Designing Reflective Thinking-Based Contextual Privacy Policy for Mobile Applications

Shuning Zhang*
zsn23@mails.tsinghua.edu.cn

Sixing Tao*
sixint@uw.edu

Eve He
eveanny.hx@gmail.com

Yuting Yang
yutingyang986@gmail.com

Ying Ma
ying.ma1@student.umimelb.edu.au

Ailei Wang
wal22@mails.tsinghua.edu.cn

Xin Yi†
yixin@tsinghua.edu.cn

Hewu Li
lihewu@cernet.edu.cn

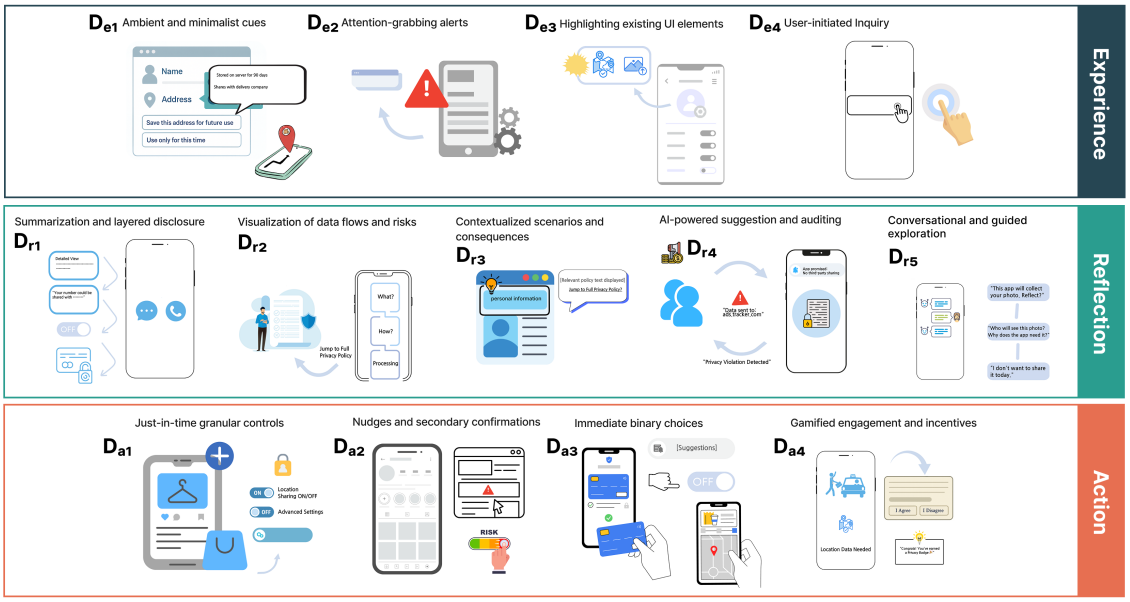


Figure 1: Synthesized design abstractions illustrating the CPP design space across the three stages of the reflective thinking-based framework: experience (D_{e1}-D_{e4}), reflection (D_{r1}-D_{r5}), and action stage (D_{a1}-D_{a4}).

Abstract

Traditional mobile privacy policies are often ignored due to their excessive length and legalistic complexity. While Contextual Privacy Policies (CPPs) present information at the point of risk, they often act as static displays that fail to engage users or prompt active decision making. To address these limitations, we construct a design space for reflective CPPs based on the Reflective Thinking Framework. Through three expert workshops, we mapped the hypothetical user interaction around CPPs to three stages of reflection: Experience (triggering surprise via contextual cues), Reflection (supporting inquiry through risk scenarios), and Action (empowering users with granular controls). We instantiated this design space with Conflect, a high-fidelity mobile prototype featuring non-intrusive sidebar alerts and dynamically synthesized risk explanations. A user study (N=48) comparing Conflect against state-of-the-art CPP designs shows that our reflective approach effectively reduces users' perceived habitual action and improves privacy comprehension. Notably, Conflect facilitates these gains without increasing cognitive load or compromising usability compared to CPPs.

Keywords

Privacy policy, Contextual privacy policy, Interaction design, Design space, Reflective thinking

*Equal contribution.
†Corresponding author.

