Time2Stop

Adaptive and Explainable Human-AI Loop for Smartphone Overuse Intervention

CHI 24

Jian Zheng
as an appreciative presenter with some *comments & questions*

Why I chose this paper?

06/18/2025

Contacting Author

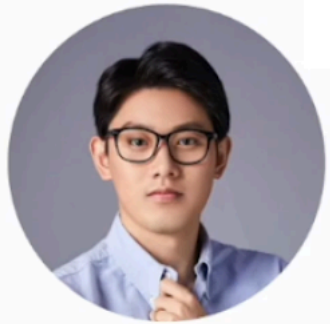
Adiba Orzikulova

adiorz@kaist.ac.kr

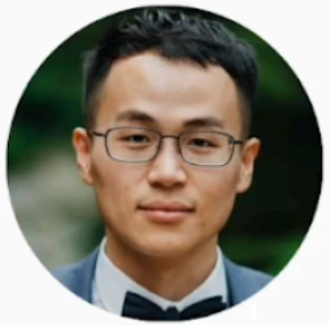
KAIST

Han Xiao
BUPT

Zhipeng Li
Tsinghua University



Yukang Yan
Carnegie Mellon University



Yuntao Wang
Tsinghua University

Yuanchun Shi
Tsinghua University

Marzyeh Ghassemi
MIT

Sung-Ju Lee
KAIST

Anind K Dey
University of Washington

Xuhai “Orson” Xu
MIT

Contents

Smartphone Overuse

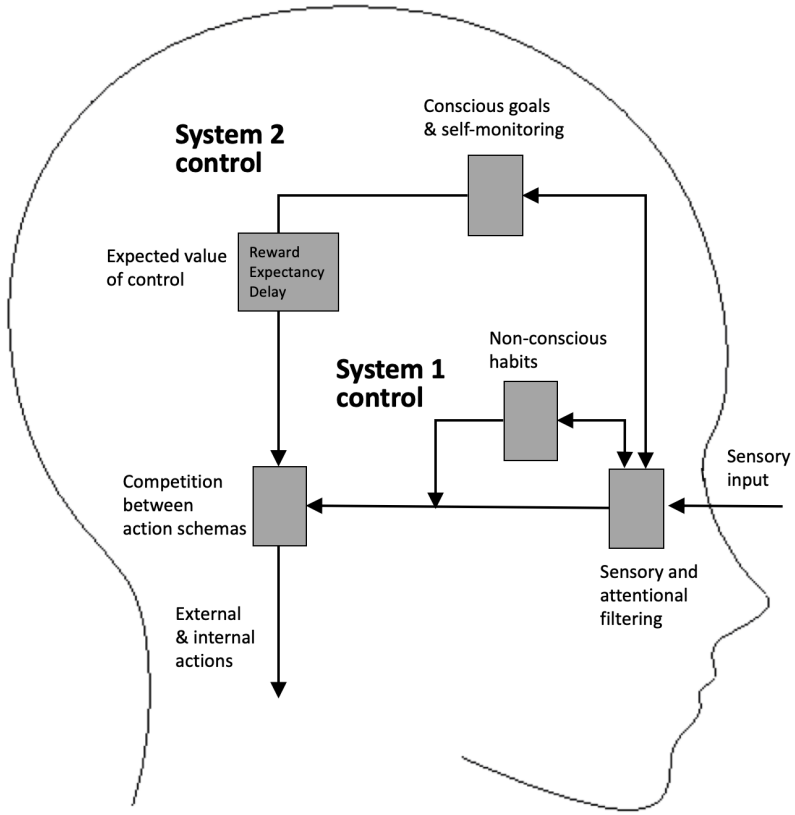
- *Problematic Smartphone Usage*
- Consequences
  - Physical health: headaches, chronic neck pain, sleep disturbance
  - Mental well-being: anxiety, depression, impaired cognitive abilities
  - Social wellness: distraction, family conflicts, performance degradation
- Existing digital intervention tools
  - Often rely on **simple criteria** like pre-determined intervals or app-specific triggers

Just-In-Time Adaptive Intervention (JITAI)

- To deliver tailored and timely support
- Dynamically adapts to users' internal and external contexts
- When the user is both **vulnerable** (susceptible to overuse) and **receptive** (able to process the intervention)
- Can be **rule-based** (predefined rules by experts) or **AI-based** (leverages ML to analyze data and identify patterns)

Explainable AI (XAI)

- Addresses the challenges of interpretability and transparency
- Helps users comprehend AI decisions, fostering trust
- Can activate System 2 thinking (reasoning and analytical system)



Dual process theory

Identifying the Research Gap

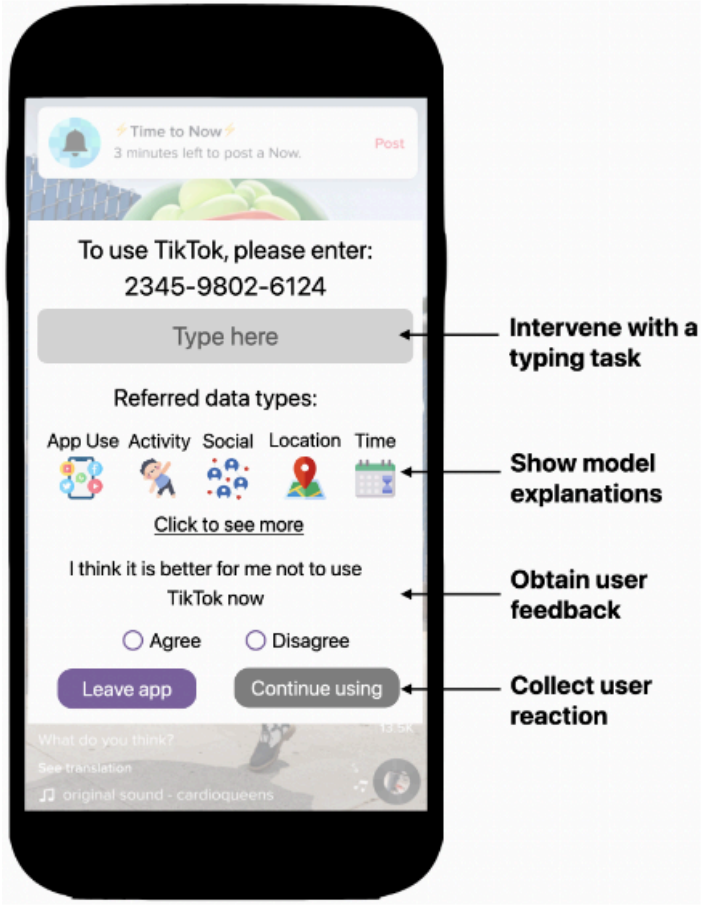
- AI-based JITAI for Smartphone Overuse
  - Most prior work are rule-based
  - Without a human-in-the-loop setup
- Integrating XAI into JITAI-based Smartphone Interventions
  - Unexplored in prior work
  - Potential to improve transparency, handle confusion, and cultivate user trust
- *What should be the control condition or the benchmark to compare with?*

What is Time2Stop?

