frequency of updates (*e.g.*, weekly or monthly) reduces the computational load and the risk of overfitting. Still, it may result in the system being slower to adapt to meaningful changes in user behavior. There is a trade-off between adaptability and stability that must be carefully considered. In addition, the trade-off is also impacted by specific applications. Interventions for mental health may require a different frequency than the ones for smartphone overuse. Future work could explore adaptive update frequencies, where the model update frequency is dynamically adjusted based on the stability of user behavior and the model's performance.

Handling "Cold-Start" in JITAI-based Interventions In our study, we devoted the first two weeks to data collection before deploying the intervention. This could be hard to achieve in real-world scenarios. To address this "cold-start" challenge, one promising future approach involves unsupervised learning [4, 18] where users will not be required to provide labels. Instead, the model will grasp smartphone usage patterns by leveraging (*e.g.*, by clustering) users' historical data. Another potential strategy involves few-shot domain adaptation [22, 23, 76, 81, 86], where we can pre-train a model with a dataset (such as from this study) and then fine-tune the model with a small amount of additional data from new users. Additionally, test-time adaptation [20, 21], an advanced domain adaptation technique, could directly utilize test-time data to adapt a global model to a new user without requiring any collected training data.

Dynamic Features for Longitudinal Model Deployment. In our study, we conducted feature selection using the first two weeks of the data and kept the feature set static throughout the experiment. However, for long-term deployment, the importance of different features may change over time. Therefore, dynamic feature selection can be applied. One potential method is selecting each model update's most relevant feature set. This may help the model to have a better performance over time. However, similar to the frequent model updates discussed above, dynamic feature selection will introduce additional complexity and computational requirements. It may also result in a less stable model, which can be challenging for explanation generation and user trust building. The same trade-off between adaptability and stability is also needed for dynamic feature selection.

Explanation Level of Details. As mentioned in the previous section, our current design of high-level intervention explanations could be too general and confusing. Sec. 7.1 discusses the potential of personalized and adaptive explanation generation. However, overly detailed explanations may inadvertently reveal sensitive information about user behavior, which can raise privacy concerns [3, 36]. It is an open research question on providing appropriate detail for model explanation. For behavior change targets that are more objective (*e.g.*, smartphone overuse), providing more detailed explanations can be a good idea. While for more abstract targets (*e.g.*, stress management), high-level and abstract explanations may be more appropriate [36].

7.3 Ethical Concerns and Risk of AI-based Intervention System

Despite the promising performance of our explainable JITAI, we also highlight the important ethical concerns of such an intelligent intervention system. These concerns must be addressed before any real-world, large-scale deployment. First, there is the risk of wrongly predicting smartphone overuse. Our best performance achieved an F1 score of 67%. It may occasionally make incorrect predictions and lead to poorly timed interventions, which are annoying or even harmful to users. For example, a false positive that incorrectly identifies a user as overusing their smartphone when they are using it for an important task may cause unnecessary stress and disrupt their workflow. Similarly, due to the limitations of the explanation method and model performance, the explanation content may not accurately reflect the actual reasons. These wrong explanations can lead to confusion and mistrust of the system and may result in users ignoring or rejecting the interventions. Therefore, it is essential to carefully evaluate the model prediction performance and explanation quality. Other than exploring more advanced ML algorithms, LLMs [8, 47, 84, 88] may offer a new method to generate appropriate and convincing explanation content. Besides, when an intervention is personalized, users, especially younger ones, could be biased towards being more receptive to adopting it [24]. Therefore, a misalignment between subjective measures (users' feedback) and objective measures (their actual behavior) in the study could exist. Our discussion about alternative user-in-the-loop labels in Section 7.2 could be a potential solution. This factor should be carefully considered when deploying interventions with explicit personalized components. Meanwhile, privacy is another critical concern. Our current system adopts a centralized learning method that merges all users' data for model training. Future work can explore edge computing methods such as federated learning [48] to address privacy concerns.

