

What Social Media Use Do People Regret? An Analysis of 34K Smartphone Screenshots with Multimodal LLM

Longjie Guo

The Information School
University of Washington
Seattle, Washington, USA
longjie@uw.edu

Xuhai Xu

Department of Biomedical
Informatics
Columbia University
New York City, New York, USA
xx2489@columbia.edu

Yue Fu

Information School
University of Washington
Seattle, Washington, USA
chrisfu@uw.edu

Xiran Lin

Global Innovation Exchange
University of Washington
Bellevue, Washington, USA
xiranlin@uw.edu

Yung-Ju Chang

Department of Computer Science
National Yang Ming Chiao Tung
University
Hsinchu, Taiwan
armuro@nycu.edu.tw

Alexis Hiniker

Information School
University of Washington
Seattle, Washington, USA
alexisr@uw.edu

Abstract

Smartphone users often regret aspects of their phone use, especially social media use. However, pinpointing specific ways in which the design of an interface contributes to regrettable use can be challenging due to the complexity of social media app features and user intentions. We conducted a one-week study with 17 Android users, using a novel method where we passively collected screenshots every five seconds, which we analyzed via a multimodal large language model to understand participants' usage activity at a fine-grained level. Triangulating this data with data from experience sampling, surveys, and interviews, we found that regret varies based on user intention, with non-intentional and social media use being especially regrettable. Regret also varies by social media activity; participants were most likely to regret viewing algorithmically recommended content and comments. Additionally, participants frequently deviated to browsing social media when their intention was direct communication, which slightly increased their regret. Our findings provide guidance to designers and policy-makers seeking to improve users' experience and autonomy.

CCS Concepts

- Human-centered computing → Empirical studies in ubiquitous and mobile computing.

Keywords

screenshots, regret, digital well-being, multimodal large language model, social media

ACM Reference Format:

Longjie Guo, Yue Fu, Xiran Lin, Xuhai Xu, Yung-Ju Chang, and Alexis Hiniker. 2025. What Social Media Use Do People Regret? An Analysis of 34K Smartphone Screenshots with Multimodal LLM. In *CHI Conference on Human Factors in Computing Systems (CHI '25)*, April 26–May 01, 2025,



This work is licensed under a Creative Commons Attribution 4.0 International License.
CHI '25, Yokohama, Japan

© 2025 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-1394-1/25/04

<https://doi.org/10.1145/3706598.3713724>

Yokohama, Japan. ACM, New York, NY, USA, 23 pages. <https://doi.org/10.1145/3706598.3713724>

1 Introduction

Despite the popularity of social media platforms, smartphone users consistently say that they engage with social media in ways they later regret. They report scrolling through content that is not worth their time and checking for updates compulsively [10, 56]. This regrettable usage is no accident; in the attention economy, online platforms are often intentionally designed to keep users on task for as long as possible and to manufacture routine phone-checking habits that keep them coming back as often as possible [19]. An increasing body of prior work argues that the concept of designing for “engagement” has received too little scrutiny and deserves greater attention from regulators [53]. This is particularly true of social media platforms, which are frequently built around attention-economy business models and provide the usage experiences that users are most likely to say they regret [10, 56].

Yet, understanding the precise relationship between users’ engagement with social media apps on the phone and their subsequent regret is incredibly challenging. First, users can engage in a variety of activities throughout their interaction in a given app, as many mobile apps today consist of a complex constellation of features. A single social media app might include, for example: a feed of algorithmically organized content, direct messaging features, discussion threads, and more, making it difficult to associate isolated design decisions with users’ usage decisions or subjective experiences. Indeed, prior research has examined the relationship between feature-level social media app usage and subsequent feelings of regret [10] and showed that regret of social media usage can vary by feature.

Second, people bring many different intentions to their app use, and prior work suggests a user’s motivation for engaging with an app may influence whether they later regret doing so [29, 37]. For instance, earlier studies have shown that users often turn to mobile apps to kill time during idle moments [9, 50, 67, 73], typically without a clear purpose. It is plausible that users’ sense of regret after such mindless use differs from their feelings after engaging with the same app for specific, goal-directed tasks, such as communicating

with a friend or seeking information. Finally, prior work suggests that whether a user's actual activity aligns with their intention may also influence regret [10, 51, 70]. Thus, in this study, we sought to answer the following research questions:

- RQ1: How does regret vary based on users' different intentions for app use?
- RQ2: How does regret vary based on the specific activities users engage in on social media apps?
- RQ3: How does regret vary based on whether users' actual activities on social media apps align with their intention?

To answer these research questions, we conducted a one-week, mixed-methods study with 17 Android users. We examined participants' phone use to enable us to understand their use of social media apps—and any regret it produces—in the larger context of all phone use. By documenting all forms of phone use and measuring associated regret (if any) across a variety of usage experiences, we were able to compare social media regret with that of other use cases. Specifically, we used the experience sampling method to capture user intentions in real-time, combined with automatic and continuous screenshot capturing every five seconds. The collected screenshots were then analyzed by a multimodal large language model [42], which is capable of interpreting visual content and generating descriptive information. We used this novel approach to systematically and efficiently analyze fine-grained phone-use behaviors in social media apps, overcoming the scalability challenges of manually coding screenshots exhaustively and the lack of flexibility of alternative approaches, such as directly retrieving UI components of an app [10, 46]. At the end of each day, participants reviewed screenshots of their own usage and labeled the sessions that they regretted. This gave us the opportunity to examine short-term, action-based regret (which prior work differentiates from actions one *failed* to take [24]). We then complemented these daily annotations with retrospective interviews, where users reflected on their phone use over the course of the study.

Our results first showed that participants' regret varied based on intended use, with non-intentional use being most regretful, and that participants regretted social media use more than other types of phone use. Then, through our LLM-powered analysis of 34,313 smartphone screenshots, we found that on social media apps, regret varied based on social media activity, and that participants found viewing algorithmically recommended content and comments most regretful. Comparing participants' initial intentions for using social media apps with the activities they ultimately engaged in reveals that participants frequently intended to communicate but ended up engaging in other social media activities instead (over 60% of the time). Additionally, participants expressed slightly more regret for sessions where they deviated from their original intentions compared to sessions where they did not. Our regression model showed that duration of app use, user intention, and proportions of recommendation- and subscription-based content and comments are significant indicators of regret. By investigating both intention and fine-grained social media activity and how they are related to regret, we provide guidance for designers and policy-makers to design experiences that respect users' autonomy and intentions.

2 Related Work

Smartphone users are presented with an infinite number of usage choices, including which apps to use, how to engage with them, and what media content to consume. Further, many platforms (especially social media platforms) rely on attention-economy business models, which has led to designs that pressure users into spending as much time with the product as possible [40]. This combination of information overload and pressure from platforms creates a context where users, at times, feel regret about their technology use and realize that their time would have been better spent elsewhere [56]. Because regret is predominantly experienced as a negative and aversive emotion [70], understanding and regulating it is highly consequential to digital well-being. Here, we describe the prior strands of research that we draw on in examining some of these regrets. This includes: prior work studying the technology use that people find problematic, interventions to address this problematic usage, technical methods for studying problematic usage behaviors, and theoretical conceptualizations of regret.

2.1 Understanding Problematic Smartphone Use

Although the widespread adoption of smartphones has brought numerous benefits, their increasing ubiquity and utility have also heightened users' attention to them. Recent research shows that many smartphone users now check their phones every few minutes, with a significant portion of these interactions being self-initiated [15, 26, 49]. A large body of research has investigated user's experience with problematic phone use [8, 10, 17, 22, 37, 48, 56]. For example, Chan used a survey to examine the relationship between phone use and subjective well-being and found that communicative uses of mobile phone are positively related to subjective well-being while non-communicative uses are negatively related to subjective well-being [8]. Tran et al. found that compulsive phone use might be triggered by unoccupied moment, tedious task, social awkwardness, and anticipation, and that users express frustration with such compulsive checking habits, unless the checking behavior resulted in experiences that they find meaningful and that transcend phone use, such as relationship-building [56].

One insight coming out of this body of work is that understanding phone use at the app level or app category level may be insufficient and too coarse-grained, since many mobile apps today, especially social media apps (which are most likely regrettable [10, 56]), offer a variety of features, where user behavior might vary depending on the feature [10] or depending on how they intend to use the app [37]. Drawing on the Uses and Gratifications theory, Lukoff et al. investigated what smartphone use is meaningful and meaningless to people by asking for their motivation and type of phone use, and found that people feel a lower sense of meaningfulness when their phone use is motivated by habitual use to pass the time and when people use their phone for entertainment and passively browsing social media [37]. By using a feature-level analysis approach and incorporating the construct of regret, Cho et al. found that users felt more regretful about social media features that comprise passive forms of usage (such as viewing social media feed) than active forms of usage (such as searching and messaging), and that they regretted habitual checking on their feed, sidetracking from original intention to recommendation-based features, and falling into

prolonged use when viewing recommendation-based content [10]. In this study, we incorporate these perspectives, and combine the experience sampling method and fine-grained understanding of user's moment-to-moment behavior via passively collected screenshots to investigate how regret varies based on what motivates people to use their phone (their intention) and what they subsequently do once they are on their phone, offering more nuanced perspectives into regrettable phone-use behavior.

While much digital well-being and problematic smartphone use research seeks to understand smartphone users and design individual-oriented solutions to reduce problematic phone use, there are also exogenous factors underlying technology overuse, as technology is often intentionally designed to optimize engagement and nudge users towards problematic use in the attention economy [19, 21]. An increasing body of prior work has argued that the concept of designing for “engagement” in technology platforms deserves more scrutiny from regulators [5, 18, 53]. Richards & Hartzog outlined several harms associated with the engagement model of digital platforms, including privacy violations, the erosion of attention (coined as “attention theft”), and detrimental impacts on mental health, relationships, and democratic processes, and argue that wrongful engagement strategies should be regulated [53]. Our study provides insights into how the design of digital platforms aiming at maximizing engagement can create regrettable experiences for their users and sheds light on how policy-makers can help improve user's experience and autonomy.

2.2 Problematic Smartphone Use Interventions

Due to the negative effects of smartphone use, many tools and intervention mechanisms aiming at reducing smartphone use have been implemented. Both Android and iOS have a default system-level tool to track app usage and set limits [25, 30]. Researchers have also demonstrated the effectiveness of various intervention techniques to reduce smartphone use [28, 31, 35, 46, 47, 64, 65]. For example, MyTime uses three mechanisms: timer, timeout, and aspiration to help decrease overuse [28]. InteractOut leverages input manipulation techniques to inhibit natural user gestures on smartphones to reduce overuse (e.g., by delaying tap and swipe) [35]. Type-Out leverages self-affirmation (letting users type statements like “I value self-control” when entering an app) to reduce smartphone overuse [65]. MindShift utilizes large language model to adaptively tailor intervention message based on user's mental state [64]. However, most of these interventions are applied at the granularity of the app or device level, where once an app is opened or a usage limit is surpassed, the proposed intervention mechanism would discourage users from continuing to use the phone. Recent work has also implemented more fine-grained feature-level interventions, where users can set limits on specific features inside social media apps (e.g., consuming content on social media feed), which has been shown to decrease passive usage related to content consumption more than app-level intervention [46]. In addition to mechanisms that use time or app launch as triggers, researchers also proposed intervention mechanisms that use machine learning models to identify best timing for intervention based on smartphone sensor and log data to deliver just-in-time intervention [47].

While we do not propose a new intervention design in this paper, our empirical findings can inform the design of intervention techniques that consider user intention and fine-grained activity to reduce problematic phone use. Our screenshot analysis approach can also inform intervention systems that aim to leverage multimodal data to deliver fine-grained and just-in-time intervention.

2.3 Capturing Smartphone Screenshots

Past work on understanding phone-use behaviors has mostly utilized methods such as interviews (e.g., [56]), phone log data (e.g., [20, 29, 55]), and experience sampling data (e.g., [37]). While these approaches have helped yield insights about phone-use behavior, they can be limited since they do not reveal user's moment-to-moment behavior precisely. Some researchers have also utilized the Android Accessibility API to detect specific social media features, but this approach requires access to an app's UI structure and manual coding of all features, making it hard to scale and not robust against app updates [10]. Researchers have also turned to smartphone screenshots, which provide richer information about user's behavior. Reeves et al. developed the Screenomics framework, where a system can capture and analyze personal experiences through passively collected screenshots every five seconds [52]. In addition to understanding user behavior, screenshots can also be used to predict user behavior. Yang et al. demonstrated that screenshots collected every five seconds can be used to predict task switching [66]. Chen et al. leveraged both screenshot data and sensor data to create a fusion model to predict time-killing behavior on smartphone [9]. The present study primarily focuses on understanding phone use behavior through screenshots (and how certain behaviors correlate with regret), not predicting certain behavior from screenshots.

While screenshot data captured in periodic intervals reveal rich information about user's behavior, one challenge of using such data is that they are hard to analyze. Recent advances in multimodal large language models (MLLMs) such as GPT-4o (and its predecessor, GPT-4V) have shown exceptional multimodal understanding capabilities across various domains and tasks [42, 68]. These models extend the capabilities of large language models (LLMs) by integrating multiple modalities (most typically, vision and language), and prior research has demonstrated that these systems excel across a spectrum of tasks, from simpler ones such as open-ended image description and object localization to more complex challenges such as understanding multi-image sequences and navigating graphical user interfaces (GUIs) [68]. The GUI understanding capability of MLLMs is of particular interest to understanding phone-use behavior, as researchers have demonstrated that, based solely on screenshots, MLLMs can understand the visual content of mobile UI screens (e.g., summarizing content or activity from screenshots) across different apps, and even operate these apps by predicting future actions [62, 68, 71]. Other studies demonstrated that MLLMs can automatically generate code based on screenshots [61, 68]. Such capabilities make MLLMs an ideal candidate to automatically analyze phone-use data recorded in screenshots, which can overcome the challenge of manual labeling and provide flexibility through prompting. In this study, we explore this possibility and introduce

a novel method to automatically categorize social media use into fine-grained activities using GPT-4o.