An intelligent, adaptive, and explainable JITAI system

- Intelligent: leverages machine learning
- Adaptive: collects user feedback to adapt the intervention model over time
- Explainable: introduces interventions with AI explanations
- JIT: intervenes at the optimal intervention timings

- *How to isolate the effect of each feature?*



System Overview

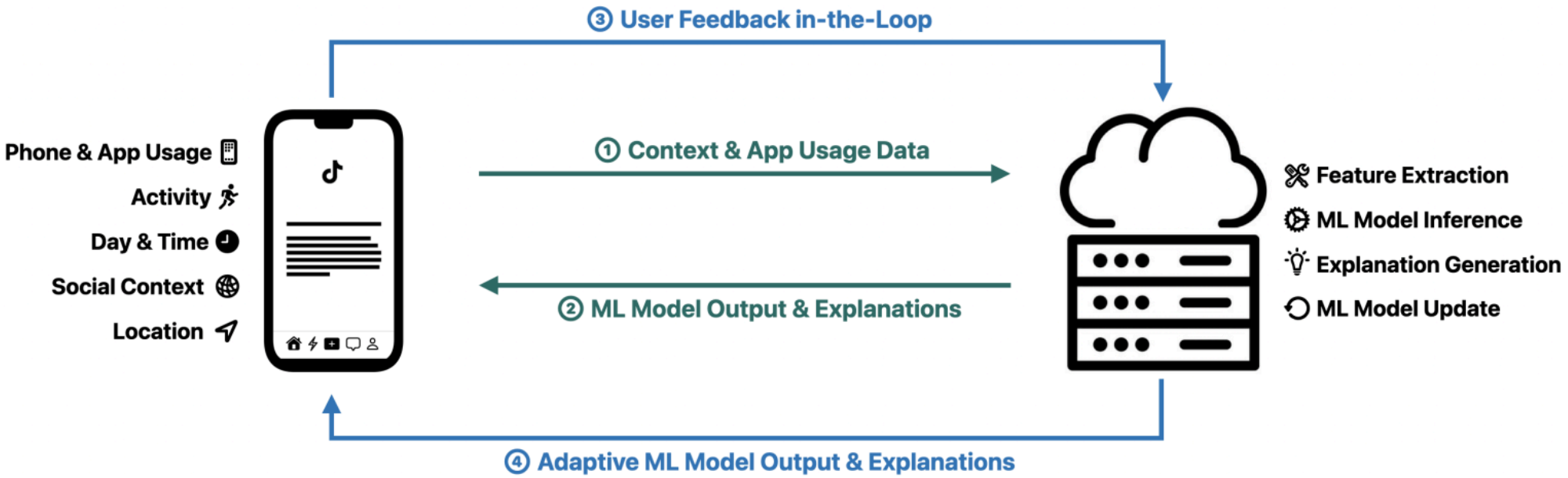


Figure 1: Time2Stop System Overview. The overall interaction flow consists of two loops. The first loop (green) includes: ① The mobile app continuously gathers contextual and app usage data (left) and transmits them to the cloud server. ② On the cloud server's end, feature extraction, ML model inference, 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.

ML Pipeline: Feature Design

(1) Phone and App Usage

- Locking & unlocking of the screen
- Battery usage: consumption rate, charge status
- App usage: count, min, max, mean, SD, sum of frequency and time
- Input interactions: scrolling, tapping, focusing
- Notifications

(2) Activity

- Stationary and mobile durations
- Ambient light

ML Pipeline: Feature Design (Cont.)

(3) Social Context

- Text message-derived features (e.g., first message time, top contacts)
- Bluetooth signals as a proxy for social contexts

(4) Location

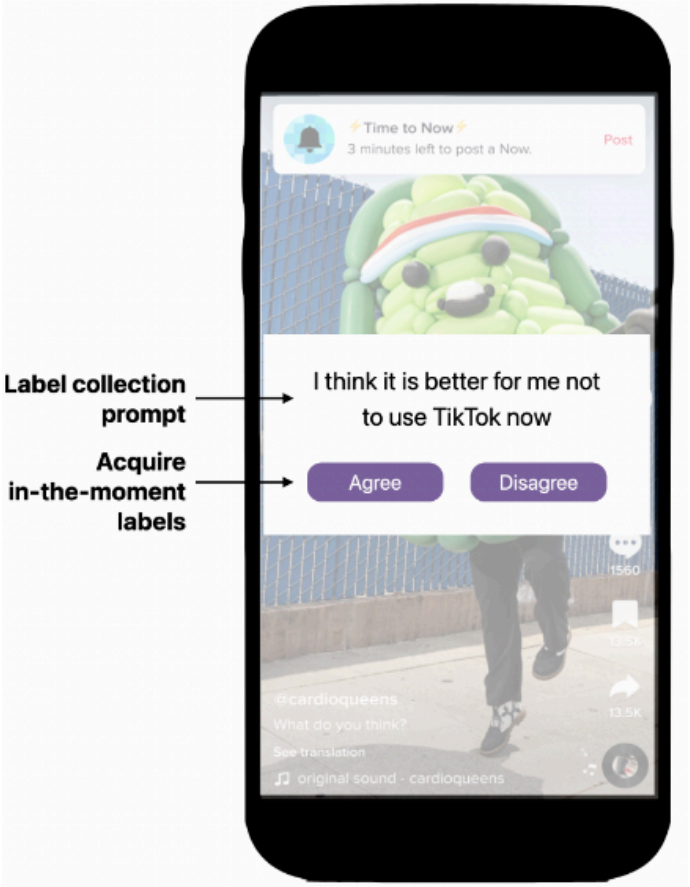
- Location type, variance, entropy *(not sure what those really are)*
- Time at the most-visited places, time at home

(5) Time

Is it collecting too many data? Will demographic data be useful?