7.4 Limitations

There are a few limitations in our work. First, we mainly focused on young adults. Our study population had a limited age range; thus, the findings of our results may not be generalizable to other population groups. Meanwhile, there is a lack of exploration on the fairness evaluation of our methods. Second, our micro-randomized study design took the daily level as the randomization unit. This enabled us to conduct a within-subject design within a feasible period. However, we could not investigate the lasting effect of different intervention methods as they were mixed. Meanwhile, our observational study may neglect unknown confounding variables beyond this paper's scope. For instance, while we employed a time-based train-test split to train the base model, we acknowledge that mitigating the 'observational effect' (the impact of monitoring and labeling) might pose a challenge. Besides, although we revealed the exact dates of different interventions during exit interviews and surveys, participants' responses might still be inaccurate or biased by their memory. Third, we updated intelligent adaptive models at

midnight, which might not align with college students' sleeping schedules. Last, our work only considered smartphone overuse as a general intervention target. The specific types of overuse, such as excessive use of social media or video gaming, were not investigated in detail in this study. Similarly, as mentioned earlier, we didn't experiment with more intervention methods other than digit typing, as the specific intervention method is not the focus of our work.

8 CONCLUSION

In this paper, we propose a novel AI-powered explainable JITAI system, Time2Stop, for smartphone overuse intervention. Our system captures user context and behavior, leverages AI to infer smartphone overuse scenarios, introduces interventions when overuse is detected, provides explanations, and updates the intervention model iteratively based on human-in-the-loop feedback to form a human-AI loop. In order to measure the effectiveness of Time2Stop, we conducted an 8-week field experiment (N=71) and compared four intervention types. Our results not only showed the advantage of the ML component (the static ML-powered version over the basic control), but more importantly, underscored the advantages of adaptive intervention types compared to the static version, with significantly better intervention accuracy (32.8%) and receptivity (8.0%). Furthermore, including explanations in our system significantly amplified its accuracy (53.8%) and receptivity (11.4%). In addition, users exhibited reduced visit frequency to apps they considered unproductive when engaged with adaptive models (7.0-8.9%). Findings from our qualitative analysis echoed the quantitative results, with users expressing a clear preference for adaptive interventions. We also observed an interesting mixed effect of explanations, which could shed light on future research direction. We further highlighted the important ethical concerns of AI-based intervention systems for real-world deployment. We envision our work can be applied beyond the field of smartphone overuse and inspire future practitioners to explore more advanced intervention techniques with a human-AI loop.

ACKNOWLEDGMENTS

We express our gratitude to all the participants who contributed to our longitudinal user study. We extend our appreciation to Jennifer Mankoff for her insights and discussions that significantly enriched this project, and to Zihan Yan for her assistance in the pilot study. This paper is based upon work supported by the VW Foundation, Quanta Computing, the Natural Science Foundation of China (NSFC) under Grant Number 62132010, Young Elite Scientists Sponsorship Program by CAST under Grant Number 2021QNRC001, Tsinghua University Initiative Scientific Research Program and Institute of Information & communications Technology Planning & Evaluation (IITP) under Grant Number 2022-0-00495.

REFERENCES

[1] Heejune Ahn, Muhammad Eka Wijaya, and Bianca Camille Esmero. 2014. A systemic smartphone usage pattern analysis: focusing on smartphone addiction issue. *Int J Multimed Ubiquitous Eng* 9, 6 (2014), 9–14.

[2] StayFree Apps. 2017. StayFree. https://play.google.com/store/apps/details?id=com.burockgames.timeclocker.

[3] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* 58 (June 2020), 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

[4] Yoshua Bengio. 2012. Deep learning of representations for unsupervised and transfer learning. In *Proceedings of ICML workshop on unsupervised and transfer learning*. JMLR Workshop and Conference Proceedings, 17–36.

[5] Jennifer L Bevan, Jeanette Pfyl, and Brett Barclay. 2012. Negative emotional and cognitive responses to being unfriended on Facebook: An exploratory study. *Computers in Human Behavior* 28, 4 (2012), 1458–1464. https://doi.org/10.1016/j.chb.2012.03.008

[6] Umang Bhatt, Alice Xiang, Shubham Sharma, Adrian Weller, Ankur Taly, Yunhan Jia, Joydeep Ghosh, Ruchir Puri, José MF Moura, and Peter Eckersley. 2020. Explainable machine learning in deployment. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 648–657.

[7] Virginia Braun and Victoria Clarke. 2012. *Thematic analysis.* American Psychological Association.