2.4 Theorizing Regret

A large body of literature in economics and psychology focuses on operationalizing and characterizing regret [3, 24, 39, 51, 69, 70]. First, there is a general consensus that, unlike many other emotions, regret is a *cognitive* emotion, where thinking and judgment are central [24, 70]. In economics, regret is defined as “*the consequence of decision-making under uncertainty*,” a reaction to the simple difference between experienced reality and rejected alternatives [3]. Research on counterfactual thinking highlights that the outcome of counterfactual realities can be imagined (rather than already known) and that the path by which a decision is made can also influence regret [39].

In Gilovich and Medvec’s seminal work on regret, the authors divide regret into regret of *action* (regretting something one has done) and *inaction* (regretting that one has not done something), and show that there is a temporal pattern to regret—namely that actions produce more regret in the short term while inactions produce more regret in the long term [24]. In other words, people experience more intense, immediate pain from a regrettable action, but over the longer arc of their life, they experience more enduring distress as a result of the ways in which they have failed to act.

Zeelenberg and Peters define regret as “*an aversive, cognitive emotion that people are motivated to regulate*” and “*a comparison-based emotion of self-blame, experienced when people realize or imagine that their present situation would have been better had they decided differently in the past*” [70]. They also explain that regret can be divided into *process regret* (regret that stems from a poor decision-making process) and *outcome regret* (regret that stems from dissatisfaction with an outcome), and that whether a decision is *justifiable* can play an important role in determining regret, independent of the decision outcome [70]. For example, intention-behavior inconsistency (not behaving in ways originally intended) can amplify regret, since behaviors that deviate from the original intention are often less justifiable, which exemplify poor decision-making processes [51, 70]. Importantly, prior work has argued that regret is not a unitary emotion and that the conceptual boundaries of regret are not always clear [24]. For example, the source of regret can vary from contexts ranging from moral transgressions to failures of self-actualization [24].

In this paper, we focus on action-based regret (and specifically, regret caused by using social media). Since this form of regret often manifests in the short term (as opposed to inaction-based regret, which is more likely to affect people over the course of their entire life span), we assessed people’s regret at the end of each day. Given that regret is not a unitary emotion and can have rich meanings, we use a rather inclusive and minimally restrictive working definition in our study, similar to prior work [24]. We asked participants whether they agreed that they felt regret about a phone use session. Our data collection method did not define regret for participants, leaving room for them to express process regret, outcome regret, or a combination of both.

3 Method

To investigate the social media use people regret, we conducted a one-week study, which consisted of an initial interview, a week of data collection in-the-wild (which included experience sampling method (ESM), passive screenshot collection, and daily questionnaires), and two follow-up interviews. Prior to launching the study, we piloted the data-collection tool and interview materials with three participants to refine the tool and materials.

3.1 Participants

3.1.1 Recruitment. We recruited 17 adults from social media channels (including X, Slack, Facebook groups, LinkedIn, Discord, and WeChat), university mailing lists, and on-campus fliers. All interested participants first responded to a screening survey after seeing a recruitment ad titled “Seeking Android Users for a Paid Research Study.” The screening survey consisted of multiple-choice and open-ended questions about their phone and phone use, the extent to which they wanted to change their phone-use habits, their motivation for joining the study, availability, potential privacy concerns regarding screenshot collection, and demographic information. We reached out to survey respondents who provided high-quality responses to the open-ended questions, reported using an Android phone as their main device with the Android version no earlier than Android 10 (which our custom-built data-collection app requires), and had access to Wi-Fi or unlimited cellular data (to make sure that they could upload data). We met with 20 potential participants for an initial interview. One was deemed ineligible to participate because they were not physically located in the U.S. The other two voluntarily decided not to participate, with one concerned about privacy, and another one unable to commit due to their schedule. The remaining 17 participants all finished the one-week in-the-wild data collection and the two follow-up interviews. All of the participants completed the study between July and August 2024.

3.1.2 Demographics. Among the 17 participants who finished the study, 4 identify as man, 12 identify as woman, and 1 identify as non-binary person. 3 of them reported they were between 18 and 24 years old, 7 between 25 and 34 years old, 5 between 35 and 44 years old, and 2 between 45 to 54 years old. In terms of ethnicity, 8 identify as White, 5 as Asian, 2 as Black or African American, 1 as Hispanic, Latino, or Spanish, and 1 as “Other.” In terms of level of education, 2 of them reported having some college education, 5 having a 4-year degree, 5 having a professional degree, and 5 having a doctoral degree.

3.1.3 Compensation. The 17 participants who completed the study all received a US\$200 Amazon gift card as a compensation. Two of the three participants who did not continue after the initial interview but finished the initial interview each received a \$10 Amazon gift card. The three pilot participants received a \$20, \$20, and \$60 Amazon gift card respectively, based on the number of days they participated in the pilot study (\$20 for each day).

3.2 Procedure

3.2.1 Initial Interview. Prior to data collection, we met with each potential participant to conduct an initial interview. In the interview, we first explained the procedure of the study (including how

frequently screenshots will be taken and our privacy-protection mechanisms), and asked participants general questions about their phone use. We then showed them the ESM survey questions we would be using for the study and assessed whether they were able to follow the instructions and understand the differences between the categories in the survey (details in Section 3.2.2). We invited participants who seemed interested and committed to join the study. For participants who consented to continue with the study, we helped them install our custom-built data-collection app (including granting the necessary permissions such as screen recording) and provided a short tutorial on how to use the app. Three of the initial interviews were conducted in-person, and the rest were done via Zoom.

3.2.2 In-the-Wild Data Collection. After the initial interview, each participant then went through the data collection process for seven days, which, as prior work suggested, is “likely to yield a fairly representative sample of the various activities individuals engage in” [27, 59], and is the duration used by one of the initial experience sampling studies [13]. We used ESM in combination with passive screenshot collection and daily questionnaires. During the week, our custom-built data-collection app ran in the background on their phone, and selectively asked about their intended use of different apps and took screenshots every five seconds in selected sessions. Although our study primarily focuses on regret in the context of social media, we still sampled other general phone use, which we used to compare with social media use.

To minimize prompting participants and taking screenshots too frequently, we decided to sample a subset of users’ phone use sessions rather than capturing all sessions. To diversify the sampled phone sessions, we adopted an algorithm that increases the likelihood of sampling a screen session (defined as the period between the screen is turned on and off) as the elapsed time since the last sampled screen session grows. The sampling intervals used were 10 and 120 minutes. Specifically, the likelihood of selecting a session within 10 minutes of the previously sampled session was set to 0, and this likelihood increased over time, reaching 100% after 120 minutes. In other words, the longer the phone went on without sampling a session, the more likely it was to sample the next detected session, and vice versa. For example, when the time difference is 65 minutes, there is a 50% chance of selecting the current screen session. The 10-minute interval was chosen based on the assumption that phone sessions occurring within this time frame are likely to share a similar context, which would conflict with our goal of sampling sessions across diverse contexts. The 120-minute interval was set so that we were guaranteed to get some sampled sessions throughout each day.

If a screen session was selected, every time the participant opened or switched to a new app (we define the period an app stays on the foreground as an *app session*, and one *screen session* contains at least one *app session*) and stayed on the app for longer than 5 seconds (we set this threshold to avoid asking participants too frequently when they fast switch between apps), the data-collection app would display a survey asking why they opened the app. When designing this ESM survey, we adopted the Uses and Gratifications (U&G) types in [37] and slightly modified the original categories by combining habitual use into the U&G types, which we called

“no specific goal.” This was mainly to reduce participants’ effort of answering two questions. Our survey question prompted the user to report their intention by asking “why are you here” and had 7 options: “no specific goal,” “get things done or self-improve,” “get information,” “communicate or interact with others,” “be entertained or have fun,” “browse social media,” and “I am not sure.” See Table 1 for the full definitions of these categories, and the abbreviations we use for the rest of the paper. The ESM prompt interface is shown in Figure 1. After the participant answered this question, the data-collection tool would start taking screenshots in the background every 5 seconds. To implement this, we used the MediaProjection API [14] on Android, which allowed our app to record screen content.

At 8 PM everyday, the app sent a notification to the participant to remind them to answer questionnaires about how regretful they felt about the app sessions captured by the data-collection tool. For each session, the app presented to them the starting time, the app name, the screenshots captured in the session, followed by a survey question measuring their regret for the session. Although there are existing standardized scales for measuring regret, such as the five-item Decision Regret Scale [7], we opted to use a simpler, one-item survey question, another common approach used in past studies to measure regret (e.g., [1, 54, 60]). This is because we needed participants to rate every app session, and we hoped that by using a lightweight survey question we would avoid overburdening participants and prevent a decline in response quality due to lengthy surveys [23]. The instruction and wording of our survey question resembles those used in [7, 60] and were adapted to the context of phone use. Also, given that regret is a rich and complex emotion (even when narrowed down only to action-based regret) that prior work defines broadly [24], we imposed minimal constraint in our survey question. Specifically, we asked: “Please think about this particular time you used your phone. Select how you feel about the following statement: I feel regret about this phone use session.” Participants answered this question on a seven-point Likert scale, from “Strongly Disagree,” to “Strongly Agree.” All of the interfaces are shown in Figure 1. One reason for assessing participants’ regret at the end of each day at 8 PM (when people are generally more available), not immediately after each session, is that we wanted to avoid disrupting their normal phone use. More importantly, since regret is inherently a cognitive-laden emotion [24] which involves comparing alternative options [70], participants may need some time to reflect on how much they regret a phone use session, which makes in-the-moment assessment potentially less reliable.

3.2.3 Two Follow-up Interviews. On two of the seven days during the week of data collection, participants engaged in a one-hour interview with a member of the research team. We scheduled the two interviews such that each participant could reflect on their phone use both on a weekday and on the weekend, as their behavior might differ between weekdays and weekends due to different daily routines. During the interview, the participants used screen sharing to go through all the sessions captured by the app from that day or the previous day, and were asked to reflect on why they regretted some experiences and valued others. As participants went through each session, the interviewer asked questions to understand what led to regret (or satisfaction), such as: “*Can you think out loud to*

Table 1: “Intended Use” categories shown in the ESM prompt and their definitions, based on Lukoff et al.’s application of Uses & Gratifications Theory [37].

Category	Abbreviation	Definition
No specific goal	<i>No Specific Goal</i>	You use your phone habitually or without a clear goal, to browse, explore, or pass the time.
Get things done or self-improve	<i>Productivity</i>	You want to achieve specific tasks and engage in activities focused on productivity or personal development.
Get information	<i>Information</i>	You want to acquire knowledge or stay updated on various topics.
Communicate or interact with others	<i>Communication</i>	You want to connect and interact with people through messaging, social networking, or video calls.
Be entertained or have fun	<i>Entertainment</i>	You want to relax and enjoy activities such as watching movies, listening to music, or playing video games.
Browse social media	<i>Social</i>	You want to consume content on social media without actively engaging, such as scrolling on feeds.
I am not sure	N/A (excluded)	You do not know how your motivation for phone use fits into these categories, you are using your phone accidentally, or you do not have time to answer this survey question.

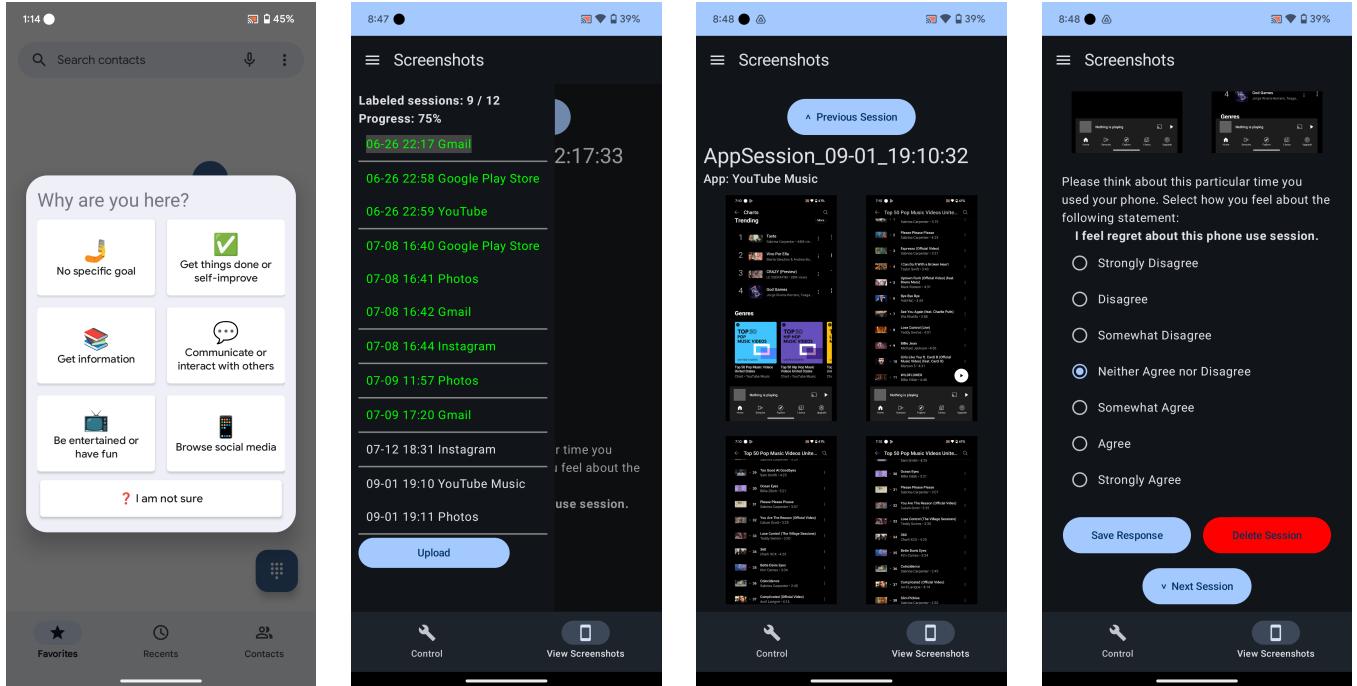


Figure 1: Screenshots of the data-collection app. The first screenshot shows the ESM prompt asking the user about their intended use when entering each app. The second screenshot shows the interface for selecting app sessions to share with the research team and uploading them (green sessions indicate sessions where the participant finished answering the regret survey question). The third screenshot shows the app session page, which presents screenshots for each captured session, along with the starting time and app name. The last screenshot shows the regret survey question, presented at the end of each app session page.

help me understand exactly how you feel about the session and the reason why you felt (or did not feel) regret about the session?" The interviewer also asked about participants' intention for app use when reviewing some sessions, such as: "*What motivated you to use this app at the time?*" At the end of each follow-up interview, the interviewer asked participants to reflect holistically about what experience they found regrettable. For example, we asked: "*Can you summarize how you answered the regret survey question? What factors did you consider when determining whether a situation is regretful or not?*" All of the follow-up interviews were conducted via Zoom.