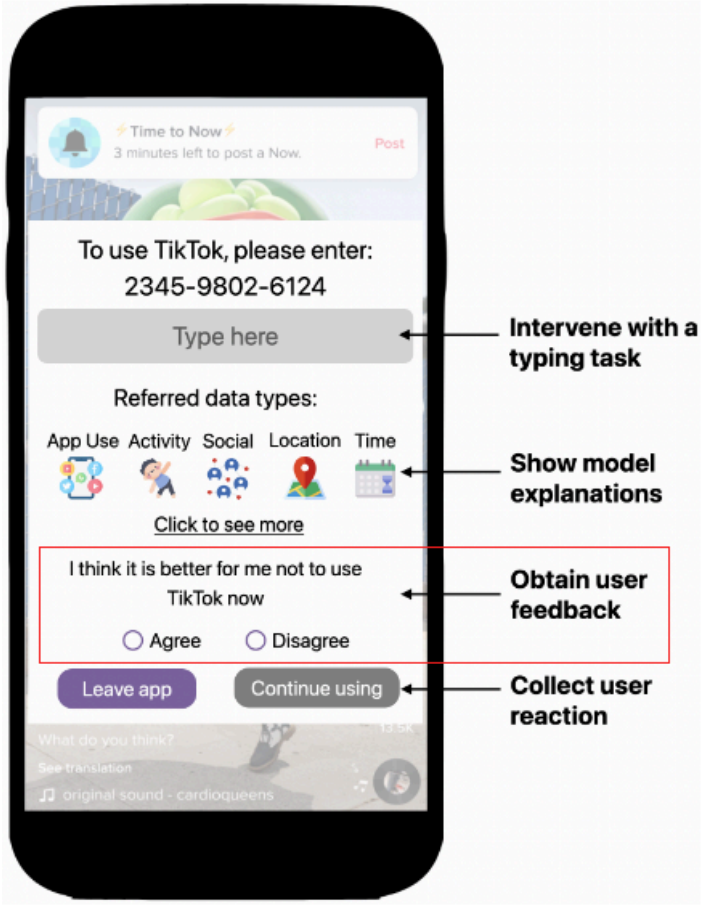
ML Pipeline: Label Collection

- Ecological momentary assessment (EMA)
  - *Experience sampling method (ESM)*
  - Whether users are overusing the phones
  - Ensures timely, contextually relevant labels
- Three in-the-moment label collection rules:
  - Entry-moment: opening a monitored app
  - Leaving-moment: leaving a monitored app
  - During Usage: every 10 minutes during usage
    - *Is 10 minutes granular enough?*
- A *unknown* cool-down interval *(5 minutes, I guess)*



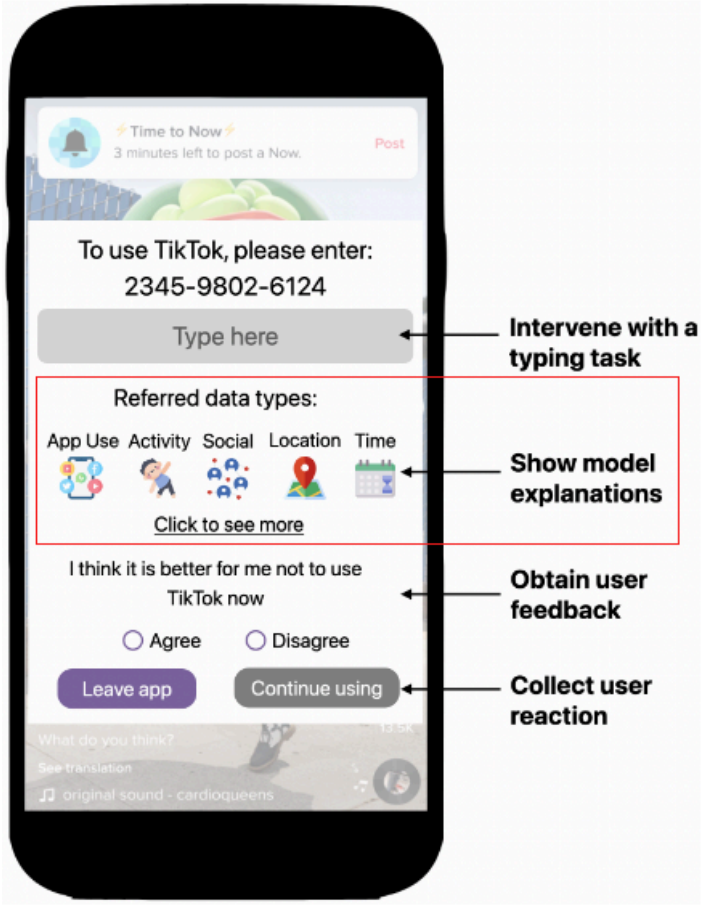
ML Pipeline: Adaptive Model Updates

- User feedback (optional) serves as new labels for updates.
- Daily updates from 12 AM to 1 AM
- Adopts decay-based sample weight assignment (recent data receives relatively higher weights)
  - Most recent day's weight: 1.0
  - Linearly decreases to 0.5 every half-week
  - To capture current trends while gradually reducing the impact of outdated information



ML Pipeline: Model Explanation Generation

- The top features contributing to an "overuse" prediction, calculated with SHapley Additive exPlanations (SHAP)
- Two Explanation Detail Levels:
  - High-level: (up to three) feature categories (e.g., "location," "activity," "app usage")
  - Low-level: feature description (e.g., "time at frequent locations")
- High-level explanations by default, low-level explanations by clicking the icons.