[8] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of Artificial General Intelligence: Early experiments with GPT-4. http://arxiv.org/abs/2303.12712

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

[10] Hyunsung Cho, DaEun Choi, Donghwi Kim, Wan Ju Kang, Eun Kyoung Choe, and Sung-Ju Lee. 2021. Reflect, not Regret: Understanding Regretful Smartphone Use with App Feature-Level Analysis. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (Oct. 2021), 1–36. https://doi.org/10.1145/3479600

[11] S Demirci, K Demirci, and M Akgonul. 2016. Headache in smartphone users: a cross-sectional study. *J Neurol Psychol* 4, 1 (2016), 5.

[12] Android Developers. 2021. Android Accessibility Service. https://developer.android.com/guide/topics/ui/accessibility.

[13] Allen L Edwards. 1951. Balanced latin-square designs in psychological research. *The American journal of psychology* 64, 4 (1951), 598–603.

[14] Malin Eiband, Hanna Schneider, Mark Bilandzic, Julian Fazekas-Con, Mareike Haug, and Heinrich Hussmann. 2018. Bringing Transparency Design into Practice. In *23rd International Conference on Intelligent User Interfaces*. ACM, Tokyo Japan, 211–223. https://doi.org/10.1145/3172944.3172961

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

[16] Linda Fischer-Grote, Oswald D Kothgassner, and Anna Felnhofer. 2019. Risk factors for problematic smartphone use in children and adolescents: a review of existing literature. *neuropsychiatrie* 33, 4 (2019), 179.

[17] Raymond Fok and Daniel S Weld. 2023. In Search of Verifiability: Explanations Rarely Enable Complementary Performance in AI-Advised Decision Making. *arXiv preprint arXiv:2305.07722* (2023).

[18] Zoubin Ghahramani. 2003. Unsupervised learning. In *Summer school on machine learning*. Springer, 72–112.

[19] Fausto Giunchiglia, Mattia Zeni, Elisa Gobbi, Enrico Bignotti, and Ivano Bison. 2018. Mobile social media usage and academic performance. *Computers in Human Behavior* 82 (2018), 177–185. https://doi.org/10.1016/j.chb.2017.12.041

[20] Taesik Gong, Jongheon Jeong, Taewon Kim, Yewon Kim, Jinwoo Shin, and Sung-Ju Lee. 2022. NOTE: Robust continual test-time adaptation against temporal correlation. *Advances in Neural Information Processing Systems* 35 (2022), 27253–27266.

[21] Taesik Gong, Yewon Kim, Taeckyung Lee, Sorn Chottananurak, and Sung-Ju Lee. 2023. SoTTA: Robust Test-Time Adaptation on Noisy Data Streams. In *Thirty-seventh Conference on Neural Information Processing Systems*. https://openreview.net/forum?id=3bdXag2rUd

[22] Taesik Gong, Yewon Kim, Adiba Orzikulova, Yunxin Liu, Sung Ju Hwang, Jinwoo Shin, and Sung-Ju Lee. 2023. DAPPER: Label-Free Performance Estimation after Personalization for Heterogeneous Mobile Sensing. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 7, 2, Article 55 (jun 2023), 27 pages. https://doi.org/10.1145/3596256

[23] Taesik Gong, Yeonsu Kim, Jinwoo Shin, and Sung Ju Lee. 2019. MetaSense: Few-shot adaptation to untrained conditions in deep mobile sensing. *SenSys 2019 - Proceedings of the 17th Conference on Embedded Networked Sensor Systems* (2019), 110–123. https://doi.org/10.1145/3356250.3360020

[24] Xitong Guo, Xiaofei Zhang, and Yongqiang Sun. 2016. The privacy–personalization paradox in mHealth services acceptance of different age groups. *Electronic Commerce Research and Applications* 16 (2016), 55–65.

[25] David H Gustafson, Fiona M McTavish, Ming-Yuan Chih, Amy K Atwood, Roberta A Johnson, Michael G Boyle, Michael S Levy, Hilary Driscoll, Steven M Chisholm, Lisa Dillenburg, et al. 2014. A smartphone application to support recovery from alcoholism: a randomized clinical trial. *JAMA psychiatry* 71, 5 (2014), 566–572.

[26] Andree Hartanto and Hwajin Yang. 2016. Is the smartphone a smart choice? The effect of smartphone separation on executive functions. *Computers in human behavior* 64 (2016), 329–336.