3.3 Data Analysis

3.3.1 Interview data. Since our initial interviews were brief orientations, we focused our qualitative data analysis on the 34 follow-up interviews (two follow-up interviews for each participant). To conduct the qualitative data analysis, we used the collaborative qualitative data analysis tool Dovetail¹, which has been used in past HCI studies [6, 32, 57] and is HIPAA, GDPR, and CCPA compliant [16], indicating robust privacy protection. All of the interview data uploaded to Dovetail was anonymized. We first used the automated transcription service in Dovetail to turn the 34 audio recordings into text. Then following the thematic analysis approach [11], two researchers first separately reviewed the transcripts and came up with initial codes in an inductive-deductive manner, generating codes both based on the data and on our research questions. Then the two researchers met and collaboratively constructed codes and applied them to all the transcripts. Our codebook is shared in Appendix B. After coding the data, one researcher pulled out important quotes from the transcripts. For each quote that appears in the paper, we use "PX, FX" to denote the participant ID and interview ID (1 being the first follow-up interview, and 2 being the second).

3.3.2 ESM and survey data. We collected data from 1,631 screen sessions and 3,946 app sessions in total, which add up to approximately 183 hours of phone use. For each app session, we had the associated app name, starting time, duration, intended use (Table 1), and regret (in a seven-point Likert scale). For our quantitative data analysis, we used analysis of variance based on mixed ordinal logistic regression, as our response variable (regret) is an ordinal variable.

3.3.3 Screenshot data. We collected 119,373 screenshots in total. We focus our screenshot data analysis on social media apps, since social media usage is one of the most common cases of problematic smartphone use [37], and these apps have mixed features such as recommendation, search, direct messaging, requiring a nuanced understanding [10], which screenshots can provide. We chose social media apps that had been used more than 50 times and at least by 3 participants from our dataset. We also excluded apps that do not have a "feed" feature and only support direct messaging (e.g., Messenger, WhatsApp). Initially, we obtained a list of 6 apps including: *Instagram*, *Facebook*, *X*, *Snapchat*, *TikTok*, and *Reddit* (in the order of their occurrences in our dataset from high to low), totaling 38,017 screenshots.

We constructed seven categories of social media activity based on common behavior of social media users, and we considered both user action and the source of content, an approach used in prior work [10, 46]. The seven categories include: **Active Communication**, **Active Search**, **Consuming Recommendation-Based Content**, **Consuming Subscription-Based Content** (this includes when the user is viewing content from accounts they follow), **Consuming Content Shared by Others**, **Viewing Comments or Discussion Thread**, and **None of the Above** (See Tables 2 and 3 for detailed definitions of these categories and example screenshots).

We used OpenAI's multimodal large language model, gpt-4o, which showed exceptional ability in visual tasks [42], to help code the screenshots. Before applying screenshot analysis to the entire dataset using large language model, we assessed the consistency between human coders and gpt-4o to assess if the model could successfully categorize the screenshots based on our coding scheme. We randomly picked 500 screenshots from the dataset and let two researchers categorize them (we also included the previous four screenshots as context). The interrater reliability between the two human raters, measured in Cohen's Kappa [12], was 0.82 (with a percentage agreement of 86%), indicating almost perfect agreement. The two raters then discussed to resolve conflicts, which led to "human consensus" categories, which we then compared with categories generated by gpt-4o using text descriptions of the screenshots with zero-shot and chain-of-thought prompting [33] (detailed process below). The interrater reliability between gpt-4o and human consensus is 0.72 (with a percentage agreement of 78%), indicating substantial agreement. When breaking our results down by app (see Table 4), we found that X had the lowest human-human consistency and the lowest raw percentage consistency among all the apps. This aligned with our two raters' observation. We found that on X, the recommendation feed ("For You") and the following feed ("Following") often look identical when the top bar is hidden, making it challenging to detect whether the user is viewing recommendation-based or subscription-based content. Since we cannot reliably detect the difference between these two activities on X, we dropped X for our data analysis. With the rest of the apps, our classification showed a human-AI consistency of 0.74 measured in Cohen's Kappa and 79% measured in raw percentage agreement, indicating substantial agreement. To further assess the validity of LLM labeling, we plotted the confusion matrix (see Figure 2) and examined the precision, recall, and F1 score for each class (Table 5). Results showed that the weighted average of F1 scores achieved nearly 0.8, with most of the classes showing an F1 score larger than 0.7 (including *Communication*, *Search*, *View_Recommendation*, *View_Subscription* and *View_Comments*). However, the precision, recall, and F1 score for the class *View_Shared* are particularly low (0.2). Out of its 6 instances that appeared in the test set, 4 were misclassified as *View_Subscription*. 11 instances of *View_Subscription* were misclassified as *View_Shared*. This might be due to the fact that, it is easy to confuse content *shared* by a subscribed account with content *posted* by a subscribed account (refer to Tables 2 and 3 for a comparison). Since this category is rare (accounting for only 1% of the data in the test set), it likely will have a minimal impact on our analysis. We still keep this category in the analysis, but caution readers that the data associated with this small category

¹<https://dovetail.com/role/researcher/>

Table 2: Social media activity categories, the corresponding abbreviation used in the rest of the paper, definition used for human and AI coding, and one example screenshot (all private messages and identifiable information have been redacted).

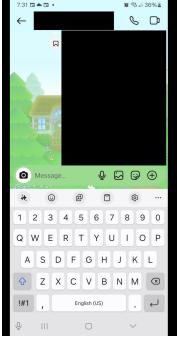
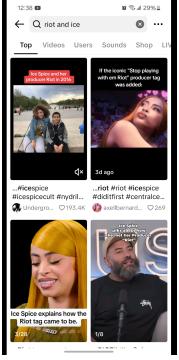
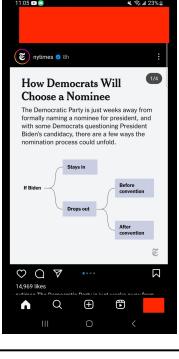
Category	Abbreviation	Definition	Example Screenshot
Active Communication	<i>Communication</i>	The screenshot includes the presence of a private messaging interface, suggesting the user is actively communicating with specific individuals or groups.	
Active Search	<i>Search</i>	The screenshot suggests the user is using the search feature to find specific information, content, articles, or items or is consuming content they found through active searching.	
Consuming Recommendation-Based Content	<i>View_Recommendation</i>	The screenshot shows explicit indicators of recommendation-based content in the user's feed, such as a "For You" tab, "Suggested Post" labels, or buttons like "Follow" and "Join" which allow users to subscribe to new content.	
Consuming Subscription-Based Content	<i>View_Subscription</i>	The screenshot shows content already followed or subscribed to by the user in their feed. The screenshot should include indicators such as an active "Following" or "Subscription" tab of the app, or signs suggesting that the user has already followed the content poster, such as the absence of buttons next to the content poster or community to follow, join, or subscribe in a feed interface.	

Table 3: (Continued) Social media activity categories, the corresponding abbreviation used in the rest of the paper, definition used for human and AI coding, and one example screenshot (all private messages and identifiable information have been redacted).

Category	Abbreviation	Definition	Example Screenshot
Consuming Content Shared by Others	<i>View_Shared</i>	The screenshot suggests the user is viewing content shared or reposted by someone they followed or opened from a private conversation.	
Viewing Comments or Discussion Thread	<i>View_Comments</i>	The screenshot suggests the user is viewing the comment section or discussion thread of a social media post.	
None of the Above	<i>Other</i>	When the user opens a link to a website, when the user sees sponsored content or ad, when it is unclear if user is viewing content posted by someone they followed or recommended to them, or when seeing these screens: home screen, notification screen, a black, dimmed, or blank screen, a screen showing a survey prompt.	

are likely erroneous. Our final social media use dataset includes 664 app sessions and 34,313 screenshots, adding up to approximately 50 hours of social media use.

To automatically code these screenshots, we first turned all screenshots into textual descriptions using gpt-4o, as we observed that reasoning directly on raw images could result in poorer performance compared to reasoning on textual descriptions of screenshots, particularly when multiple screenshots are involved. We asked the model to “*describe all of the visual elements (including all of the UI components and content) and the user’s activity in the screenshot in great detail*.” Importantly, when asking the model to generate descriptions for each screenshot, we also provided the

previous screenshot in the same app session (if applicable) as context and asked the model to describe “*how the user transitioned from the first screenshot to the second*” to help the model identify transitions and relationships between screenshots. The full prompt is included in Appendix A.1. All screenshots were sent to gpt-4o with 1/16 of the original resolution (1/4 of the original width and height respectively), which provided a good balance of cost/speed and fidelity.

Then, using the textual descriptions generated by gpt-4o, we asked the model to categorize each screenshot into one of seven categories of social media activity using structured outputs [44]. We employed zero-shot prompting with chain-of-thought reasoning,

Table 4: Human-human and human-AI interrater reliability by app based on Cohen’s Kappa and raw percentage agreement. X (formerly Twitter) was later dropped due to having the lowest human-human consistency caused by ambiguity of the interface. The row in bold represents reliability for the remaining apps.

App	% in the Test Set	Human-Human Consistency		Human-AI Consistency	
		Cohen’s Kappa	Raw %	Cohen’s Kappa	Raw %
All	100.0%	.816	85.6%	.715	77.6%
All (-X)	86.8%	.842	87.8%	.736	79.5%
Facebook	33.8%	.845	88.8%	.664	75.1%
Instagram	31.6%	.814	86.7%	.757	82.3%
TikTok	15.6%	.792	85.9%	.749	83.3%
X	13.2%	.648	71.2%	.575	65.2%
Reddit	4.8%	.830	91.7%	.528	79.2%
Snapchat	1.0%	1.000	100.0%	.545	80.0%

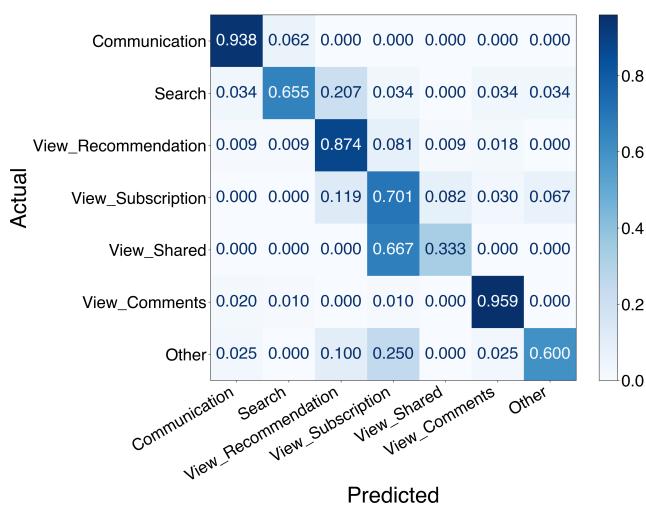


Figure 2: Normalized confusion matrix. Each row represents one actual class, and each column represents one predicted class. The number in each cell represents the proportion of a predicted class for a given actual class.

which have been shown to elicit multi-step reasoning and help reach correct answers [33]. When providing the textual descriptions, we also included the descriptions of the previous four screenshots in the same app session (if applicable) as context, since some critical details used for categorization may be hidden when looking at only one screenshot. Note that the four previous screenshots were also provided when human coders were coding the screenshots used for assessing consistency. We provided the definition of each of the

Table 5: Evaluation metrics for each class (representing one type of social media activity), including precision, recall, and f1 score. “Support” indicates the number of actual occurrences of the class in the test set. Micro (accuracy), macro, and weighted averages are also reported.

Class	Precision	Recall	F1 Score	Support
<i>Communication</i>	.750	.938	.833	16
<i>Search</i>	.864	.655	.745	29
<i>View_Recommendation</i>	.789	.874	.829	111
<i>View_Subscription</i>	.790	.701	.743	134
<i>View_Shared</i>	.143	.333	.200	6
<i>View_Comments</i>	.922	.959	.940	98
<i>Other</i>	.706	.600	.649	40
Accuracy	.795	.795	.795	434
Macro Average	.709	.723	.706	434
Weighted Average	.806	.795	.797	434

social media categories to the model, including visible indicators of the categories (such as a highlighted “For You” tab for Consuming Recommendation-Based Content), similar to how a human coder would classify these screenshots. We then complemented this with a few general rules, such as asking the model to pay attention to the previous screenshots to infer user’s current activity. The full prompt is included in Appendix A.2. See Figure 3 for a graphical example of the MLLM coding process.

3.4 Ethical Considerations

We carefully considered how to protect participants’ privacy, given the invasive and potentially sensitive nature of our data collection. We implemented several mechanisms to give participants control over their data. First, we ensured that screenshots were only stored locally and shared with the research team only when the participant actively elected to upload them. All of the uploaded screenshots were then stored in a private Amazon S3² bucket, which automatically applies server-side encryption [2] and can only be accessed by using the administrator credentials of the bucket owner. Second, participants could view all of the screenshots captured by the app and delete an entire session of screenshots at any time. They were also asked to review and delete any screenshots they did not want us to see when reflecting on the data and filling in daily questionnaires at the end of each day. Third, if a participant anticipated that they would be doing something private on their phone (such as checking their bank account), they were allowed to temporarily stop data collection and resume participating later.