Proceedings on Privacy Enhancing Technologies 2026(4), 814–836

https://doi.org/10.56553/popets-2026-0146

1 Introduction

Privacy policies serve as the primary mechanism for informing users about data practices in mobile applications [1, 23]. However,

their effectiveness is undermined by policies' excessive length and complexity. Research indicates that the average privacy policy contains approximately 4,000 words, which takes the average adult an estimated 16 minutes to read [8]. Consequently, 74% of users reported that they completely bypass these policies, opting instead for the "quick join" option without engaging with content [43]. This pattern is largely driven by the cognitive burden imposed by the verbose and legalistic nature of traditional privacy notices. In addition, traditional privacy policies are typically presented only once at app launch, making it difficult for users to connect policy information to privacy-relevant situations that arise during use [45].

To address these limitations, researchers developed "just-in-time" notices [52] and Contextual Privacy Policies (CPPs) [44, 62]. **As introduced by Feth [22], CPPs present privacy information at the point of user interaction, contextualizing abstract data practices rather than relegating them to separate, static documents.** While CPPs reduce information overload, timely delivery alone does not ensure meaningful engagement. Many existing CPPs still function primarily as disclosure mechanisms, offering limited support for helping users interpret privacy implications, reflect on potential consequences, or act on that understanding. This suggests a need to move beyond information delivery toward designs that support contextual reflective engagement of privacy implications.

In this paper, we address this need by constructing a design space for CPPs grounded in the *Reflective Thinking Framework* [21, 30]. **Drawing on established literature [18, 53], we define reflective thinking as a conscious process of examining one's assumptions and considering alternatives when faced with unexpected situations, before taking action.** We use this specific framework as it provides the theoretical potential to mitigate user habituation, and bridge the cognitive gap between passive privacy awareness and informed action. Since contextual cues are essential for stimulating reflection, this approach aligns with the objectives of contextual privacy policy design. We use this framework to structure interactions that disrupt habitual engagement and facilitate intentional decision-making. Our investigation therefore progress from the conceptualization of this design space to its practical instantiation, and subsequent empirical evaluation. Specifically, we address the following research questions (RQs):

RQ1: What design dimensions constitute a reflective-thinking-based design space for mobile CPPs?

RQ2: What are the interface elements of a reflective-thinking-based CPP?

RQ3: How users perceive privacy risks and interactive experience within this framework?

To address RQ1, we theoretically grounded our approach in a reflective thinking framework. We conducted three workshops with experienced researchers to construct a design space for mobile CPPs. Through a multi-stage process, participants identified six problems, with *disconnect between abstract policy text and concrete user actions* as the most critical problem to solve. We synthesized these findings into design patterns across three reflective thinking stages: experience, reflection, and action. For the experience stage, participants favored non-intrusive methods like *ambient and minimalist cues*. For the reflection stage, there was a strong consensus on translating policy into *contextualized, risk-based scenarios*, supplemented by *summarization and layered disclosure*, to make abstract risks tangible. Finally, for the action stage, *just-in-time granular control* was rated highest, underscoring the importance of empowering users with immediate action choices.

Informed by these findings, we developed Conflect, a high-fidelity CPP prototype that operationalizes this reflective thinking framework (Figure 3). For the experience stage, it appears by default as a minimal sidebar that, when activated, highlights relevant UI elements to ground the user's experience in the immediate context. For the reflection stage, it presents notifications that frame data practices as tangible, risk-based scenarios, and allow users to view the corresponding policy excerpts to facilitate reflection. For the action stage, it nudges users toward actionable choices, such as adjusting specific permissions, to complete the reflective cycle.

To evaluate the impact of this approach (RQ3), we conducted a user study (N=48) comparing Conflect against a state-of-the-art CPP for mobile applications [45]. We found that Conflect significantly disrupts automatic, habitual behaviors and improves users' privacy comprehension. While evidence for fostering deeper, critical reflection was limited within this short-term study, Conflect effectively enhanced privacy awareness without increasing cognitive load or compromising overall usability.