[27] Alexis Hiniker, Sungsoo (Ray) Hong, Tadayoshi Kohno, and Julie A. Kientz. 2016. MyTime: Designing and Evaluating an Intervention for Smartphone Non-Use. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 4746–4757. https://doi.org/10.1145/2858036.2858403

[28] Joyce Ho and Stephen S Intille. 2005. Using context-aware computing to reduce the perceived burden of interruptions from mobile devices. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 909–918.

[29] Robert R Hoffman, Shane T Mueller, Gary Klein, and Jordan Litman. 2018. Metrics for explainable AI: Challenges and prospects. *arXiv preprint arXiv:1812.04608* (2018).

[30] Wilhelm Hofmann, Malte Friese, and Fritz Strack. 2009. Impulse and Self-Control From a Dual-Systems Perspective. *Perspectives on Psychological Science* 4, 2 (March 2009), 162–176. https://doi.org/10.1111/j.1745-6924.2009.01116.x

[31] Paul Hur, HaeJin Lee, Suma Bhat, and Nigel Bosch. 2022. Using Machine Learning Explainability Methods to Personalize Interventions for Students. *International Educational Data Mining Society* (2022).

[32] Shamsi T Iqbal and Brian P Bailey. 2010. Oasis: A framework for linking notification delivery to the perceptual structure of goal-directed tasks. *ACM Transactions on Computer-Human Interaction (TOCHI)* 17, 4 (2010), 1–28.

[33] Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. 2020. Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–14. https://doi.org/10.1145/3313831.3376219

[34] Jaejeung Kim, Chiwoo Cho, and Uichin Lee. 2017. Technology supported behavior restriction for mitigating self-interruptions in multi-device environments. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 3 (2017), 1–21.

[35] Jaejeung Kim, Joonyoung Park, Hyunsoo Lee, Minsam Ko, and Uichin Lee. 2019. LocknType: Lockout Task Intervention for Discouraging Smartphone App Use. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300927

[36] Taewan Kim, Haesoo Kim, Ha Yeon Lee, Hwarang Goh, Shakhboz Abdigapporov, Mingon Jeong, Hyunsung Cho, Kyungsik Han, Youngtae Noh, Sung-Ju Lee, and Hwajung Hong. 2022. Prediction for Retrospection: Integrating Algorithmic Stress Prediction into Personal Informatics Systems for College Students' Mental Health. In *CHI Conference on Human Factors in Computing Systems*. ACM, New Orleans LA USA, 1–20. https://doi.org/10.1145/3491102.3517701

[37] Predrag Klasnja, Eric B Hekler, Saul Shiffman, Audrey Boruvka, Daniel Almirall, Ambuj Tewari, and Susan A Murphy. 2015. Microrandomized trials: An experimental design for developing just-in-time adaptive interventions. *Health Psychology* 34, S (2015), 1220.

[38] Minsam Ko, Seungwoo Choi, Koji Yatani, and Uichin Lee. 2016. Lock n' LoL: Group-Based Limiting Assistance App to Mitigate Smartphone Distractions in Group Activities. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 998–1010. https://doi.org/10.1145/2858036.2858568

[39] Minsam Ko, Subin Yang, Joonwon Lee, Christian Heizmann, Jinyoung Jeong, Uichin Lee, Daehee Shin, Koji Yatani, Junehwa Song, and Kyong-Mee Chung. 2015. NUGU: A Group-Based Intervention App for Improving Self-Regulation of Limiting Smartphone Use. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing* (Vancouver, BC, Canada) *(CSCW '15)*. Association for Computing Machinery, New York, NY, USA, 1235–1245. https://doi.org/10.1145/2675133.2675244

[40] Minsam Ko, Subin Yang, Joonwon Lee, Christian Heizmann, Jinyoung Jeong, Uichin Lee, Daehee Shin, Koji Yatani, Junehwa Song, and Kyong-Mee Chung. 2015. NUGU: a group-based intervention app for improving self-regulation of limiting smartphone use. In *Proceedings of the 18th ACM conference on computer supported cooperative work & social computing*. 1235–1245.

[41] Geza Kovacs. 2019. *HabitLab: In-the-wild Behavior Change Experiments at Scale*. Stanford University.

[42] Geza Kovacs, Zhengxuan Wu, and Michael S Bernstein. 2018. Rotating online behavior change interventions increases effectiveness but also increases attrition. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–25.

[43] Florian Künzler, Varun Mishra, Jan-Niklas Kramer, David Kotz, Elgar Fleisch, and Tobias Kowatsch. 2019. Exploring the state-of-receptivity for mhealth interventions. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 4 (2019), 1–27.