During the initial interview and before the participants explicitly consented to join the study, we were fully transparent about when and how passive screenshots would be taken during the study, and how the privacy protection mechanisms mentioned above work. We also explained to participants that only researchers in the team

²<https://docs.aws.amazon.com/s3/>

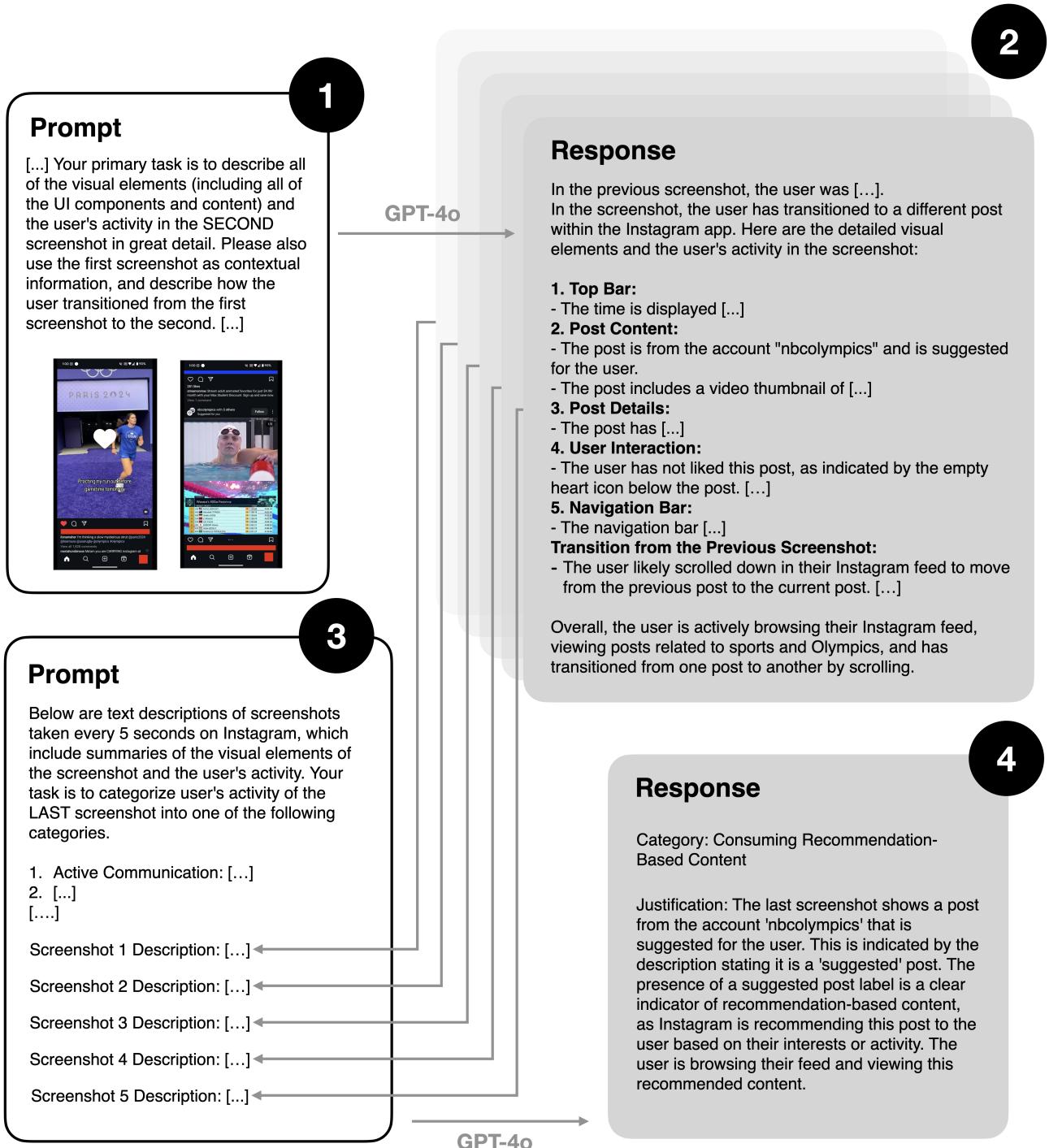


Figure 3: The process of using GPT-4o to code screenshots. The associated activity category for each screenshot was obtained using the steps shown in the diagram. (1) An image-to-text prompt was first constructed, which included instructions to describe all visual elements on the screenshot and how the user transitioned from the previous screenshot in the same session (if one exists). (2) The model sent back detailed text description of the screenshot. (3) Combining the text descriptions of the previous four screenshots, we constructed the second text-only prompt which asked the model to categorize the user's activity. (4) The model sent back the category it identified and its justification for choosing that category.

will be able to access the raw screenshot data, and that we may use the screenshots to train our own machine learning models and use OpenAI's GPT models to analyze the screenshots. We cited OpenAI's policy which states that they do not store image data uploaded via their API or use the data for training their model [45]. To ensure that the participants fully understood how the screenshot data would be collected and used, we also asked them if they needed any clarification or had any concerns regarding data privacy before having them join the study. Since our study required participants to constantly review and reflect on their phone use behavior, which can cause emotional stress, we emphasized that their participation was voluntary, and that they could drop out of the study at any time if they did not feel comfortable, in which case they would still receive pro-rated compensation for any portion of the study they had completed. In terms of data sharing with OpenAI, we eventually used 34,312 raw screenshots encoded in the base64 format in our requests, along with the associated app name. No other information about the participants was shared with OpenAI. This study was reviewed by our institutional review board (IRB) and deemed exempt.

4 Results

4.1 How Regret Varies by User Intention

To address RQ1, we began by analyzing all 3,946 sampled app sessions of general phone use to explore the relationship between regret and user intention. These sessions included any app usage, except for Settings and the data-collection app itself.

4.1.1 Regret varies based on intended use. We found that the user's stated intention for opening an app predicted the extent to which they would later regret doing so (see Figure 4). An analysis of variance based on mixed ordinal logistic regression indicated a statistically significant effect of *Intended Use* on *Regret* ($p < .001$). Pairwise comparisons (also shown in Figure 4) indicated that participants were significantly more likely to regret their use when it was motivated by *No Specific Goal* than when it was motivated by any other intention (i.e., any of *Communication* ($p < .001$), *Entertainment* ($p < .001$), *Information* ($p < .001$), *Productivity* ($p < .001$), and *Social* ($p < .01$)). Post hoc comparisons also revealed that the *Regret* scores for each pair of *Intended Use* except between *Entertainment* and *Information* were significantly different ($p < .001$).

Overall, these results indicate that participants were most likely to regret phone use when they went to an app without a specific goal, followed by when they wanted to browse social media. Interestingly, participants felt similarly regretful when they were intentionally seeking either entertainment or information, with each being less regrettable than browsing social media. Participants tended to express less regret when their intention in using an app was to communicate or interact with others, and least regretful when they wanted to get things done or self-improve.

4.1.2 Non-intentional social media use is more regretful than intentional social media use. Given that participants were most likely to regret their phone use after using an app without having a goal, we examined which apps they used when they reported not having a goal. Figure 5 shows the top 10 apps with the highest proportion of user-reported intention being *No Specific Goal*, after removing

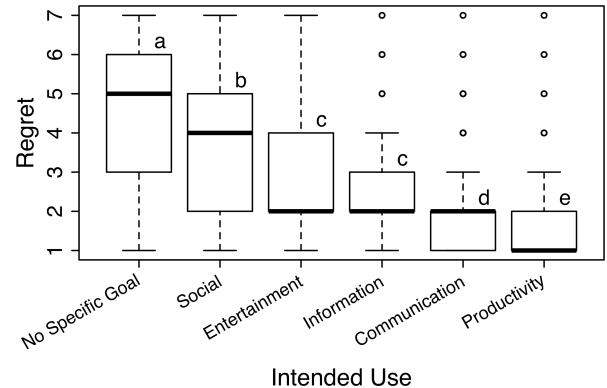


Figure 4: Box plots of *Regret* by *Intended Use*. *Regret* values are responses to the prompt: “I feel regret about this phone use session” (1=Strongly Disagree, 7=Strongly Agree). Letters (a, b, c, ...) indicate pairs that are NOT significantly different from each other. For example, we did not find a statistically significant difference between *Entertainment* and *Information*, but did find a statistically significant difference between *No Specific Goal* and *Social*. For pairs with statistically significant difference, $p < .001$ for all of the pairs except between *No Specific Goal* and *Social*, where $p < .01$.

apps that have been used less than 20 times overall (due to their overall low frequency in the dataset). Eight of these apps have some form of social media, which aligns with participants' perception, as expressed in interviews, that they usually go to social media apps when habitually checking their phone. For example, P8 said: “[Instagram] is one of the apps that I just open as a habit for no reason. And I don't even know why.” (P8, F2)

Although participants were more likely to regret sessions motivated by no specific goal than they were to regret sessions where they intended to browse social media (as described in 4.1.1), in both cases they predominantly used social media apps (see Figure 5 and Figure 6). Participants explained that they felt good about the times when they intentionally chose to browse social media for pleasure, but they regretted instances where they mindlessly turned to it out of habit or as a knee-jerk avoidance mechanism because of procrastination:

“I'm gonna go on Facebook and check out social media, but [if] it's like, I'm giving myself a nice break, then I'm less likely to regret it than if I'm just doing it to zone out or to procrastinate an off computer test or an off phone task... [If] I'm doing it because I want to go on social media, then I seem less likely to regret it, than it's like if I didn't really have a specific purpose. I just want to go on my phone to, like, avoid doing other things.” (P7, F1)

Thus, both our qualitative and quantitative results suggest that non-intentional, habitual social media use was more regretful than intentional, deliberate social media use.

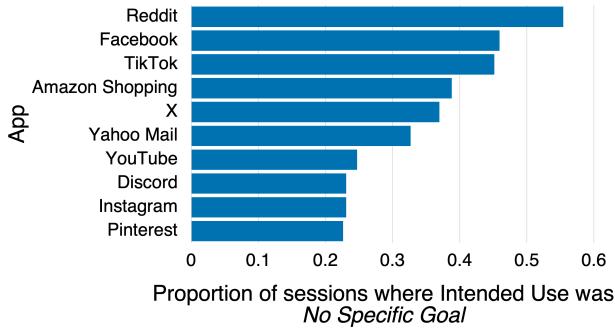


Figure 5: Top 10 Apps with the highest proportion of sessions where user-reported intended use was *No Specific Goal*, after filtering out apps that have been used less than 20 times overall.

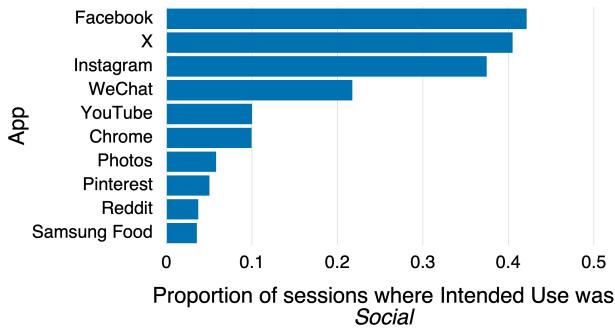


Figure 6: Top 10 Apps with the highest proportion of sessions where user-reported intended use was *Social*, after filtering out apps that have been used less than 20 times overall.

4.2 How Regret Varies by Type of Social Media Activity

To answer RQ2, we next examined participants' fine-grained interactions within social media apps, given that these were the apps they were most likely to regret using. As described in Section 3.3.3, we used a multimodal LLM to analyze all screenshots from five social media apps: Instagram, Facebook, Snapchat, TikTok, and Reddit (adding up to 664 app sessions and 34,313 screenshots), evaluating each screenshot for the “activity” it represents (one of seven possible choices).

4.2.1 Social media activity and regret. To understand how regret is related to different social media activities, we tested the relationship between these two variables. Using the most prevalent activity in each app session to represent the session, we conducted an analysis of variance based on mixed ordinal logistic regression, which indicated a statistically significant effect of *Activity* on *Regret* $p < .001$. Pairwise comparisons indicated that regret Likert scores for *View_Recommendation* vs. all the other activities except *View_Comments*, including *View_Subscription* ($p < .01$), *Other* ($p < .001$), *Search* ($p < .001$), and *Communication* ($p < .001$),

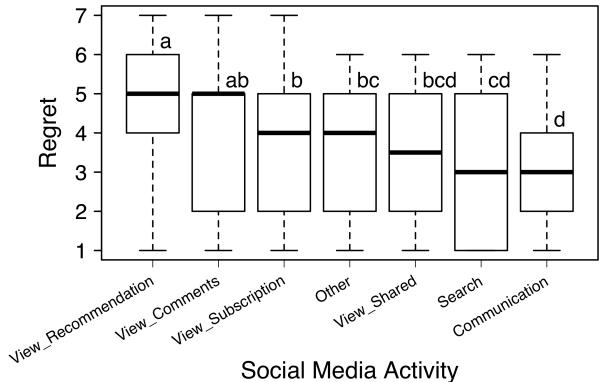


Figure 7: Box plots of *Regret* by *Social Media Activity*. *Regret* values are responses to the prompt: “I feel regret about this phone use session” (1=Strongly Disagree, 7=Strongly Agree). Letters (a, b, c, ...) indicate pairs that are NOT significantly different from each other. For example, we did not find a statistically significant difference between *View_Recommendation* and *View_Comments*, but did find a statistically significant difference between *View_Recommendation* and *View_Subscription*. The “Other” category may include anything that does not easily fall into any of the other categories, such as advertisements, notifications, settings menus, etc. Note that the LLM performance on classifying *View_Shared* was low (often confused with *View_Subscription*), so the regret associated with it may not be accurate.

were statistically significantly different, that the regret scores for *View_Comments* vs. *Communication* ($p < .001$) and *Search* ($p < .01$) were statistically significantly different, that the regret scores for *View_Subscription* vs. *Communication* ($p < .001$) and *Search* ($p < .05$) were statistically significantly different, and that the regret scores for *Other* vs. *Communication* ($p < .01$) were statistically significantly different. Figure 7 shows the box plots of regret scores for each activity category, along with results of post-hoc pairwise comparisons. These results indicate that viewing recommendation-based content and comments or discussion threads are the most regrettable experiences of social media for the participants, followed by viewing subscription-based content, while communication, and search are less regrettable.