In summary, this paper makes three contributions:

- A theoretically grounded design space for mobile CPPs based on the *Reflective Thinking Framework*.
- Development of Conflect, a prototype that instantiates this design space to engage users in active reflection.
- Empirical findings showing that a reflective approach reduces habitual action and improves privacy comprehension, while maintaining comparable usability.

2 Background & Related Work

2.1 Privacy Policy Visualization

Initial efforts in privacy policy visualization centered on machine-readable formats like the Platform for Privacy Preferences (P3P) and tools such as PrivacyBird [17], which aimed to be both human-understandable and machine-readable. However, achieving widespread adoption proved challenging. Consequently, research shifted toward condensing policy information into digestible formats, such as privacy nutrition labels [32, 33] and related tabular forms [33, 47]. While these formats were later enhanced with interactivity [48] and adapted for mobile applications [65], a key limitation remained that the visualizations were not presented at the precise moment a privacy risk occurred. This temporal disconnect limited their effectiveness in informed consent.

A parallel stream of research explored expressive and engaging modalities, including comics [59], and serious games [58]. Although these approaches successfully increased user engagement, they introduced practical challenges, such as lengthy interaction times and difficulties in automatic generation.

Finally, researchers explores embedding privacy notices in a contextual form, as it directly correlates the visualization with the specific privacy risk at hand [45]. This line of work has progressed from initial conceptualizations [44] to practical implementations on websites [62] and mobile apps [45]. However, existing contextual notices still lack the interface design to promote user reflection on

privacy risks, particularly on screen-constrained mobile devices. Addressing this gap, we adopt a reflective thinking perspective to design and visualize CPPs for mobile applications.

2.2 Contextual Privacy Policy

CPPs deliver policy information during relevant user interactions, addressing the limits of traditional monolithic policies [9, 22]. Early work proposed reorganizing policy text around interaction contexts [9] and tailoring statements to user actions [22], but these efforts prioritized information structure over interaction design.

Subsequent research validated the effectiveness of CPPs. Patil et al. [46] highlighted that privacy feedback is most effective when delivered at the right time and at a manageable rate, which CPPs support. Bergmann et al. [7] similarly found that presenting information in its corresponding context increases awareness. CPPs have also been shown to improve transparency and support more user-centric communication [40, 44]. Ortloff et al. [44] further showed this through a concept showcase across seven websites, manually identifying contextual placements. While these studies confirmed the benefits of contextual information delivery, they did not investigate how the *design* of the CPP itself could encourage deep user reflection on potential privacy risks.

Besides design, research also focused on automating CPP's generation. Zimmeck and Bellovin [70] introduced Privee, an architecture for automatic privacy policy analysis, while Mysore Sathyendra et al. [51] developed classification models to identify opt-out choices in policies. Harkous et al. [27] introduced Polisis, an automated framework using hierarchical neural network classifiers for high-level and fine-grained queries, offering a promising approach for annotating entire policies.

Recent works produced end-to-end tools capable of automatically generating and injecting contextual notices. Windl et al. [62] proposed PrivacyInjector, a production AI tool designed to automatically generate contextual privacy policies for websites. Pan et al. [45] extended CPPs to mobile applications, presenting a novel algorithm for generating CPPs in that context. However, a research gap remains regarding how interaction design of CPP influences users' reflection and behavior, which our paper addresses.

2.3 Reflective Thinking

Dewey [18] defines reflection as an active inquiry triggered by unexpected situations to solve problems. Building on this, Schön [53] distinguishes between reflection-on-action (common in personal informatics [38]) and reflection-in-action. Specifically, reflection-in-action is sparked by a "surprise" that disrupts automatic user behaviors, fostering conscious inquiry and transformation [53]. This framework has been widely adopted and expanded by subsequent researchers [4, 56].

To translate this "surprise" into practice, Bentvelzen et al. [6] identify four design resources: temporal perspective, conversation, comparison, and discovery. For instance, conversation facilitates dialogue with others, chatbots, or oneself, while discovery reveals new patterns and insights. Operationally, design strategies like slow technology [26] and design frictions [16, 50] can effectively introduce these moments of surprise.