[44] Min Kwon, Joon-Yeop Lee, Wang-Youn Won, Jae-Woo Park, Jung-Ah Min, Changtae Hahn, Xinyu Gu, Ji-Hye Choi, and Dai-Jin Kim. 2013. Development and validation of a smartphone addiction scale (SAS). *PloS one* 8, 2 (2013), e56936.

[45] Vivian Lai, Yiming Zhang, Chacha Chen, Q Vera Liao, and Chenhao Tan. 2023. Selective explanations: Leveraging human input to align explainable ai. *arXiv preprint arXiv:2301.09656* (2023).

[46] Liette Lapointe, Camille Boudreau-Pinsonneault, and Isaac Vaghefi. 2013. Is Smartphone Usage Truly Smart? A Qualitative Investigation of IT Addictive Behaviors. In *2013 46th Hawaii International Conference on System Sciences*. 1063–1072. https://doi.org/10.1109/HICSS.2013.367

[47] Shiyang Li, Jianshu Chen, Yelong Shen, Zhiyu Chen, Xinlu Zhang, Zekun Li, Hong Wang, Jing Qian, Baolin Peng, Yi Mao, et al. 2022. Explanations from large language models make small reasoners better. *arXiv preprint arXiv:2210.06726* (2022).

[48] Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. 2020. Federated Learning: Challenges, Methods, and Future Directions. *IEEE Signal Processing Magazine* 37, 3 (May 2020), 50–60. https://doi.org/10.1109/MSP.2020.2975749

[49] Peng Liao, Kristjan Greenewald, Predrag Klasnja, and Susan Murphy. 2020. Personalized HeartSteps: A Reinforcement Learning Algorithm for Optimizing Physical Activity. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 1 (March 2020), 1–22. https://doi.org/10.1145/3381007

[50] Mengchen Liu, Jiaxin Shi, Kelei Cao, Jun Zhu, and Shixia Liu. 2017. Analyzing the training processes of deep generative models. *IEEE transactions on visualization and computer graphics* 24, 1 (2017), 77–87.

[51] Markus Löchtefeld, Matthias Böhmer, and Lyubomir Ganev. 2013. AppDetox: Helping Users with Mobile App Addiction. In *Proceedings of the 12th International Conference on Mobile and Ubiquitous Multimedia* (Luleå, Sweden) *(MUM '13)*. Association for Computing Machinery, New York, NY, USA, Article 43, 2 pages. https://doi.org/10.1145/2541831.2541870

[52] Tao Lu, Hongxiao Zheng, Tianying Zhang, Xuhai Xu, and Anhong Guo. 2024. InteractOut: Leveraging Interaction Proxies as Input Manipulation Strategies for Reducing Smartphone Overuse. In *Proceedings of the 2024 CHI conference on human factors in computing systems*. Association for Computing Machinery, New York, NY, USA, 1–18.

[53] Kai Lukoff, Ulrik Lyngs, Karina Shirokova, Raveena Rao, Larry Tian, Himanshu Zade, Sean A. Munson, and Alexis Hiniker. 2023. SwitchTube: A Proof-of-Concept System Introducing "Adaptable Commitment Interfaces" as a Tool for Digital Wellbeing. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 197, 22 pages. https://doi.org/10.1145/3544548.3580703

[54] Kai Lukoff, Ulrik Lyngs, Himanshu Zade, J. Vera Liao, James Choi, Kaiyue Fan, Sean A. Munson, and Alexis Hiniker. 2021. How the Design of YouTube Influences User Sense of Agency. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 368, 17 pages. https://doi.org/10.1145/3411764.3445467

[55] Scott M. Lundberg and Su In Lee. 2017. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems* 2017-Decem, Section 2 (2017), 4766–4775.

[56] Ulrik Lyngs, Kai Lukoff, Laura Csuka, Petr Slovák, Max Van Kleek, and Nigel Shadbolt. 2022. The Goldilocks level of support: Using user reviews, ratings, and installation numbers to investigate digital self-control tools. *International journal of human-computer studies* 166 (2022), 102869.

[57] Ulrik Lyngs, Kai Lukoff, Petr Slovak, Reuben Binns, Adam Slack, Michael Inzlicht, Max Van Kleek, and Nigel Shadbolt. 2019. Self-control in cyberspace: Applying dual systems theory to a review of digital self-control tools. In *proceedings of the 2019 CHI conference on human factors in computing systems*. 1–18.