These results align with what the participants shared in the follow-up interviews. Many participants said that not all social media usage was equally regrettable to them. Specifically, participants said that activities like direct messaging were meaningful and not regrettable at all because of the human connection they brought. For example, P7 said:

“One thing I noticed during this research study is I almost never regretted, like Messenger or phone call or like something where I’m interacting with another person that I know... It feels meaningful.” (P7, F2)

When it comes to passively consuming social media content, participants said that the source of the content tends to lead to different levels of regret. Specifically, some participants said that recommendation-based content was more regrettable than subscription-based content. For example, one participant described recommendation-based content as “random” and more regrettable than subscription-based content and content they intentionally searched for:

“So random content, like what Twitter gives me or Facebook. Like almost all of it was random. It wasn’t people I knew, right. So like that content, I end up regretting more than if it’s content of like people I know or a work thing, like an article I was reading. [...] Even content that I specifically look for, I regret a little less I think than the random stuff.” (P11, F1)

Another participant shared a similar perspective, adding that they had short attention span when looking at recommended videos:

“So it’s very random content. Like it’s always changing a lot... It’s not only changing videos, it’s changing content like styles or genres... My attention span is short and I’m not interacting with the videos and the content is very random. So I’m like, willingly engage, like staying in this place where the content isn’t like valuable.” (P9, F2)

Some participants mentioned that reading comments on social media, especially political ones, can be a negative experience to them: *“When I’m in the news, sometimes when I go into comment sections, then I’d be more likely to regret the content because especially with political topics, there’s so many obnoxious comments in the comment section.”* (P7, F2)

Thus, both of our quantitative and qualitative results show that there are systematic patterns linking specific social media activity to regret. Behaviors like consuming content that is algorithmically recommended are associated to higher regret and are perceived as less valuable and relevant by our participants, whereas they found active communication to be meaningful.

4.2.2 Visualizing social media activity over the duration of an app session. Given that people’s regret varies based on specific activities on social media, we sought to understand how those activities are distributed in different apps and over the duration of an app session. We adopted the approach of segmenting sessions based on time elapsed within every app session in [10] and created a timeline for each of the five social media apps we examined, with each showing the breakdown of users’ activity at each time chunk of the app session (see Figure 8). Each timeline revealed a distinctive behavioral signature in the way users engaged with the app. For example, on Instagram, participants tended to begin their usage session by reading and sending direct messages and viewing content from accounts they follow. After approximately one minute, the balance of activities began to shift, and participants increasingly viewed recommended content from accounts they do not follow. In contrast, on Reddit, participants tended to begin by browsing their feed (which consisted of a mix content from accounts they follow and other recommended content from accounts they do not follow) at first, and then dove into specific discussion threads. On

TikTok, participants mostly consumed recommended content from accounts they do not follow throughout the duration of the session.

These timelines show that different apps consist of different percentage of each type of social media content, and there are distinct patterns of how users move between regretful and less regretful activities, suggesting that the design of these social media apps can influence user’s tendency to engage in behaviors they would later regret. For example, the start of an Instagram session tends to be filled with low regret activities, such as direct messaging and viewing content from followed accounts, but users slowly shift into less valuable activities as their sessions went on. Reddit users spend the majority of their time reading other users’ comments, a choice we found to be highly regrettable. TikTok users consistently engaged with a high proportion of recommended content, the most regrettable behavior among our seven categories, from the moment they were on the app.

4.2.3 Factors that predict regret. To understand what factors can predict regret, we fit a cumulative logistic regression model. We modeled the ratio of each activity within a session (from 0 to 1) and *Intended Use* as fixed effects and *User ID* as a random effect. We added *Duration* and *App* as fixed effects to control for the effect of the time spent on each app session, as most participants reported that it is an important factor that influences their regret.

We found that *Duration* ($p < .001$), *Intended Use* ($p < .001$), *Ratio_View_Recommendation* ($p < .001$), *Ratio_View_Comments* ($p < .01$), and *Ratio_View_Subscription* ($p < .05$) all had a significant effect on *Regret*. These results show that the amount of time spent on an app, participant’s intention when entering the app, and the amount of time they viewed recommendation-based content, comments, and subscription-based content can predict their subsequent regret. Notably, we did not find a significant effect of *App* on *Regret* ($p = .058$), indicating that after controlling for intended use, time spent on app, and specific app activity, the app itself does not predict regret. These results align with our results in 4.1.1 and 4.2.1, and they further show that user intention and the amount of recommendation-based, subscription-based content and comment consumption are all individually important predictors of regret when holding other factors constant.

4.3 How Regret Is Influenced by Deviation from Communication to Browsing Social Media

Given that the two variables we examined above, intended use and social media activity, indicate what people want to do and what people actually do respectively, a natural hypothesis (connected to RQ3) that arises is that people may feel more regret when their actual behavior deviates from their original intention than when the behavior aligns with their intention, as suggested by prior work [10] and the theory on regret regulation, which states that intention-behavior inconsistency increases process regret [51, 70]. Indeed, participants mentioned that when they deviated from their initial purpose and stayed longer on the app, it made them feel regret: *“That’s when I put somewhat agree to regretting... because it starts off with good intentions with like, I’m just trying to entertain myself, and then it ends with like, okay, let’s hear a little bit longer than I should have.”* (P2, F1)

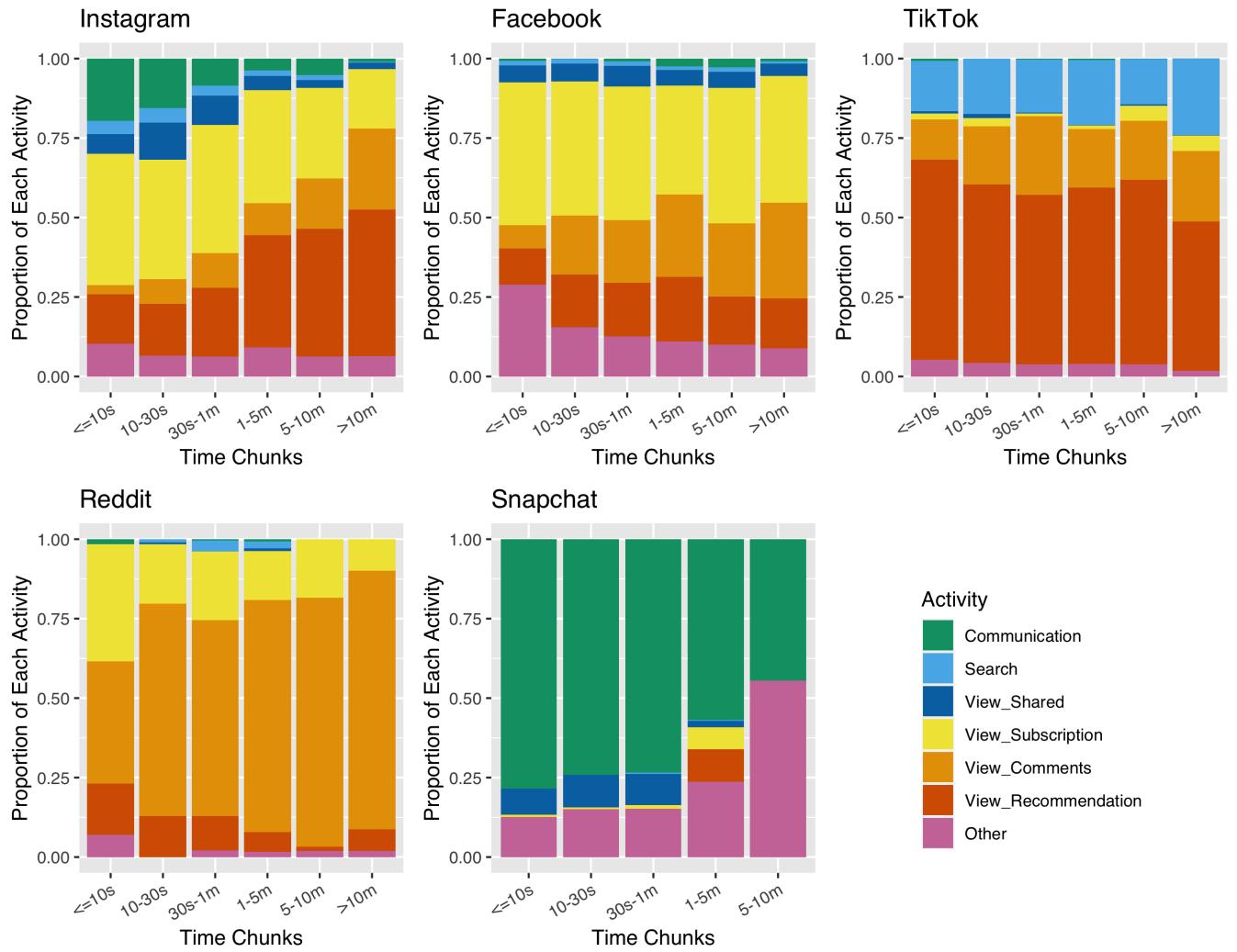


Figure 8: Activity in each social media app in time chunks. Each time chunk contains proportions of each activity in that specific time frame. Note that the LLM performance on classifying *View_Shared* was low (often confused with *View_Subscription*), so its representation in the timeline it may not be accurate.

However, what specific activity counts as “deviation” from an initial purpose can be challenging to pinpoint. For example, if participants reported wanting to get information, they may or may not want to get information from accounts they follow or through active search. To investigate this pattern of deviation, we chose a specific case where the mapping between intended use and activity is clear: when participants reported wanting to passively browse social media or actively communicate or interact with others. In the follow-up interviews, participants brought up that when they started with an initial purpose to communicate, sometimes they ended up browsing on their feed instead, a choice they later regretted. For example, when reviewing a session in a follow-up interview, P3 said: “I think I planned to communicate with someone but I didn’t. I lost my purpose.” (P3, F1) Similarly, P4, at the end of the last follow-up interview, summarized this as an indicator of regret: “Even if I

spend a session on Instagram and I do DM at the beginning of that session, like that’s a good indicator that I’m not gonna regret it. But then if I only do that for like 30 seconds and then I spend half an hour on the app, then I’m much more likely to regret it.” (P4, F2)

To quantify this deviation, we first visualized the proportion of screenshots associated with each activity within each intended use (see Figure 9). In sessions where users started with an intention to communicate or interact with others (the first bar in Figure 9), 37.3% of the screenshots were labeled as *Communication* by LLM, the rest included *View_Recommendation* (20.5%), *View_Subscription* (13.3%), *Other* (11.6%), *View_Comments* (7.4%), *View_Shared* (6.7%, although this might be inaccurate due to mislabeling), and *Search* (3.1%). In sessions where users started with an intention to browse social media (the last bar in Figure 9), 36.8% were *View_Subscription*, 29.6% were *View_Recommendation*, 15.7% were *View_Comments*,

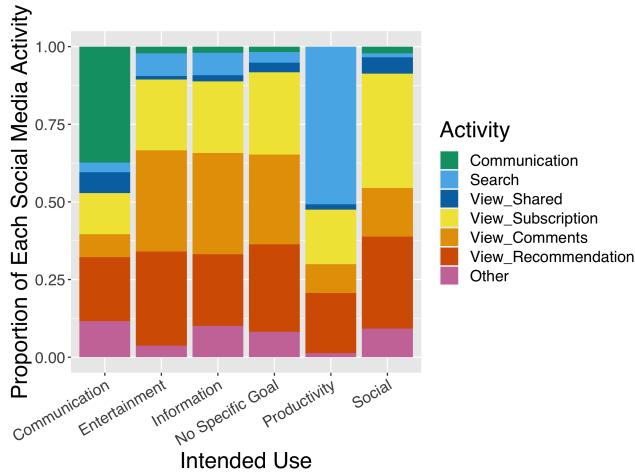


Figure 9: Proportion of social media activity for each intended use (which participants provided *prior to* entering every social media app). Note that the LLM performance on classifying *View_Shared* was low (often confused with *View_Subscription*), so the proportion associated with it may not be accurate.

9.2% were *Other*, 5.2% were *View_Shared*, 2.2% were *Communication*, and 1.3% were *Search*. These results suggest that while participants frequently deviated to non-communication behavior when having an intention to communicate (over 60% of the time), they rarely started with an intention to browse social media but ended up communicating (around 2% of the time).

To examine how deviation from communication to browsing social media may be related to regret, we set up our data analysis in the following way. We extracted three groups of sessions from our dataset. The first group includes sessions where the reported intended use was *Communication*, and the only activity from that session was *Communication* (i.e., LLM identified direct communication from all the screenshots). The second group includes sessions where the reported intended use was *Communication*, but activities other than direct communication were present (i.e., the user deviated from their intention to communicate). As a comparison, we also added a third group, which includes sessions where the intention was to browse social media. Figure 10 shows box plots of *Regret* for these three groups. These three groups add up to 317 app sessions, with each group having 46, 90, 181 sessions respectively. We found a statistically significant difference of *Regret* between the three groups ($p < .001$). Pairwise comparisons indicated that the *Communication* and *Communication Deviated to Social Media* groups were statistically significantly different ($p < .05$), and that both groups were statistically significantly different from *Social Media* ($p < .001$). The median regret scores for *Communication* and *Communication Deviated to Social Media* are both 3, whereas the median regret score for *Social Media* is 5. As another measure of central tendency, average regret scores for each group are 2.87, 3.02, and 3.95.