Notably, the privacy paradox [35], and the well-known problems that users are reluctant to read privacy policies [43] have bothered researchers for years, which reflective thinking could help. For example, Terpstra et al. [60] argue that Human-Computer Interaction (HCI) should intentionally design the friction to stimulate a user's conscious deliberation on privacy. They also proposed a design-for-privacy model where such friction triggers reflective thinking, pairing it with meaningful controls that empower users to act on their privacy preferences [61].

To operationalize these theoretical principles, we apply the *Experience-Reflection-Action (ERA)* [30] framework, and map each phase to specific CPP design elements. For the *experience* stage which acts as the trigger, the theoretical "surprise" could be operationalized via specific UI frictions or alerts, such as context-aware interruptions. This serves as the triggering condition to disrupt the user's automatic workflow. For the *reflection* stage, CPP could implement specific features, such as the dialogue reflection highlighted in Fleck and Fitzpatrick's [24] framework, that encourage users to explore the "why" and "how" of data usage. For the *action* stage, the system could map insights to actionable controls, allowing users to modify granular system permissions immediately following reflective phase.

3 Constructing a Mobile CPP Design Space via Reflective Thinking-Guided Workshops

3.1 Operationalization of Reflective Thinking Framework

To ensure the generated designs were grounded in reflective thinking, we utilized the ERA framework to formulate specific design constraints for the workshops. We operationalized the *experience* stage as the requirement for a "surprising" interruption, challenging participants to design triggers that effectively shift users from automatic information consumption to conscious attention, distinguishing this approach from easily ignored traditional notices. For the *reflection* stage, we called for visualizations that reveal the latent consequences of data sharing, which compel users to actively think about risks rather than passively observe them. Finally, we framed the *action* stage as "informed control", providing mechanisms for users to exercise immediate agency over their data to ensure that cognitive engagement translates into concrete outcomes.

3.2 Study Setup

Recruitment and Participants. We recruited 10 participants through personal contacts, snowball sampling [25], and targeted email invitations to authors who had published on privacy policy visualization in major HCI and privacy venues within the past five years. To capture the interdisciplinary nature of this design space, spanning usable privacy, policy comprehension, and interface design, we recruited participants with overlapping expertise across multiple domains. Participant demographics and background are summarized in Table 1. The study was approved by the IRB of our university, and participants received 350 Chinese Yuan (CNY), which aligned with local compensation standards.

Study Design. We conducted three small expert workshops (3–4 participants each) rather than a single large focus group to reduce

Table 1: Participant demographics, including domain expertise, gender, age, country of residence, years of research or design experience, self-rated User-Centered Design (UCD) familiarity on a 1–10 scale, and publication count in top-tier venues within the past five years. *Note: P1 operates in industry, where paper publication is not a primary requirement.*

PID	Expertise	Gender	Age	Country	Exp.	UCD	Papers
P1	HCI Researcher	F	28	China	4	7	0
P2	HCI Researcher; Privacy/Security Researcher	F	28	Australia	3	7	11
P3	UI/UX Designer	F	25	New Zealand	3	7	0
P4	Privacy/Security Researcher	M	28	USA	4	8	12
P5	Privacy/Security Researcher	M	24	Germany	2	4	2
P6	HCI Researcher; Privacy/Security Researcher	F	22	USA	4	7	10
P7	Software Engineering Researcher	F	40	Australia	13	5	102
P8	HCI Researcher; Privacy/Security Researcher	M	27	USA	7	9	14
P9	HCI Researcher; Privacy/Security Researcher	M	21	China	5	8	15
P10	HCI Researcher	F	24	China	3	7	2
Mean (SD)			*26.7 (5.0)*		*4.8 (3.1)*	*6.9 (1.4)*	*16.8 (30.5)*

social pressure, mitigate dominance, and increase individual participation [12], while still enabling cross-session validation.

All workshops followed the same moderator guide and materials to ensure consistency (see Appendix C and D). Prior to the sessions, participants received a concise CPP primer, a figure from Pan et al. [45] illustrating the differences between CPPs and runtime or install-time notices, and an overview of the reflective-thinking framework with stage-specific prompts. We also provided two example CPP designs that walked through the reflective-thinking stages (*experience* → *reflection* → *action*), clearly framed as illustrative to avoid anchoring. As a pre-task, each participant prepared two CPP design ideas for mobile app usage scenarios. They were asked to focus on how their designs could facilitate reflective thinking on mobile devices, and to illustrate the interaction flow using the three reflective stages.