Examples of Explanantion

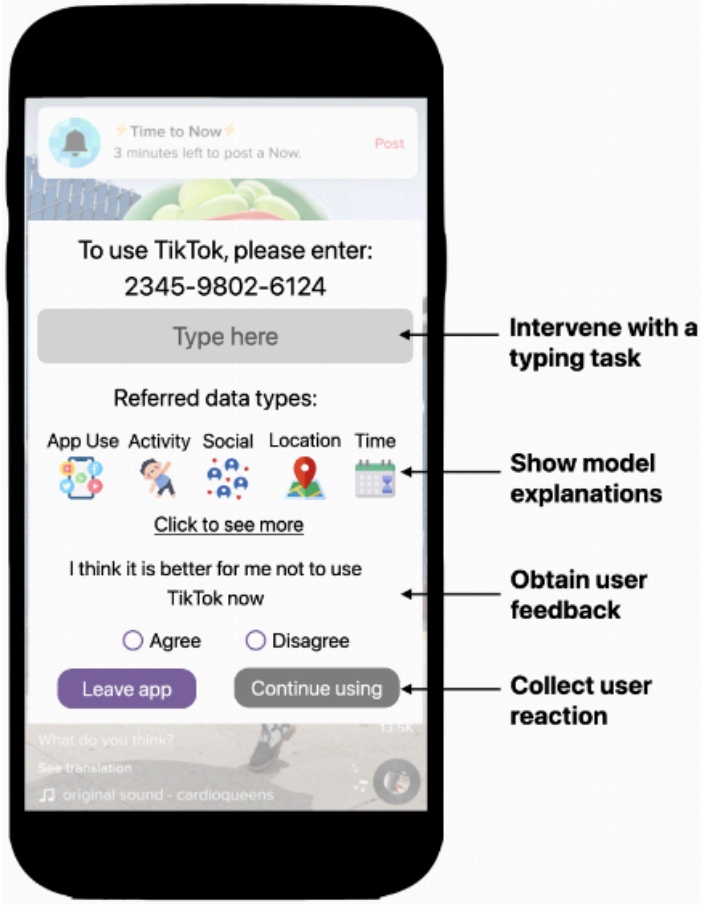
Table 2: Examples of Feature Explanations at Different Explanation Levels.

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

- *How would you feel upon seeing each explanation?*

Intervention

- Users select their "monitored apps"
- ML model predicts **"overuse" or not** based on current context and app usage upon app launching and every five minutes afterward.
  - *Is five minutes granular enough?*
- If "overuse", type 12 randomly generated digits to **continue using** *(even if the feedback is "Disagree")*, or directly **leave app**
- Neither easily circumvented nor overly restrictive
  - *How restrictive or annoying is it?*



System Overview

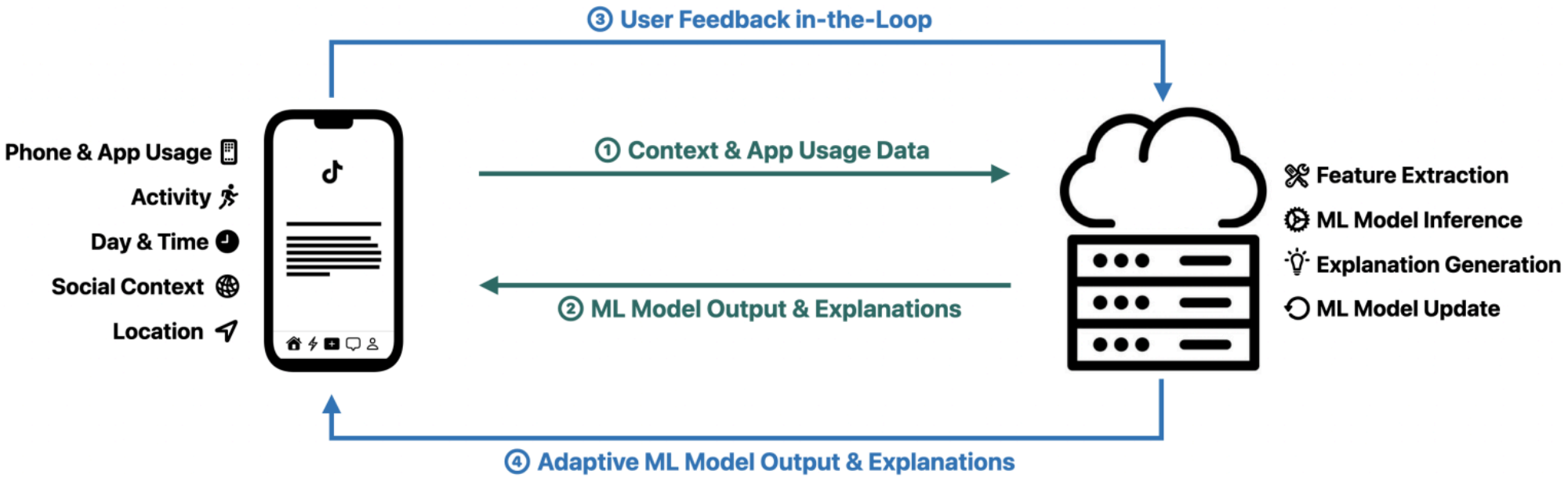


Figure 1: Time2Stop System Overview. The overall interaction flow consists of two loops. The first loop (green) includes: ① The mobile app continuously gathers contextual and app usage data (left) and transmits them to the cloud server. ② On the cloud server's end, feature extraction, ML model inference, 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.

Participants

- Recruited from university community forums
- Selected based on Android smartphone use and high Smartphone Addiction Scale (SAS) score (>120)
- 71 participants completed the study (48 females, 23 males; aged 21.8 ± 2.3, range 18-27)
  - Initial 127 participants
  - 49 discontinued due to personal reasons, software/hardware issues, compatibility, privacy, or battery concerns
  - 7 more excluded for insufficient data
  - *71/127 = 56%*

Procedure

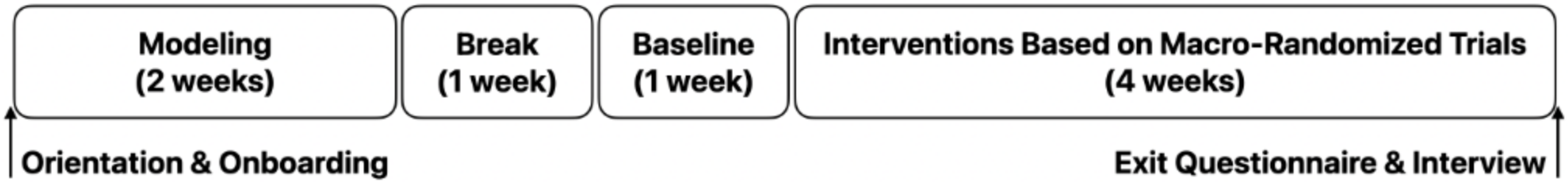


Figure 4: Field Experiment Flowchart

- Modeling (2 weeks): to collect features and labels for the initial models training
- Break (1 week): to mitigate carry-over effects from label collection
  - Label collection may affect phone usage.
- Baseline (1 week): to collect app usage data with no intervention
- Intervention (4 weeks): daily micro-randomized trials of four interventions
  - *What about the label collection for models updating here?*

Intervention

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)	✓	✓	✓

- *What should the Control group be like based on the research gap?*
- *Other possibilities: Non-Personalized (Global), Personalized-w-Exp*

Intervention (Cont.)