These results indicate that participants tended to feel just slightly more regretful when they deviated from active communication to

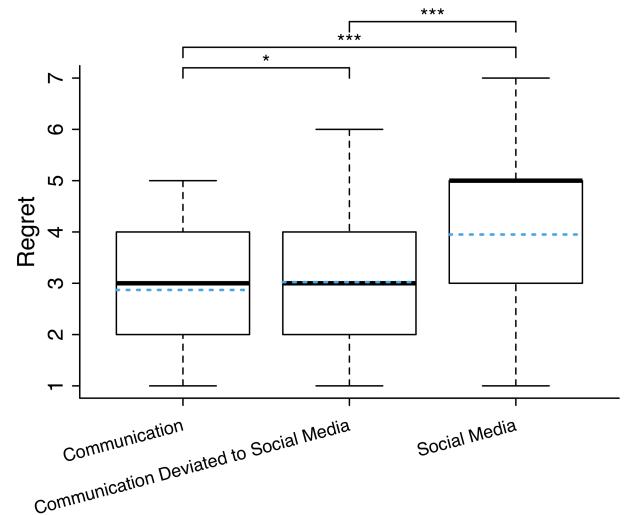


Figure 10: Box plots of Regret for three groups. The first group, labeled as “Communication,” includes sessions where the user intention was to communicate or interact with others, and the subsequent activity (from the screenshots) shows communication only. The second group, labeled as “Communication Deviated to Social Media,” includes sessions where the user intention was to communicate or interact with others, but the subsequent activity includes things other than communication, suggesting the user has deviated from their intention to browsing social media. The last group, labeled as “Social Media,” includes sessions where the user intention was to browse social media. Significant levels of difference between each pair are indicated in the plot, where * means $p < .05$ and *** means $p < .001$. The blue dashed line in each box plot represents the mean for that group, and the black solid line represents the median.

passive browsing of social media. But regardless of whether they deviated, sessions where the user intention was to browse social media was much more regretful than sessions where the intention was to communicate or interact with others.

5 Discussion

5.1 Characterizing Regrettable Social Media Use

Through our mixed-methods analysis, our results robustly showed that people’s intentions when using an app is an important predictor of later regret. We consistently saw that non-intentional phone use is related to an increased sense of regret later on. This aligns with prior work suggesting that ritualistic use (i.e., where the user habitually uses their phone to pass the time) is associated with a lower sense of meaning [37]. This effect persists even after holding other factors constant: what app the user is using, the specific activity they engage in, and how much time they spend on the app. Theories of regret explain that people regret actions that are less justifiable and those that stem from poor decision-making processes [24, 70]. This may explain why people experience more

regret over phone-use sessions that are not motivated by specific user intentions. Picking up the phone for no specific reason is harder to justify than doing so in pursuit of a specific intention. Cho et al. argue that people regret their phone use when it reflects the impulsivity of choosing a small, short-term reward (e.g., browsing their phone) over a bigger one (e.g., engaging in productive activities) [10]. This can also explain why people regret moments of non-intentional phone use, as this non-intentional use is often characterized by a lack of purpose and habitual or reflexive behavior, in which the user may be prone to impulsiveness. Given the importance of the user's intention in shaping their regret, what remains to be understood is the precise nature of such non-intentional use. For example, participants in our study said that they felt the urge to engage in non-intentional use when they were procrastinating and trying to avoid doing other activities. How other factors (such as the user's own mental state, their context, and design factors like push notifications) might trigger such non-intentional use is worth exploring in future work.

Our findings also highlight the benefits of examining the different ways that people engage with social media, rather than treating all social media use holistically. We found that people's regret can vary depending on the dominant activity in a social media usage session; participants were most likely to feel regret after viewing content that was algorithmically recommended to them or after browsing comments or discussion threads. And they felt least regretful after they had communicated with other people. These findings contribute to the ongoing debate on how passive and active social media use contributes to well-being [58], providing evidence that active communication (through direct messaging) is associated with lower negative affect than other activities.

Our findings can be interpreted both in terms of outcome regret and process regret. In terms of outcomes, Cho and colleagues explain that users seek four types of rewards from social media use, including: social rewards, informational rewards, personal interests, and entertainment rewards [10]. Activities like searching for information and direct messaging may offer particularly high informational rewards and social rewards respectively, and participants described these activities as more "meaningful." This data suggests they tend to induce less outcome regret than other activities.

In considering process regret, it is worth noting that when recommended content is displayed to users by a social media algorithm (without users explicitly choosing to view this content), this process can feel less justifiable, which induces more process regret. Indeed, in the interviews, participants described such content as "random," indicating that they often run into recommended content in an arbitrary, unplanned way, without a conscious or justified reason stemming from a well-considered decision-making process. These two independent sources of regret may jointly contribute to the higher regret associated with viewing algorithmically recommended content and comments, and the lower regret associated with active search and communication.

Past work has found through interviews that when users "sidetrack," deviating from their original intentions, they experience more regret about their usage [10]. This is supported by Zeelenberg and Pieters' work on regret regulation, which shows that intention-behavior inconsistency increases regret [51, 70]. In this study, we

examined this phenomenon within the context of a specific usage case, where users' behavior deviates from their intention to communicate or interact with people. We characterized the prevalence of such deviation, finding that when participants intended to communicate with other people, 60% of the time they deviated from this intention and browsed social media. Although this deviation was significantly associated with increased regret relative to non-deviation sessions, this increase was slight. And these sessions were far less regretful than sessions where users set out to browse social media in the first place. One potential explanation for this result is that if people start with an intention they judge positively (such as communicating with a friend), and they go on to fulfill this intention, they may feel satisfied with their experience overall, even if they later deviate to passive browsing. As noted earlier, we chose a specific form of deviation in this study, and future work can explore alternative forms to further model the relationship between intention-activity deviation and regret (e.g., by analyzing how activity changes and deviates throughout an app usage session).

5.2 Implications for Design and Policy

We encountered systematic patterns linking particular behaviors with later feelings of regret (see Figure 7) and linking the design of an app with particular behaviors (see Figure 8). For example, Instagram users began their sessions by viewing content from people they follow but were gradually funneled into viewing algorithmically recommended content from unknown accounts; Reddit users gradually descended into comment black holes; TikTok users engaged with mostly recommended content the moment they landed on the app. All of these patterns reflect a shift away from behaviors users value and toward behaviors they do not.

These findings have important implications for regulating attention-economy designs. Regulation of the technology industry strives to protect users from harm, including deceptive and manipulative practices that encourage users to act against their own best interests [41]. Historically, this work has targeted financial and data-privacy harms (e.g., [4, 38]), identifying, for example, manipulative designs that encourage users to make purchases they do not need or intend to make. A growing body of academic work has called for the investigation of attentional harm [36] and shown that attention-capture designs are widespread [40]. However, demonstrating manipulation and harm has been comparatively challenging with respect to users' engagement, because the fact that a user chooses to engage is argued to be a reflection of their preferences [53]. Our results provide continuing evidence that users' engagement decisions are not always a reflection of their intrinsic desires, but instead reflect a lack of justifiability that increases regret and goes against their best interests.

Further, our findings point to a clear link between (process) regret and engagement decisions that are app-driven (see Figure 11). When participants' decision to initiate engagement came from an intrinsic goal that they were able to articulate, they were far less likely to regret their use than when they could not come up with a reason for why they were there. And once engaged, they were far less likely to regret staying engaged when they looked at content and messages from people they know and have chosen to follow; they were much

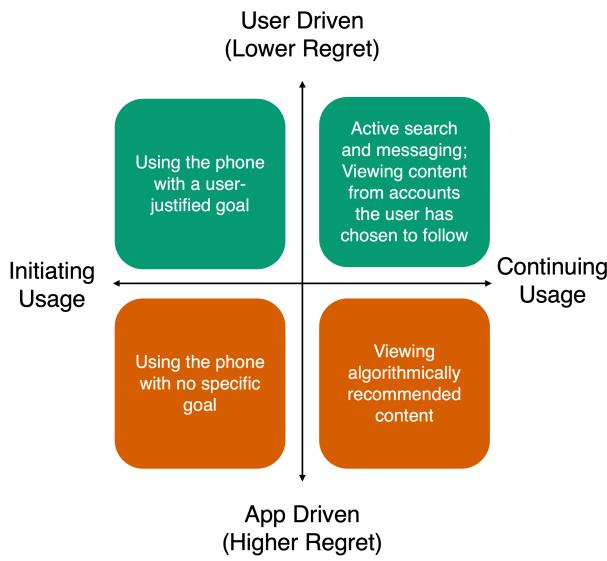


Figure 11: Regret matrix. We divide regretful behavior along two axes: app or user driven and the stage of usage (initiating or continuing usage). The two green blocks indicate experiences associated with lower regret, while the two orange blocks indicate experiences associated with higher regret.

more likely to regret staying engaged when the app pushed content of its own choosing. These findings have important implications for social media apps. To maximize users' well-being and reduce aversive feelings such as regret, app designers should not only consider the outcome of a user's engagement (for example, whether they will value the content being recommended to them), but also aim to reduce process regret that stems from unjustified, unplanned, and unconscious use. For example, apps could explicitly assess the user's intention upon app entry, adjust the interface to respect that intention, and avoid interspersing features and content in a way that may distract users and lead to sidetracking. The SwitchTube system, which introduced the notion of "adaptive commitment interface," is an example of such an experience, where the interface aligns with the user's goal [36]. These findings also point to a need for tools and benchmarks that measure a user's intention as they engage with an interface. If users consistently lose the thread of their intention, the app in question could benefit from redesign and regulation. Regulations that require companies to assess users' intentionality and to remove features that undermine it have the potential to reduce attentional harm and designs that funnel users into behaviors they regret.

5.3 Using Screenshots and Multimodal Large Language Model to Understand Phone-Use Behavior

In this study, we introduced a novel method to analyze fine-grained phone-use behavior through passive screenshot collection and automatic analysis using an MLLM. We showed that this method

achieved a substantial agreement (a Cohen's Kappa of 0.736) with human consensus labels, an accuracy of 79.5%, and a weighted average F1 score of 79.7%. Compared to previous methods that similarly analyzed fine-grained behaviors on social media platforms using the Android Accessibility API (e.g., [10, 46]), our approach has a few advantages. First, our approach does not require access to an app's UI structure during data collection and therefore does not require a strict rule-based mechanism to determine the feature that the user is interacting with. Second, our approach is more robust against app updates and can be applied to a wide range of apps more easily by simply changing the prompt to LLM. However, one limitation of our approach compared to using the Accessibility API is that LLM performance can degrade when certain categories share similar visual features with one another. For example, in this study, the LLM classification performance for viewing shared content from friends was much lower than for other categories. In this study, we employed zero-shot chain-of-thought prompting when using an MLLM. It would be valuable for future work to explore other ways to leverage MLLMs to improve the classification performance, potentially through few-shot chain-of-thought prompting [63] and vision fine-tuning [43].

For researchers interested in adopting our approach, we generally recommend keeping a short window of screenshots (e.g., around five screenshots) rather than one screenshot during categorization (both for human and LLM), since there might be details missing from just one screenshot. For example, if the user scrolled past the title of a post, it might be hard to tell whether the content is from someone they followed or the platform's recommendation algorithm. Having screenshots from seconds ago can be helpful in such cases. In this study, we used text descriptions of five screenshots (25 seconds) for our categorization task. In rare instances, we still found that we needed more screenshots to be able to classify correctly. For example, if a user opened shared content from a conversation, it might be necessary to trace back to earlier screenshots to see the origin of the content. However, we noticed that adding too many screenshots as context can degrade the current model's performance (e.g., more than 10 screenshots), although we recognize that this might change as models improve.

We note that collecting raw screenshots of phone use may present privacy concerns for participants. It would be valuable for future work to investigate privacy-preserving approaches to extract relevant information from real-time screenshots on-device. Past work has shown that on-device models can be trained to detect UI elements [72] and extract information from screenshots [34]. Similar approaches could potentially replace the first step of our screenshot analysis process (in Figure 3) and make the pipeline more privacy-preserving.

While our approach was developed for this empirical study, we believe that it can also be adopted for smartphone overuse intervention systems. Given that screenshots contain rich information about a user's activity, future work can explore combining screenshot analysis with passive sensing data to predict problematic phone use and deliver adaptive interventions tailored to users' in-the-moment contexts. For such a system to be deployed in the wild, privacy is an important consideration, and on-device approaches mentioned above might be a promising solution to alleviate privacy concerns. For a just-in-time intervention system, speed is also an important

factor to consider. In our study, we found that retaining 1/16 of the original resolution of the screenshots and taking screenshots every 5 seconds was an ideal combination, but a just-in-time system might have a higher threshold for speed and responsiveness. Future work can further explore variations of the screenshot collection interval and level of detail of screenshots to find what might be optimal for a real-time system.

5.4 Limitations

Our study has a few limitations. First, due to the complexity and novelty of our study procedure, which involves both passive screenshot collection and three interviews per participant, we slowly rolled out the study and included only 17 participants in the U.S. While our participants are relatively diverse in terms of age and racial identity, a large proportion of them are women. Future work can apply similar methods to a bigger and more balanced sample and in different contexts to examine whether the findings of this study still hold robustly across different population. Second, in the screenshot analysis, we only examined five social media apps. While social media apps are most commonly associated with problematic phone use, and the apps we chose are among the most popular in the U.S., we believe our screenshot analysis technique can be applied to a broader range of apps. For example, in Figure 5, Amazon shopping also showed a high proportion of non-intentional sessions, and online shopping apps also typically have a wide range of passive (e.g., feed for recommended products) and active features (e.g., search). Therefore, future work can investigate fine-grained phone-use behaviors beyond the scope of social media apps and investigate their impact on user's experience. Finally, since participants could delete entire sessions of screenshots due to privacy concerns, some of their behavior may not be reflected in our data.