[58] Amama Mahmood, Junxiang Wang, Bingsheng Yao, Dakuo Wang, and Chien-Ming Huang. 2023. LLM-Powered Conversational Voice Assistants: Interaction Patterns, Opportunities, Challenges, and Design Guidelines. *arXiv preprint arXiv:2309.13879* (2023).

[59] Lakmal Meegahapola, William Droz, Peter Kun, Amalia De Götzen, Chaitanya Nutakki, Shyam Diwakar, Salvador Ruiz Correa, Donglei Song, Hao Xu, Miriam Bidoglia, et al. 2023. Generalization and Personalization of Mobile Sensing-Based Mood Inference Models: An Analysis of College Students in Eight Countries. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 4 (2023), 1–32.

[60] Varun Mishra, Florian Künzler, Jan-Niklas Kramer, Elgar Fleisch, Tobias Kowatsch, and David Kotz. 2021. Detecting receptivity for mHealth interventions in the natural environment. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies* 5, 2 (2021), 1–24.

[61] Varun Mishra, Florian Künzler, Jan-Niklas Kramer, Elgar Fleisch, Tobias Kowatsch, and David Kotz. 2021. Detecting Receptivity for mHealth Interventions in the Natural Environment. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 2 (June 2021), 1–24. https://doi.org/10.1145/3463492

[62] Alberto Monge Roffarello and Luigi De Russis. 2019. The Race Towards Digital Wellbeing: Issues and Opportunities. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3290605.3300616

[63] Inbal Nahum-Shani, Mashfiqui Rabbi, Jamie Yap, Meredith L. Philyaw-Kotov, Predrag Klasnja, Erin E. Bonar, Rebecca M. Cunningham, Susan A. Murphy, and Maureen A. Walton. 2021. Translating Strategies for Promoting Engagement in Mobile Health: A Proof-of-Concept Micro-Randomized Trial. *Health psychology : official journal of the Division of Health Psychology, American Psychological Association* 40, 12 (Dec. 2021), 974–987. https://doi.org/10.1037/hea0001101

[64] Inbal Nahum-Shani, Shawna N Smith, Bonnie J Spring, Linda M Collins, Katie Witkiewitz, Ambuj Tewari, and Susan A Murphy. 2018. Just-in-Time Adaptive Interventions (JITAIs) in Mobile Health: Key Components and Design Principles for Ongoing Health Behavior Support. *Annals of Behavioral Medicine* 52, 6 (May 2018), 446–462. https://doi.org/10.1007/s12160-016-9830-8

[65] Inbal Nahum-Shani, Shawna N Smith, Bonnie J Spring, Linda M Collins, Katie Witkiewitz, Ambuj Tewari, and Susan A Murphy. 2018. Just-in-time adaptive interventions (JITAIs) in mobile health: key components and design principles for ongoing health behavior support. *Annals of Behavioral Medicine* 52, 6 (2018), 446–462.

[66] Urbandroid (Petr Nálevka). 2015. Digital Detox: Focus & Live. https://play.google.com/store/apps/details?id=com.urbandroid.ddc.

[67] Adiba Orzikulova, Hyunsung Cho, Hye-Young Chung, Hwajung Hong, Uichin Lee, and Sung-Ju Lee. 2023. FinerMe: Examining App-level and Feature-level Interventions to Regulate Mobile Social Media Use. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW2, Article 274 (oct 2023). https://doi.org/10.1145/3610065

[68] Subramani Parasuraman, Aaseer Thamby Sam, Stephanie Wong Kah Yee, Bobby Lau Chik Chuon, and Lee Yu Ren. 2017. Smartphone usage and increased risk of mobile phone addiction: A concurrent study. *International journal of pharmaceutical investigation* 7, 3 (2017), 125.

[69] Chunjong Park, Junsung Lim, Juho Kim, Sung-Ju Lee, and Dongman Lee. 2017. Don't bother me. I'm socializing! A breakpoint-based smartphone notification system. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. 541–554.

[70] Shaokan Pi. 2015. Forest. https://www.forestapp.cc/.