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)	✓	✓	✓

- Control: intervened based on a fixed probability derived from each participant's percentage of **overuse** during the initial modeling phase (37.8% on average)
- Personalized: the user's own data has a higher weight (10:1) to other users' data

Intervention (Cont.)

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)	✓	✓	✓

- Interface for Control, Personalized, and Adaptive-wo-Exp was identical *(i.e., no explanation)* to reduce bias.
  - *Which should be the baseline for the Adaptive-w-Exp? No Exp, or random Exp?*

Order of the Four Conditions

Micro-Randomized Trials (MRT).

- A technique optimized for JITAI within mHealth.
- Within-subject design: Each participant experienced all four intervention types.
- Intervention type was altered on a daily basis.
- Latin Square design (n=4) used to diversify the order and reduce order effects.
- Participants were **not informed** of the specific order or dates of intervention types during the study to reduce cognitive bias.
  - *But were they blind to what condition they were in each day?*

Evaluation Metrics — Quantitative

1. Intervention Accuracy: proportion of interventions marked as "Agree"
2. Intervention Receptivity: proportion of interventions responded by "Leaving app"
    - *What are accuracy and receptivity? What is the relationship between them?*
3. App Usage Duration: **total** time spent on monitored applications
4. App Visit Frequency: number of times monitored applications were launched
    - *Not about stopping, but **starting** the usage*

- *total time = average session length * visit frequency.*
- *How will Time2Stop affect your phone usage behavior, if you were a participant?*
- *Total screen time including the un-monitored apps?*

Evaluation Metrics — Qualitative

- Post-Study Questionnaire *(qualitative?)*
  - Participants were informed of the exact dates for each intervention type at the end *(of the four-week)* to aid recall.
  - To rank the four intervention types based on their preferences.
  - To rate perceived accuracy, effectiveness, and level of trust for each type.
- Semi-structured Exit Interviews
  - Questions: "What do you think of the four intervention techniques?" "Reason behind preference ranking?" "Thoughts on explanations?"
  - Analyzed using thematic analysis

I will not follow the structure of the paper in reporting the results.

Data Overview

- Collected Data
  - 497,458 minutes of usage from 149 apps (17 ± 5 apps per person)
  - 207,898 app sessions
  - 75,670 modeling-phase labels (60.5% entry, 24.5% using, 14.9% exit)
    - *75670 / 71 / 14 = 76.13 per person per day*
    - *Why such difference between the three? Why fewer "using" than "entry"?*
  - 39,188 intervention-phase labels (user feedback)
    - *39188 / 71 / 28 = 19.71 per person per day*

Model, Feature, & Intervention Frequency

- **Random Forest** performed better (F1 = 66.7%) than NB, LR, SVM, DT, & KNN
- Key Features
  - Phone unlock duration *(current session length)*
  - Movement patterns (travel distance, static ratio)
  - Nighttime usage
- Control sent more interventions than the other three conditions
  - *How to interpret this?*

Quantitative Results

- Four *(or five)* metrics
  - Accuracy ("Yes, I agree I am overusing it.")
  - Receptivity (Choosing "leaving app")
  - Visit frequency (I opened Twitter **10** times today)
  - *Session length (Each time I spent **5** minutes in Twitter)*
  - Total time (I spent **50** minutes in Twitter today)
- Four groups: Control, Personalized, Adaptive-wo-Exp, Adaptive-w-Exp
- What is effect of Adaptation and Explanation?
  - Click here to make your prediction

Adaptation → Accuracy & Receptivity

- Normalized data against Control, *original data not provided*
- Adaptive = Adaptive-wo-Exp and Adaptive-w-Exp merged
  - *Or should we just use Adaptive-wo-Exp?*
- Accuracy: Adaptive outperformed Personalized
- Receptivity: Comparison between Adaptive and Personalized not reported
- *Is the effect coming from the adaptation or explanation?*

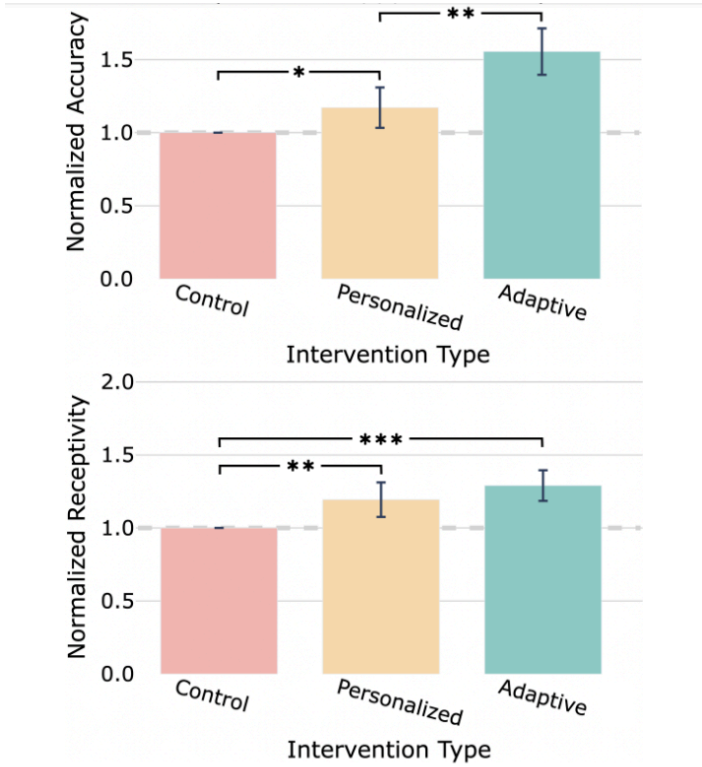


Figure 5: Intervention Accuracy (Top) and Receptivity (Bottom) Comparison across Three Intervention Types. Error bar indicates standard deviation. The same below. The two adaptive versions (with and without explanation) are merged into `Adaptive` to highlight better that adaptive ML-based methods had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

Adaptation → Accuracy & Receptivity *(Mine)*

- *If we compare Adaptive-wo-Exp (blue) with Personalized (yellow)*
  - *Accuracy: not reported*
  - *Receptivity: not reported*
- *I will not say that adaptation alone improved accuracy or receptivity.*