6 Conclusion

In this study, we combined experience sampling, surveys, retrospective interviews, and passively collected screenshots analyzed via a multimodal large language model to examine how regret varies depending on user's intention and activity on social media. We found that regret of phone use varies by user intention, and our participants felt most regretful when using their phone without a specific goal. We also found that on social media, participants regretted more when they viewed content that was algorithmically generated and comments or discussion threads and less when they actively communicated with people and searched for information. Additionally, we found that over 60% of the time people deviated from their intention to communicate to social media browsing, which slightly increased their regret. We argue that designers and policy-makers who seek to improve user's experience and autonomy can assess and measure user's intentionality and reduce features and designs that might funnel users into regrettable use. Our screenshot analysis approach can also be adopted in just-in-time and fine-grained intervention systems that seek to reduce smartphone overuse.

Acknowledgments

Special thanks to all the participants for completing the intensive data-collection process and contributing their insights to this study. We also thank Anind Dey, Je-Wei Hsu, Shahan Ali Memon, Adiba

Orzikulova, and Nic Weber for their insightful suggestions at various stages of this paper. Additionally, we thank the anonymous reviewers for their detailed and thoughtful feedback, which helped us refine this work. Alexis Hiniker is a special government employee for the Federal Trade Commission. The content expressed in this manuscript does not reflect the views of the Commission or any of the Commissioners.

References

- [1] Jens Allaert, Rudi De Raedt, and Marie-Anne Vanderhasselt. 2019. When choosing means losing: Regret enhances repetitive negative thinking in high brooders. *Journal of Experimental Social Psychology* 85 (2019), 103850. <https://doi.org/10.1016/j.jesp.2019.103850>
- [2] Amazon Web Services. 2024. *Using Encryption for Amazon S3*. <https://docs.aws.amazon.com/AmazonS3/latest/userguide/UsingEncryption.html> Accessed: 2024-12-04.
- [3] David E. Bell. 1982. Regret in Decision Making under Uncertainty. *Operations Research* 30, 5 (1982), 961–981. <https://doi.org/10.1287/opre.30.5.961> arXiv:<https://doi.org/10.1287/opre.30.5.961>
- [4] Christoph Bösch, Benjamin Erb, Frank Kargl, Henning Kopp, and Stefan Pfattheicher. 2016. Tales from the dark side: Privacy dark strategies and privacy dark patterns. *Proceedings on Privacy Enhancing Technologies* (2016). <https://doi.org/10.1515/popets-2016-0038>
- [5] Bianca Bosker. 2016. The Binge Breaker. *The Atlantic* (November 2016). <https://www.theatlantic.com/magazine/archive/2016/11/the-binge-breaker/501122/>
- [6] Kyle Boyd, Patrick McAllister, Maurice Mulvenna, Raymond Bond, Hui Wang, Ivor Spence, Guanfeng Wu, and Abbas Haider. 2023. Designing Multimodal Video Search by Examples (MVSE) user interfaces: UX requirements elicitation and insights from semi-structured interviews. In *Proceedings of the European Conference on Cognitive Ergonomics 2023* (Swansea, United Kingdom) (ECCE '23). Association for Computing Machinery, New York, NY, USA, Article 4, 8 pages. <https://doi.org/10.1145/3605655.3605665>
- [7] Jamie C. Brehaut, Annette M. O'Connor, Timothy J. Wood, Thomas F. Hack, Laura Siminoff, Elisa Gordon, and Deb Feldman-Stewart. 2003. Validation of a Decision Regret Scale. *Medical Decision Making* 23, 4 (2003), 281–292. <https://doi.org/10.1177/0272989X03256005> PMID: 12926578.
- [8] Michael Chan. 2015. Mobile phones and the good life: Examining the relationships among mobile use, social capital and subjective well-being. *New Media & Society* 17, 1 (2015), 96–113. <https://doi.org/10.1177/1461444813516836>
- [9] Yu-Chun Chen, Yu-Jen Lee, Kuei-Chun Kao, Jie Tsai, En-Chi Liang, Wei-Chen Chiu, Faye Shih, and Yung-Ju Chang. 2023. Are You Killing Time? Predicting Smartphone Users' Time-killing Moments via Fusion of Smartphone Sensor Data and Screenshots. 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 647, 19 pages. <https://doi.org/10.1145/3544548.3580689>
- [10] Hyunsung Cho, DaEun Choi, Donghwi Kim, Wan Ju Kang, Eun Kyong Choe, and Sung-Ju Lee. 2021. Reflect, not Regret: Understanding Regretful Smartphone Use with App Feature-Level Analysis. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 456 (oct 2021), 36 pages. <https://doi.org/10.1145/3479600>
- [11] Victoria Clarke and Virginia Braun. 2017. Thematic analysis. *The Journal of Positive Psychology* 12, 3 (2017), 297–298. <https://doi.org/10.1080/17439760.2016.1262613>
- [12] Jacob Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement* 20, 1 (1960), 37–46. <https://doi.org/10.1177/00131644600200104>
- [13] Mihaly Csikszentmihalyi, Reed Larson, and Suzanne Prescott. 1977. The ecology of adolescent activity and experience. *Journal of youth and adolescence* 6, 3 (1977), 281–294. <https://doi.org/10.1007/BF02138940>
- [14] Android Developers. 2024. Media Projection. <https://developer.android.com/media/grow/media-projection>. Accessed: 2024-09-10.
- [15] Tilman Dingler and Martin Pielot. 2015. I'll be there for you: Quantifying Attentiveness towards Mobile Messaging. In *Proceedings of the 17th International Conference on Human-Computer Interaction with Mobile Devices and Services*. 1–5.
- [16] Dovetail. 2024. Trust Center - Dovetail. <https://trust.dovetail.com/> Accessed: 2024-12-04.
- [17] Jon D. Elhai, Robert D. Dvorak, Jason C. Levine, and Brian J. Hall. 2017. Problematic smartphone use: A conceptual overview and systematic review of relations with anxiety and depression psychopathology. *Journal of Affective Disorders* 207 (2017), 251–259. <https://doi.org/10.1016/j.jad.2016.08.030>
- [18] European Parliament. 2023. New EU rules needed to address digital addiction. *European Parliament News* (2023). <https://www.europarl.europa.eu/news/it/press-releases/2023/03/20230320STO00001>

- room/20231208IPR15767/new-eu-rules-needed-to-address-digital-addiction Accessed: 2023-09-11.
- [19] Nir Eyal. 2014. *Hooked: How to build habit-forming products*. Penguin.
- [20] Hossein Falaki, Ratul Mahajan, Srikanth Kandula, Dimitrios Lymberopoulos, Ramesh Govindan, and Deborah Estrin. 2010. Diversity in smartphone usage. In *Proceedings of the 8th International Conference on Mobile Systems, Applications, and Services* (San Francisco, California, USA) (*MobiSys '10*). Association for Computing Machinery, New York, NY, USA, 179–194. <https://doi.org/10.1145/1814433.1814453>
- [21] Marco Fasoli. 2021. The overuse of digital technologies: human weaknesses, design strategies and ethical concerns. *Philosophy & Technology* 34, 4 (2021), 1409–1427. <https://doi.org/10.1007/s13347-021-00463-6>
- [22] Jesse Fox and Jennifer J. Moreland. 2015. The dark side of social networking sites: An exploration of the relational and psychological stressors associated with Facebook use and affordances. *Computers in Human Behavior* 45 (2015), 168–176. <https://doi.org/10.1016/j.chb.2014.11.083>
- [23] Mirta Galesic and Michael Bosnjak. 2009. Effects of Questionnaire Length on Participation and Indicators of Response Quality in a Web Survey. *Public Opinion Quarterly* 73, 2 (05 2009), 349–360. <https://doi.org/10.1093/poq/nfp031> arXiv:<https://academic.oup.com/poq/article-pdf/73/2/349/5462698/nfp031.pdf>
- [24] Thomas Gilovich and Victoria Medvec. 1995. The Experience of Regret: What, When, and Why. *Psychological review* 102 (04 1995), 379–95. <https://doi.org/10.1037/0033-295X.102.2.379>
- [25] Google. 2024. Focus your time with tech - Digital Wellbeing. <https://wellbeing.google/get-started/focus-your-time-with-tech/>. Accessed: 2024-09-10.
- [26] Maxi Heitmayer and Saadi Lahou. 2021. Why are smartphones disruptive? An empirical study of smartphone use in real-life contexts. *Computers in Human Behavior* 116 (2021), 106637.
- [27] Joel M Hektnar, Jennifer Anne Schmidt, and Mihaly Csikszentmihalyi. 2007. *Experience sampling method: Measuring the quality of everyday life*. Sage.
- [28] 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>
- [29] Alexis Hiniker, Shwetak N. Patel, Tadayoshi Kohno, and Julie A. Kientz. 2016. Why would you do that? predicting the uses and gratifications behind smartphone-usage behaviors. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (Heidelberg, Germany) (*UbiComp '16*). Association for Computing Machinery, New York, NY, USA, 634–645. <https://doi.org/10.1145/2971648.2971762>
- [30] Apple Inc. 2024. Use Screen Time on your iPhone or iPad. <https://support.apple.com/en-us/108806>. Accessed: 2024-09-10.
- [31] 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>
- [32] Hyung-Kwon Ko, Gwanmo Park, Hyeon Jeon, Jaemin Jo, Juho Kim, and Jinwook Seo. 2023. Large-scale Text-to-Image Generation Models for Visual Artists' Creative Works. In *Proceedings of the 28th International Conference on Intelligent User Interfaces* (Sydney, NSW, Australia) (*IUI '23*). Association for Computing Machinery, New York, NY, USA, 919–933. <https://doi.org/10.1145/3581641.3584078>
- [33] Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large Language Models are Zero-Shot Reasoners. In *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (Eds.), Vol. 35. Curran Associates, Inc., 22199–22213. https://proceedings.neurips.cc/paper_files/paper/2022/file/8bb0d291acd4acf06ef12099c16f326-Paper-Conference.pdf
- [34] Sumit Kumar, Gopi Ramena, Manoj Goyal, Debi Mohanty, Ankur Agarwal, Benu Changmai, and Sukumar Moharana. 2020. On-Device Information Extraction from Screenshots in form of tags. In *Proceedings of the 7th ACM IKDD CoDS and 25th COMAD* (Hyderabad, India) (*CoDS COMAD 2020*). Association for Computing Machinery, New York, NY, USA, 275–281. <https://doi.org/10.1145/3371158.3371200>
- [35] Tao Lu, Hongxiao Zheng, Tianying Zhang, Xuhai “Orson” Xu, and Anhong Guo. 2024. InteractOut: Leveraging Interaction Proxies as Input Manipulation Strategies for Reducing Smartphone Overuse. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (*CHI '24*). Association for Computing Machinery, New York, NY, USA, Article 245, 19 pages. <https://doi.org/10.1145/3613904.3642317>
- [36] 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>
- [37] Kai Lukoff, Cissy Yu, Julie Kientz, and Alexis Hiniker. 2018. What Makes Smartphone Use Meaningful or Meaningless? *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 1, Article 22 (mar 2018), 26 pages. <https://doi.org/10.1145/3191754>
- [38] Arunesh Mathur, Gunes Acar, Michael J. Friedman, Eli Lucherini, Jonathan Mayer, Marshini Chetty, and Arvind Narayanan. 2019. Dark Patterns at Scale: Findings from a Crawl of 11K Shopping Websites. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 81 (nov 2019), 32 pages. <https://doi.org/10.1145/3359183>
- [39] Dale T. Miller, William Turnbull, and Cathy McFarland. 1990. Counterfactual Thinking and Social Perception: Thinking about What Might Have Been. Advances in Experimental Social Psychology, Vol. 23. Academic Press, 305–331. [https://doi.org/10.1016/S0065-2601\(08\)60322-6](https://doi.org/10.1016/S0065-2601(08)60322-6)
- [40] Alberto Monge Roffarello, Kai Lukoff, and Luigi De Russis. 2023. Defining and Identifying Attention Capture Deceptive Designs in Digital Interfaces. 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 194, 19 pages. <https://doi.org/10.1145/3544548.3580729>
- [41] Arvind Narayanan, Arunesh Mathur, Marshini Chetty, and Mihir Kshirsagar. 2020. Dark patterns: past, present, and future. *Commun. ACM* 63, 9 (aug 2020), 42–47. <https://doi.org/10.1145/3397884>
- [42] OpenAI. 2024. Hello GPT-4o. <https://openai.com/index/hello-gpt-4o/>. Accessed: 2024-09-10.
- [43] OpenAI. 2024. Introducing vision to the fine-tuning API. <https://openai.com/index/introducing-vision-to-the-fine-tuning-api/> Accessed: 2024-09-10.
- [44] OpenAI. 2024. Structured Outputs Guide. <https://platform.openai.com/docs/guides/structured-outputs>. Accessed: 2024-09-10.
- [45] OpenAI. 2024. Vision Guide. <https://platform.openai.com/docs/guides/vision>. Accessed: 2024-09-10.
- [46] Adiba Orzikulova, Hyunsung Cho, Hye-Young Chung, Hwjung 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), 30 pages. <https://doi.org/10.1145/3610065>
- [47] 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* (Honolulu, HI, USA) (*CHI '24*). Association for Computing Machinery, New York, NY, USA, Article 250, 20 pages. <https://doi.org/10.1145/3613904.3642747>
- [48] Antti Oulasvirta, Tye Rattenbury, Lingyi Ma, and Eeva Raita. 2012. Habits make smartphone use more pervasive. *Personal and Ubiquitous computing* 16 (2012), 105–114. <https://doi.org/10.1007/s00779-011-0412-2>
- [49] Martin Pielot, Karen Church, and Rodrigo De Oliveira. 2014. An in-situ study of mobile phone notifications. In *Proceedings of the 16th international conference on Human-computer interaction with mobile devices & services*, 233–242.
- [50] Martin Pielot, Tilman Dingler, Jose San Pedro, and Nuria Oliver. 2015. When attention is not scarce - detecting boredom from mobile phone usage. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '15)*. Association for Computing Machinery, New York, NY, USA, 825–836. <https://doi.org/10.1145/2750858.2804252>
- [51] Rik Pieters and Marcel Zeelenberg. 2005. On bad decisions and deciding badly: When intention-behavior inconsistency is regrettable. *Organizational Behavior and Human Decision Processes* 97, 1 (2005), 18–30. <https://doi.org/10.1016/j.obhdp.2005.01.003>
- [52] Byron Reeves, Nilam Ram, Thomas N Robinson, James J Cummings, C Lee Giles, Jennifer Pan, Agnese Chiatti, MJ Cho, Katie Roehrick, Xiao Yang, et al. 2021. Screenomics: A framework to capture and analyze personal life experiences and the ways that technology shapes them. *Human-Computer Interaction* 36, 2 (2021), 150–201. <https://doi.org/10.1080/07370024.2019.1578652> PMID: 33867652
- [53] Neil Richards and Woodrow Hartzog. 2024. Against Engagement. *Boston University Law Review* 104 (2024), 1151. <https://ssrn.com/abstract=4845648>
- [54] Adi Sagi and Nehemia Friedland. 2007. The Cost of Richness: The Effect of the Size and Diversity of Decision Sets on Post-Decision Regret. *Journal of personality and social psychology* 93 (10 2007), 515–24. <https://doi.org/10.1037/0022-3514.93.4.515>
- [55] Choonsung Shin and Anind K. Dey. 2013. Automatically detecting problematic use of smartphones. In *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (Zurich, Switzerland) (*UbiComp '13*). Association for Computing Machinery, New York, NY, USA, 335–344. <https://doi.org/10.1145/2493432.2493443>
- [56] Jonathan A. Tran, Katie S. Yang, Katie Davis, and Alexis Hiniker. 2019. Modeling the Engagement-Disengagement Cycle of Compulsive Phone 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, 1–14. <https://doi.org/10.1145/3290605.3300542>
- [57] Joseph Tu, Derrick Wang, Lydia Choong, Arnold Dian Jr. Abistado, Ally Suarez, Stuart Hallifax, Katja Rogers, and Lennart Nacke. 2024. Rolling in Fun, Paying the Price: A Thematic Analysis on Purchase and Play in Tabletop Games. *ACM Games* (Oct. 2024). <https://doi.org/10.1145/3700628> Just Accepted.