During each workshop, participants first shared and iterated on their designs through group discussion, focusing on interaction flows, specific privacy issues they aimed to address, and design rationales behind their choices. They were asked to explain how their designs prompted reflective thinking, and how reflective processes could help overcome common challenges of existing CPPs. The group then evaluated the advantages and disadvantages of various privacy policy problems for CPP interventions, identifying promising design directions based on their anticipated efficacy and feasibility. Each session concluded with open-ended feedback. All workshops were held online, and were audio- and video-recorded, with materials shared through a collaborative Figma workspace[1].

To inform our design space, researchers inductively derived design abstractions by mapping the raw workshop suggestions to higher-level conceptual dimensions. To synthesize findings across sessions and reduce bias, we conducted a follow-up cross-session prioritization survey. We illustrated the synthesized dimensions with anonymized idea cards, each representing a distinct design across a single stage. Participants ranked them by perceived effectiveness and feasibility.

Data Analysis. We conducted thematic analysis [10] of our transcribed text materials. The visual materials were first translated into text descriptions following established guidance [3]. Participants' design artifacts were treated as qualitative data, and abstracted into synthesized design representations that capture shared principles. One primary author first reviewed all the transcripts and constructed the initial codebook. Then the primary author and a secondary author discussed the initial codebook and reached consensus. These two authors coded the transcripts and iteratively refined the codebook, with regular discussions to resolve disagreements. Through this process, they mapped the raw design suggestions to higher-level conceptual categories, which constituted the structural dimensions of the design space. Given the inductive nature of this process and the documented drawbacks of calculating inter-rater reliability in this context [41], we ensured quality through these discussions rather than inter-rater reliability scores. For the cross-session prioritization, we aggregated the ranking results using Borda Count, a common method for combining ranked preferences [42]. This prioritization served as an important rationale for instantiating the design space into a concrete prototype.

3.3 Results

We present our findings in three parts: (i) a summary of key problems with current privacy policies that reflective thinking-based CPP could address, (ii) a reflective thinking-based CPP design space, (iii) a prioritization of designs within the design space.

3.3.1 Problems of Privacy Policies. Participants identified six practical problems (Pr1-6) inherent in current privacy policies (Table 2). By using a UCD approach to surface these problems and map them onto the ERA framework, we show how reflective design could resolve these problems.

Pr1: Disconnect between context and action. While well-documented in existing literature [45, 62], we identify this problem as a critical failure of the *action* stage, as users lack the immediate contextual triggers to initiate reflection. The spatial and temporal separation of policies, often hidden in legal menus or presented only at sign-up, prevents users from associating policy with in-situ operations. As P1 noted, policies are *"always displayed on the sign-up page,"* but users rarely connect this to subsequent operations within the app's workflow.

Pr2: Failure to communicate tangible risk. Current policies are written in abstract language that describes procedures rather than communicating real-world consequences, representing a failure of the *reflection* stage. Users struggle to translate "data sharing"

[1] http://figma.com/

into tangible consequences. P5 argued that the policy needs to explain the "worst-case situation", such as the risk of being stalked if the location information was shared, to foster conscious inquiry.

Pr3: Erosion of user agency and motivation. The binary "take-it-or-leave-it" nature of current policies induces "privacy resignation" [19], signaling a failure in the *action* stage. P7 explained that the lack of options beyond "accept all data collection or stop using the app entirely" causes users to *"lose a lot of motivation to read the policy"*.

Pr4: Information overload and poor readability. The density of legal text exceeds cognitive capacity, forcing users into heuristic processing, such as skipping, rather than systematic reflective processing [48, 65]. This disrupts the *experience* stage. P9 hoped to design simple, visual cues to convey information "at a glance" without the need for "lengthy text descriptions".