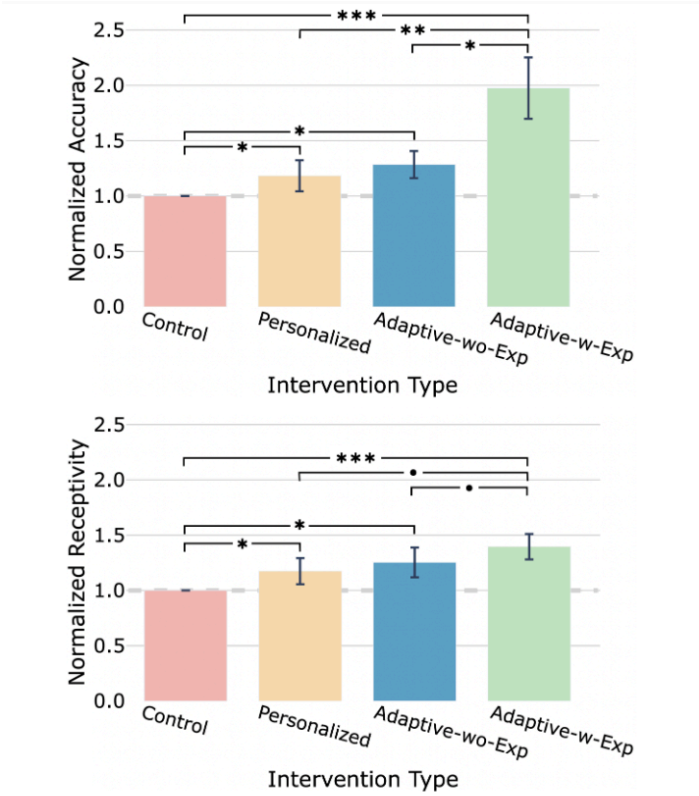


Figure 6: Intervention Accuracy (Top) and Receptivity (Bottom) Comparison across Four Intervention Types. The two versions of `Adaptive` are divided (*Adaptive-w-Exp* and *Adaptive-wo-Exp*) to better highlight that adding explanations can further enhance the performance of interventions.

Adaptation → Usage Behavior

- Normalized data *(against what? Why baseline is not 1.0?)*
- Visit frequency
  - Adaptive-wo-Exp (blue) was fewer (8.9%) than **baseline** ($p < 0.01$)
  - Adaptive-w-Exp (green) was marginally fewer (7.0%) than **baseline** ($p = 0.07$)
  - "This showed the advantage of the two adaptive methods over the Personalized and Control methods." (Page 12, Section 6.4.1) *really?*
- Total time: no significant difference

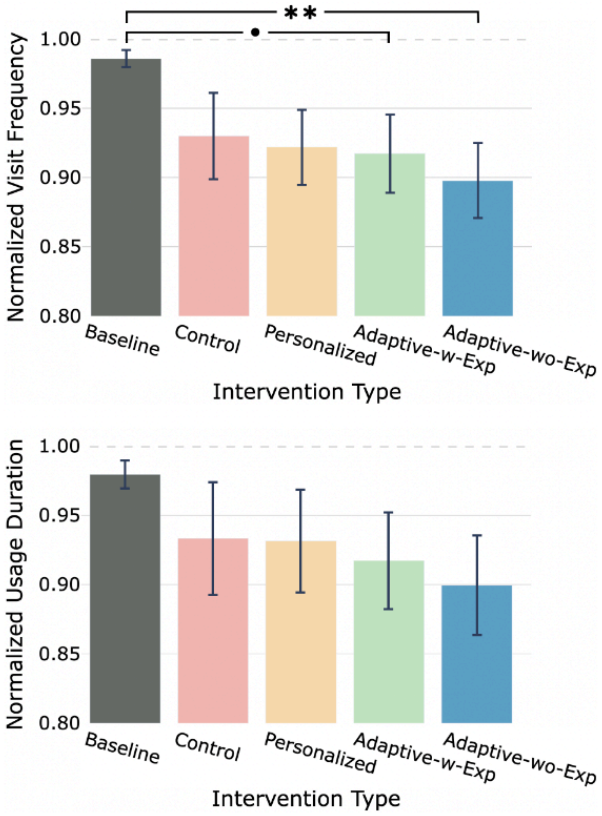


Figure 8: App usage visit frequency (Top) and usage duration (Bottom). The two adaptive methods reduced the most app visit frequency and usage duration. Interestingly, in contrast to Figure 5-7, showing explanations did not augment the performance from the perspective of app usage behavior.

Adaptation → Usage Behavior *(Mine)*

- *What about session length?*
- *I calculated the session length:*

Group	visit frequency	total time (minute)	session length (second)
baseline	95.0	214.0	135.2
control	88.3	199.7	135.6
personalized	87.6	199.2	136.5
adaptive-wo-Exp	85.3	192.4	135.4
adaptive-w-Exp	87.1	196.2	135.2

- *I guess there was no difference in session length among groups.*

Summary of Adaptation's Effect

- Accuracy
  - Adaptive-w-Exp and Adaptive-wo-Exp combined, better then Personalized
  - Adaptive-wo-Exp alone, better than Control but not better than Personalized
- Receptivity. Better than Control but not better than Personalized
- Visit frequency. Fewer than baseline, but not Control or Personalized.
- Session length. No effect
- Total time. No effect

"Adaptive models significantly outperform the baseline methods on intervention accuracy and receptivity." (Abstract)
Adaptation alone did not improve any metrics beyond Personalized

Explanation → Accuracy & Receptivity

- Accuracy: Adaptive-w-Exp outperformed Adaptive-wo-Exp
- Receptivity: Adaptive-w-Exp marginally outperformed Adaptive-wo-Exp

Explanation improved accuracy and receptivity.