- [58] Patti M Valkenburg, Irene I van Driel, and Inc Beyens. 2022. The associations of active and passive social media use with well-being: A critical scoping review. *New Media & Society* 24, 2 (2022), 530–549. <https://doi.org/10.1177/14614448211065425>
- [59] Niels van Berkel, Denzil Ferreira, and Vassilis Kostakos. 2017. The Experience Sampling Method on Mobile Devices. *ACM Comput. Surv.* 50, 6, Article 93 (Dec. 2017), 40 pages. <https://doi.org/10.1145/3123988>
- [60] Frenk Van Harreveld, Joop van der Pligt, and Loran Nordgren. 2008. The relativity of bad decisions: Social comparison as a means to alleviate regret. *British Journal of Social Psychology* 47, 1 (2008), 105–117. <https://doi.org/10.1348/014466607X260134>
- [61] Yuxuan Wan, Chaozheng Wang, Yi Dong, Wenxuan Wang, Shuqing Li, Yintong Huo, and Michael R. Lyu. 2024. Automatically Generating UI Code from Screenshot: A Divide-and-Conquer-Based Approach. arXiv:2406.16386 [cs.SE] <https://arxiv.org/abs/2406.16386>
- [62] Junyang Wang, Haiyang Xu, Jiaob Ye, Ming Yan, Weizhou Shen, Ji Zhang, Fei Huang, and Jitao Sang. 2024. Mobile-Agent: Autonomous Multi-Modal Mobile Device Agent with Visual Perception. arXiv:2401.16158 [cs.CL] <https://arxiv.org/abs/2401.16158>
- [63] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (Eds.), Vol. 35. Curran Associates, Inc., 24824–24837. https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abc4-Paper-Conference.pdf
- [64] Ruolan Wu, Chun Yu, Xiaole Pan, Yujia Liu, Ningning Zhang, Yue Fu, Yuhan Wang, Zhi Zheng, Li Chen, Qiaolei Jiang, Xuhai Xu, and Yuanchun Shi. 2024. MindShift: Leveraging Large Language Models for Mental-States-Based Problematic Smartphone Use Intervention. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '24). Association for Computing Machinery, New York, NY, USA, Article 248, 24 pages. <https://doi.org/10.1145/3613904.3642790>
- [65] 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 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) (CHI '22). Association for Computing Machinery, New York, NY, USA, Article 442, 17 pages. <https://doi.org/10.1145/3491102.3517476>
- [66] Xiao Yang, Nilam Ram, Thomas Robinson, and Byron Reeves. 2019. Using Screenshots to Predict Task Switching on Smartphones. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) (CHI EA '19). Association for Computing Machinery, New York, NY, USA, 1–6. <https://doi.org/10.1145/3290607.3313089>
- [67] Xiu-Juan Yang, Qing-Qi Liu, Shuai-Lei Lian, and Zong-Kui Zhou. 2020. Are bored minds more likely to be addicted? The relationship between boredom proneness and problematic mobile phone use. *Addictive Behaviors* 108 (2020), 106426.
- [68] Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. 2023. The Dawn of LMMs: Preliminary Explorations with GPT-4V(ision). arXiv:2309.17421 [cs.CV] <https://arxiv.org/abs/2309.17421>
- [69] Marcel Zeelenberg. 1999. Anticipated regret, expected feedback and behavioral decision making. *Journal of behavioral decision making* 12, 2 (1999), 93–106. [https://doi.org/10.1002/\(SICI\)1099-0771\(199906\)12:2<93::AID-BDM311>3.0.CO;2-S](https://doi.org/10.1002/(SICI)1099-0771(199906)12:2<93::AID-BDM311>3.0.CO;2-S)
- [70] Marcel Zeelenberg and Rik Pieters. 2007. A Theory of Regret Regulation 1.0. *Journal of Consumer Psychology* 17, 1 (2007), 3–18. https://doi.org/10.1207/s15327663jcp1701_3
- [71] Chi Zhang, Zhao Yang, Jiaxuan Liu, Yucheng Han, Xin Chen, Zebiao Huang, Bin Fu, and Gang Yu. 2023. AppAgent: Multimodal Agents as Smartphone Users. arXiv:2312.13771 [cs.CV] <https://arxiv.org/abs/2312.13771>
- [72] Xiaoyi Zhang, Lilian de Greef, Amanda Swearngin, Samuel White, Kyle Murray, Lisa Yu, Qi Shan, Jeffrey Nichols, Jason Wu, Chris Fleizach, Aaron Everitt, and Jeffrey P Bigham. 2021. Screen Recognition: Creating Accessibility Metadata for Mobile Applications from Pixels. 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 275, 15 pages. <https://doi.org/10.1145/3411764.3445186>
- [73] Jian Zheng and Ge Gao. 2024. Fragmented Moments, Balanced Choices: How Do People Make Use of Their Waiting Time?. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–14. <https://doi.org/10.1145/3613904.3642608>

A LLM Prompts

A.1 Screenshot to Text Description (Step 1 in Figure 3)

```
{"role": "system", "content": "You are a helpful assistant specializing in visual content analysis of smartphone screenshots."}, {"role": "user", "content": ["Here are two smartphone screenshots, taken 5 seconds apart within the same app Instagram. Your primary task is to describe all of the visual elements (including all of the UI components and content) and the user's activity in the SECOND screenshot in great detail. Please also use the first screenshot as contextual information, and describe how the user transitioned from the first screenshot to the second. Do not describe the first screenshot in detail. In your response, refer to the first screenshot as 'the previous screenshot', and the second screenshot as 'the screenshot'.", {"image": ...}, {"image": ...}]}
```

A.2 Text Description to Category (Step 3 in Figure 3)

```
{"role": "system", "content": "You are a helpful assistant specializing in understanding text descriptions of smartphone screenshots."}, {"role": "user", "content": "Below are text descriptions of screenshots taken every 5 seconds on Instagram, which include summaries of the visual elements of the screenshot and the user's activity. Your task is to categorize user's activity of the LAST screenshot into one of the following categories. Justify your choices with a step-by-step rationale."}
```

Active Communication: The screenshot includes the presence of a private messaging interface, suggesting the user is actively communicating with specific individuals or groups.

Active Search: The screenshot suggests the user is using the search feature to find specific information, content, articles, or items or is consuming content they found through active searching.

Consuming Recommendation-Based Content: The screenshot shows explicit indicators of recommendation-based content in the user's feed, such as a "For You" tab, "Suggested Post" labels, or buttons like "Follow" and "Join" which allow users to subscribe to new content.

Consuming Subscription-Based Content: The screenshot shows content already followed or subscribed to by the user in their feed. The screenshot should include indicators such as an active 'Following' or 'Subscription' tab of the app, or signs suggesting that the user has already followed the content poster, such as the absence of buttons next to the content poster or community to follow, join, or subscribe in a feed interface.

Consuming Content Shared by Others: The screenshot

suggests the user is viewing content shared or reposted by someone they followed or opened from a private conversation.

Viewing Comments or Discussion Thread: The screenshot suggests the user is viewing the comment section of discussion thread of a social media post.

None of the Above: When the user opens a link to a website, when the user sees sponsored content or ad, when it is unclear if user is viewing content posted by someone they followed or recommended to them, or when seeing these screens: home screen, notification screen, a black, dimmed, or blank screen, a screen showing a survey prompt.

RULES:

1. Pay attention to how the user transitioned from earlier screenshots. Look for evidence from the four previous screenshots when needed. For example, if the last screenshot shows some content in full screen, without an indicator of where the content came from, look at previous screenshots where the same content appeared and find the relevant indicators.
2. Do NOT always assume that the user is Consuming Subscription-Based Content when you do not see a follow button. Check if the user previously opened the video from a conversation, or did active searching to find

the result, and look at previous screenshots to find out if the follow button was present when the user was viewing the same piece of content.

3. Do NOT consider general social media features such as likes, comments, shares, or hashtags as indicators of recommendation-based content.
4. If the user is looking at a post with a comment section but the majority of the screenshot is not about the comments, do not categorize the user's activity as Viewing Comments or Discussion Thread.
5. If the user is looking at multiple social media posts in their feed, pay attention to the one the user is most likely looking at, such as the one in the center and showing the full content. Ignore partially visible posts in the feed.

Screenshot Descriptions:

```
Screenshot 1 Description: {...}  
Screenshot 2 Description: {...}  
Screenshot 3 Description: {...}  
Screenshot 4 Description: {...}  
Screenshot 5 Description: {...}  
"}
```

B Interview Codebook

See Tables 6.

Table 6: Interview codebook.

Code	Secondary Code	Definition	Example
Session Duration	Extensive Use	The participant spent a prolonged period on their phone.	"I think it was also about the length—it was just so long. It felt unnecessarily prolonged. So, yes, I regretted that session." (P12, F1)
	Brief and Intermittent Use	The participant spent a short time in a given session.	"I started watching that and scrolling a bit through my homepage. It didn't get completely out of hand—I wasn't scrolling for 30 minutes or anything. I kept it relatively controlled." (P9, F1)
Intended Use	No Specific Goal	The participant engaged with their phone habitually or without a clear goal, or engaged in "doomscrolling".	"I think sometimes it's almost unconscious to open a messaging app. For example, in the theater, I guess I was maybe expecting a message, so I opened it. But then I just closed it without doing anything because there wasn't any message." (P8, F2)
	Having a Specific Goal	The participant engaged with their phone with a clear intention.	"Someone asked me a question about which division my school operates in, so that's what I was doing. No regret there—it was a very specific intention." (P12, F2)
	Intention-Behavior Inconsistency	Phone use deviates from the user's original purpose.	"You could see it as me trying to complete my task, which was finding a video to help me fall asleep, like rain sounds. But I couldn't get to that until I ended up checking YouTube Shorts beforehand. That part didn't have any meaning to me" (P9, F1)
Social Media Activity	Active Communication	The participant used direct messaging features to communicate or interact with others.	"This is WhatsApp with my family and I definitely don't regret it because I wanted to get updates on what my brother and his girlfriend are doing because they're traveling right now and it's, uh, important for me to communicate with my family." (P4, F2)
	Active Search	The participant searched for a specific piece of information.	"I wanted to see what deals were available, and I also wanted to find out what restaurants were there, so I could actually get information." (P1, F1)
	Consuming Recommendation-Based Content	The participant viewed social media content algorithmically recommended by the platform.	"So random content, like what Twitter gives me or Facebook. Like almost all of it was random. It wasn't people I knew, right. So like that content, I end up regretting more than if it's content of like people I know or a work thing, like an article I was reading." (P11, F1)
	Consuming Subscription-Based Content	The participant viewed social media content posted by accounts they followed.	"Every time I see her, I watch her latest video. So that's interesting. It must have been about just feeling like I was spending too much time on social media because I like watching her videos." (P7, F1)
	Consuming Content Shared by Others	The participant viewed social media content shared by their friends.	"I opened a TikTok because my friend sent it to me. I don't regret it because I needed to watch it to reply to her message. It was pretty short, and I did find it entertaining." (P4, F1)
	Viewing Comments or Discussion Thread	The participant viewed comments or discussions from others on various topics.	"Sometimes when I go into comment sections, then I'd be more likely to regret the content because especially with political topics, there's so many obnoxious comments in the comment section." (P7, F1)
External Factors	Bedtime Phone Use	Phone use during or before bedtime that affected sleep.	"When I'm playing on social media at night, when I should be sleeping, when I'm having problems sleeping, tends to be when I regret it more. I realized during this that I feel more regretful about being on social media in the middle of the night, especially when it's just looking at videos and memes." (P6, F1)
	Using the Phone at Work	Phone use during work time.	"If I'm stressed with work, I think that leads to more regret because like my time which I should be working and then I feel like I wasted my time." (P8, F2)
	Using the Phone during Commuting	Phone use during the time the participant was traveling between places.	"If I'm on social media while I'm on the bus or light rail or walking to work, I never regret that time because it's not like there's many other productive things I could realistically be doing while I'm commuting." (P4, F1)