[71] Mashfiqui Rabbi, Min Hane Aung, Mi Zhang, and Tanzeem Choudhury. 2015. MyBehavior: Automatic Personalized Health Feedback from User Behaviors and Preferences using Smartphones. *Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing* September (2015), 707–718. https://doi.org/10.1145/2750858.2805840 ISBN: 9781450317702.

[72] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why should i trust you?" Explaining the predictions of any classifier. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 13-17-Augu (2016), 1135–1144. https://doi.org/10.1145/2939672.2939778

[73] William Riley, Jami Obermayer, and Jersino Jean-Mary. 2008. Internet and mobile phone text messaging intervention for college smokers. *Journal of American College Health* 57, 2 (2008), 245–248.

[74] Matthew Saponaro, Ajith Vemuri, Greg Dominick, and Keith Decker. 2021. Contextualization and individualization for just-in-time adaptive interventions to reduce sedentary behavior. In *Proceedings of the conference on health, inference, and learning*. 246–256.

[75] Saul Shiffman, Arthur A Stone, and Michael R Hufford. 2008. Ecological momentary assessment. *Annu. Rev. Clin. Psychol.* 4 (2008), 1–32.

[76] Jake Snell, Kevin Swersky, and Richard Zemel. 2017. Prototypical networks for few-shot learning. *Advances in neural information processing systems* 30 (2017).

[77] Robert Thomson and Jordan Richard Schoenherr. 2020. Knowledge-to-information translation training (kitt): An adaptive approach to explainable artificial intelligence. In *Adaptive Instructional Systems: Second International Conference, AIS 2020, Held as Part of the 22nd HCI International Conference, HCII 2020, Copenhagen, Denmark, July 19–24, 2020, Proceedings 22*. Springer, 187–204.

[78] Screen Time. 2020. Screen Time. https://support.apple.com/en-us/HT208982.

[79] Ofir Turel, Alexander Serenko, and Nick Bontis. 2008. Blackberry addiction: Symptoms and outcomes. *AMCIS 2008 Proceedings* (2008), 73.

[80] Helena Vasconcelos, Matthew Jörke, Madeleine Grunde-McLaughlin, Tobias Gerstenberg, Michael S Bernstein, and Ranjay Krishna. 2023. Explanations can reduce overreliance on ai systems during decision-making. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (2023), 1–38.

[81] Yaqing Wang, Quanming Yao, James T Kwok, and Lionel M Ni. 2020. Generalizing from a few examples: A survey on few-shot learning. *ACM computing surveys (csur)* 53, 3 (2020), 1–34.

[82] Digital Wellbeing. 2018. Digital Wellbeing. https://www.android.com/digital-wellbeing/.

[83] Paweł W. Woźniak, Przemysław Piotr Kucharski, Maartje M.A. de Graaf, and Jasmin Niess. 2020. Exploring Understandable Algorithms to Suggest Fitness Tracker Goals That Foster Commitment. In *Proceedings of the 11th Nordic Conference on Human-Computer Interaction: Shaping Experiences, Shaping Society* (Tallinn, Estonia) *(NordiCHI '20)*. Association for Computing Machinery, New York, NY, USA, Article 35, 12 pages. https://doi.org/10.1145/3419249.3420131

[84] Ruolan Wu, Chun Yu, Xiaole Pan, Yujia Liu, Ningning Zhang, Yue Fu, Wang Yuhan, Zheng Zhi, Chen Li, Jiang Qiaolei, Xuhai Xu, and Yuanchun Shi. 2024. MindShift: Leveraging Large Language Models for Mental-States-Based Problematic Smartphone Use Intervention. In *Proceedings of the 2024 CHI conference on human factors in computing systems*. Association for Computing Machinery, New York, NY, USA, 1–18. https://doi.org/10.1145/3613904.3642790

[85] Xuhai Xu, Prerna Chikersal, Afsaneh Doryab, Daniella K. Villalba, Janine M. Dutcher, Michael J. Tumminia, Tim Althoff, Sheldon Cohen, Kasey G. Creswell, J. David Creswell, Jennifer Mankoff, and Anind K. Dey. 2019. Leveraging Routine Behavior and Contextually-Filtered Features for Depression Detection among College Students. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 3 (Sept. 2019), 1–33. https://doi.org/10.1145/3351274