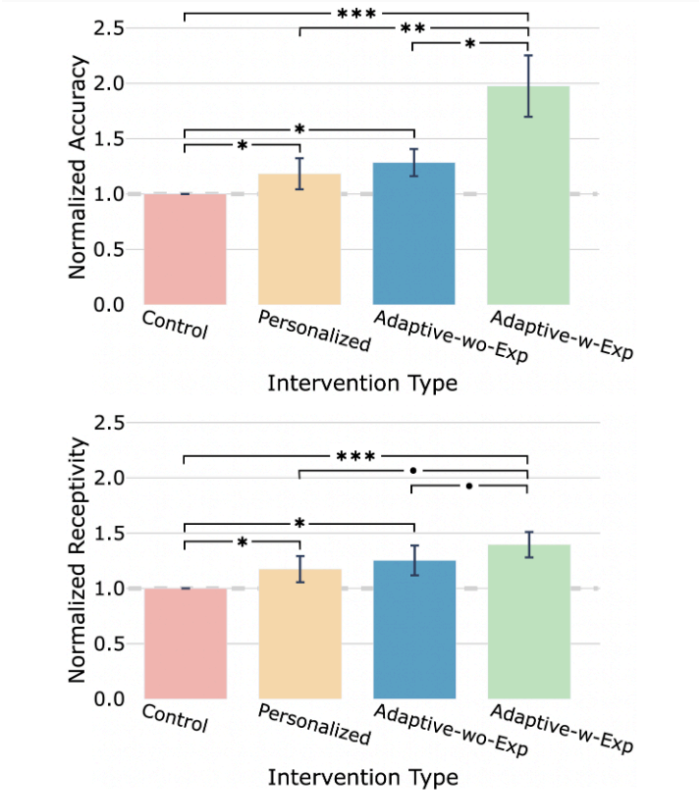


Figure 6: Intervention Accuracy (Top) and Receptivity (Bottom) Comparison across Four Intervention Types. The two versions of Adaptive are divided (*Adaptive-w-Exp* and *Adaptive-wo-Exp*) to better highlight that adding explanations can further enhance the performance of interventions.

Explanation → Usage Behavior

- Visit frequency
  - Adaptive-w-Exp (green) was marginally fewer (7.0%) than **baseline** (p = 0.07)
- Total time: no significant difference
- *Session length: no significant difference*

The comparison between Adaptive-w-Exp and Adaptive-w-Exp suggested that adding explanation increased (insignificantly) the visit frequency and total time.

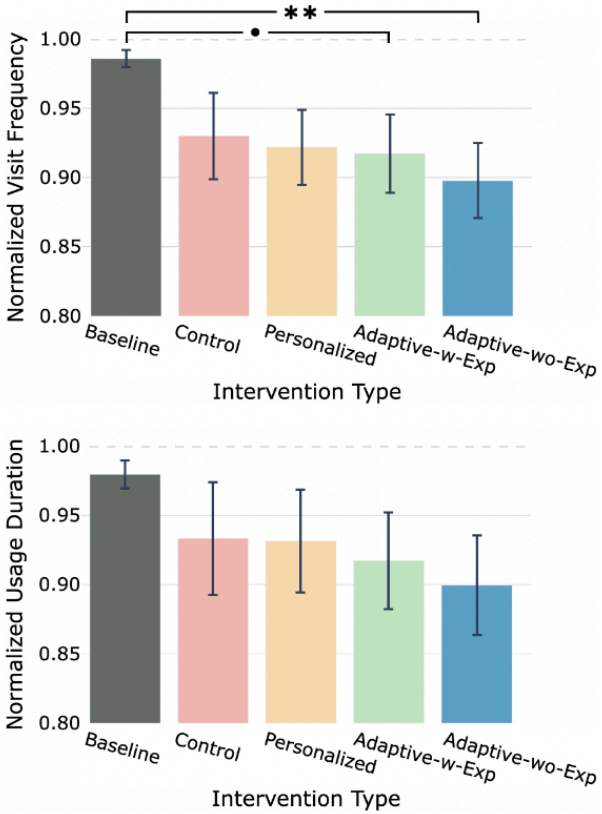


Figure 8: App usage visit frequency (Top) and usage duration (Bottom). The two adaptive methods reduced the most app visit frequency and usage duration. Interestingly, in contrast to Figure 5-7, showing explanations did not augment the performance from the perspective of app usage behavior.

Results from the Questionnaire

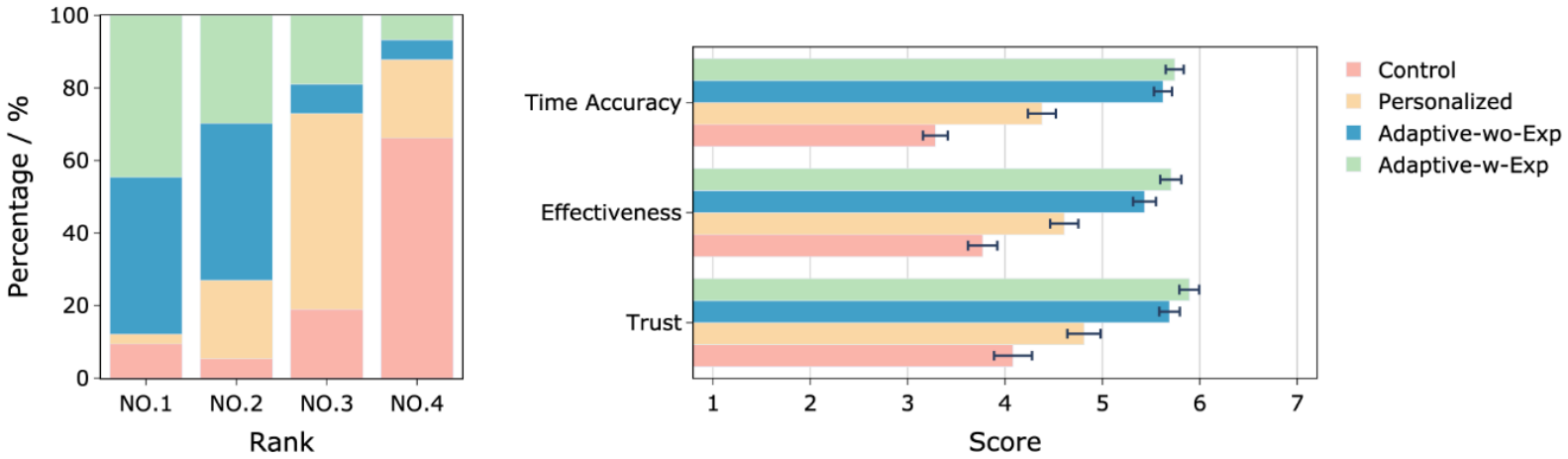


Figure 9: Survey Results Summary. (Left) User Preference Rankings among The Four Intervention Methods. (Right) User Ratings on Intervention Time Accuracy, Perceived Effectiveness, and Level of Trust to Different Methods. The two adaptive methods received the highest user subjective preference and ratings.

"the intervention order was presented to participants. They then filled out the questionnaire" (Page 9, Section 5.2)

- *It will be more convincing if interventions were refereed as A, B, C, & D.*
- *Is Adaptive-w-Exp ranked higher than Adaptive-wo-Exp?*

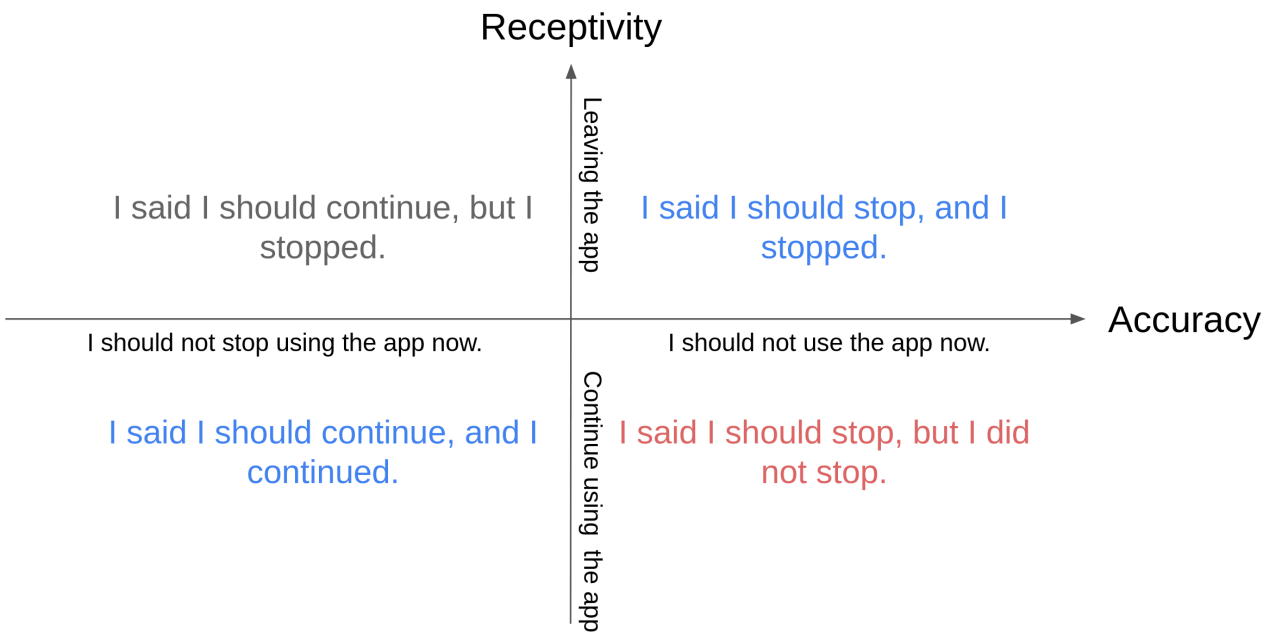
Results from the Interview

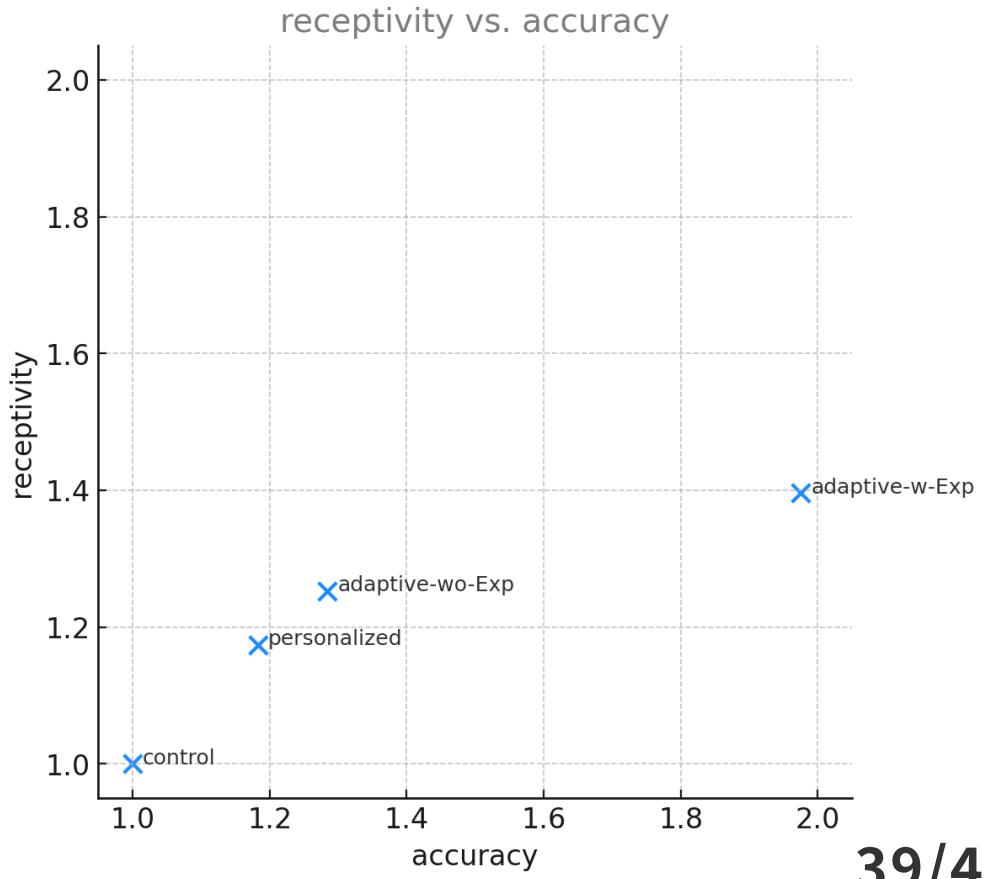
- General preference for adaptive intervention methods
  - Participants "felt the difference" as adaptive versions "had been learning" about their behavior, leading to more "comfortable" intervention timing over time.
- Mixed reaction to the explanation
  - Some found explanations fostered self-awareness, prompted reflection, and built trust.
  - Others viewed them as "overly broad," confusing, or "unnecessary."

Summary of the results

- Control condition (random intervention) reduced visit frequency (and thus total time) but not session length.
- Personalized condition outperformed Control in accuracy and receptivity but not in usage behavior change.
- Adaptive-wo-Exp had similar performance to Personalized condition.
- Adaptive-w-Exp outperformed Adaptive-wo-Exp in accuracy and (marginally) receptivity but not in usage behavior (actual insignificantly worse than Adaptive-wo-Exp).

What are accuracy, receptivity, & the gap in between?





More Advanced Explanation Generation?

"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." (Page 15, Section 7.1)

- *Is explanation always necessary and beneficial?*

Thank you!

- The slides are made with Marp, Marp for VS Code, and Marp CLI.
  - Theme academic by Kaito Sugimoto, slightly modified by me
  - Some scripts are generated by ChatGPT.
- I will keep the repository public until the end of the week if anyone is interested.