[86] Xuhai Xu, Jun Gong, Carolina Brum, Lilian Liang, Bongsoo Suh, Shivam Kumar Gupta, Yash Agarwal, Laurence Lindsey, Runchang Kang, Behrooz Shahsavari, Tu Nguyen, Heriberto Nieto, Scott E Hudson, Charlie Maalouf, Jax Seyed Mousavi, and Gierad Laput. 2022. Enabling Hand Gesture Customization on Wrist-Worn Devices. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*. ACM, New Orleans LA USA, 1–19. https://doi.org/10.1145/3491102.3501904

[87] Xuhai Xu, Xin Liu, Han Zhang, Weichen Wang, Subgiya Nepal, Kevin S Kuehn, Jeremy Huckins, Margaret E Morris, Paula S Nurius, Eve A Riskin, Shwetak Patel, Tim Althoff, Andrew Campell, Anind K Dey, and Jennifer Mankoff. 2023. GLOBEM: Cross-Dataset Generalization of Longitudinal Human Behavior Modeling. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 4 (2023), 32.

[88] Xuhai Xu, Bingshen Yao, Yuanzhe Dong, Hong Yu, James Hendler, Anind K Dey, and Dakuo Wang. 2023. Leveraging large language models for mental health prediction via online text data. *arXiv preprint arXiv:2307.14385* (2023).

[89] Xuhai Xu, Han Zhang, Yasaman Sefidgar, Yiyi Ren, Xin Liu, Woosuk Seo, Jennifer Brown, Kevin Kuehn, Mike Merrill, Paula Nurius, Shwetak Patel, Tim Althoff, Margaret E Morris, Eve Riskin, Jennifer Mankoff, and Anind K Dey. 2022. GLOBEM Dataset: Multi-Year Datasets for Longitudinal Human Behavior Modeling Generalization. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*. 18.

[90] Xuhai Xu, Tianyuan Zou, Han Xiao, Yanzhang Li, Ruolin Wang, Tianyi Yuan, Yuntao Wang, Yuanchun Shi, Jennifer Mankoff, and Anind K Dey. 2022. TypeOut: Leveraging Just-in-Time Self-Affirmation for Smartphone Overuse Reduction. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*. ACM, New Orleans LA USA, 1–17. https://doi.org/10.1145/3491102.3517476

[91] Ziqi Yang, Xuhai Xu, Bingsheng Yao, Shao Zhang, Ethan Rogers, Stephen Intille, Nawar Shara, Dakuo Wang, et al. 2023. Talk2Care: Facilitating Asynchronous Patient-Provider Communication with Large-Language-Model. *arXiv preprint arXiv:2309.09357* (2023).

[92] Sedat Yasin, Erman Altunisik, and Ali Zeynal Abidin Tak. 2022. Digital Danger in Our Pockets: Effect of Smartphone Overuse on Mental Fatigue and Cognitive Flexibility. *The Journal of Nervous and Mental Disease* (2022), 10–1097.

[93] Zhan Zhang, Yegin Genc, Dakuo Wang, Mehmet Eren Ahsen, and Xiangmin Fan. 2021. Effect of ai explanations on human perceptions of patient-facing ai-powered healthcare systems. *Journal of Medical Systems* 45, 6 (2021), 64.

[94] Linbo Zhuang, Lisheng Wang, Dongming Xu, Zhiyong Wang, and Renzheng Liang. 2021. Association between excessive smartphone use and cervical disc degeneration in young patients suffering from chronic neck pain. *Journal of Orthopaedic Science* 26, 1 (2021), 110–115.

A EXPLANATION EXAMPLES

Table 2: Examples of Feature Explanations at Different Explanation Levels.

Model Feature	Readable Name	Explanation	
		High-level	**Low-level**
numViewScrolledCurrentAppCategory	Number of Scrolls in Current App Category	Phone & App Use	Number of Interactions
sumDurationDischarge	Battery Discharge Duration	Phone & App Use	Battery Usage
durationMobile	Duration of Being Mobile	Activity	Duration of Being Mobile
avgLux	Average Lux in Light Conditions	Activity	Light Conditions
countScansMostFrequentDevice	Number of Frequently Scanned Devices	Social	Number of Nearby Devices
timeFirstSent	Time of First Sent Message	Social	Time of Sent Message
timeAtTopOneLocation	Time Spent at Top One Location	Location	Time at Frequent Locations
minLengthStayAtClusters	Minimum Stay at Frequent Locations	Location	Time at Frequent Locations
isNight	Whether it is the Night Time	Time	the Night Time