Pr5: Mismatch between policy and practice. Skepticism regarding whether stated policies match actual app behaviors creates distrust that inhibits engagement. P4 mentioned the real-world "rumors" of apps selling personal information, a practice that is unlikely to be officially stated in policies.

Pr6: Outdated content for modern risks. Static documents fail to address emerging anxieties, such as data use for AI training, leaving users without the necessary information to reflect on modern privacy threats. P1 argued that many current policies are a "product of the last decade."

3.3.2 A Design Space For Reflective Thinking-Based CPPs. To overcome these barriers, we synthesized a design space for CPP, which is mapped to the ERA framework [30, 53]. This framework transforms CPPs into dynamic reflection supports: *Experience* designs trigger "surprise" to disrupt habituation, *Reflection* designs support sense-making , and *Action* designs enable control (see Figure 2).

Experience Stage. Designs in this stage aim to induce the "surprise" necessary to shift users from automatic task execution to conscious awareness, addressing *disconnect between context and action* (Pr1) and *information overload* (Pr4).

D_{e1}: Ambient and minimalist cues. This design uses subtle indicators integrated into the user interface to mitigate *information overload* (Pr4). It signals the presence of a privacy-relevant action without disrupting the primary task. For example, P3 proposed using numerical indicators to show data frequency.

D_{e2}: Attention-grabbing alerts. For high-stakes data transactions, this approach uses prominent visual cues to actively capture user attention, interrupt habituation, and signal immediate tangible risk (Pr2). P5 suggested red exclamation for intrusive collection.

D_{e3}: Highlighting existing UI elements. By visually emphasizing otherwise overlooked on-screen elements (Pr1), this design anchors abstract privacy concerns to concrete interface objects, thereby triggering reflection. P6's design highlighted personal information like a child's name to make abstract concepts tangible.

D_{e4}: User-initiated inquiry. This empowers users to proactively seek information, supporting reflection driven by user curiosity. P7 designed a floating button that users could drag onto UI elements to retrieve relevant information on demand.

Reflection Stage. This dimension supports users' cognitive process of "inquiry", translating abstract data practices into comprehensible insights to address *failure to communicate tangible risk* (Pr2), *information overload* (Pr4), *mismatch between policy and practice* (Pr5) and *outdated content* (Pr6).

Table 2: The aggregated ranking scores (Borda Count) of designs and problems.

(a) Designs of experience stage.

Item	Score
D_{e3}: Highlight existing UI elements	20
D_{e1}: Ambient and minimalist cues	19
D_{e2}: Attention-grabbing alerts	11
D_{e4}: User-initiated inquiry	10

(b) Designs of reflection stage.

Item	Score
D_{r3}: Contextualized scenarios and consequences	32
D_{r1}: Summarization and layered disclosure	24
D_{r2}: Visualization of data flows and risks	23
D_{r5}: AI-powered suggestion and auditing	11
D_{r4}: Conversational and guided exploration	10

(c) Designs of action stage.

Item	Score
D_{a1}: Just-in-Time granular controls	23
D_{a2}: Immediate binary choices	18
D_{a3}: Nudges and secondary confirmations	14
D_{a4}: Gamified engagement and incentives	5

(d) Problems.

Item	Score
Pr1: Disconnect between context and action	40
Pr4: Information overload and poor readability	33
Pr2: Failure to communicate tangible risk	24
Pr5: Mismatch between policy and practice	22
Pr3: Erosion of user agency and motivation	18
Pr6: Outdated content for modern risks	13

D_{r1}: Summarization and layered disclosure. To counter information overload and facilitate users' conscious inquiry (Pr4), this design structures content hierarchically. P10 proposed multi-step disclosure that expands from simple prompts to detailed summaries.

D_{r2}: Visualization of data flows and risks. This design fosters reflection by transforming abstract data practices into concrete mental models (addressing Pr2, Pr4). Visual aids, such as P9's maps revealing travel patterns or P8's data flowcharts, bridge the gap between technical text and perceived reality.

D_{r3}: Contextualized scenarios and consequences. This design facilitates situated reflection by grounding data practices in relatable narratives rather than procedural descriptions. By illustrating specific outcomes (addressing Pr2, Pr6), such as P6's narrative warnings of "worst-case scenarios", the interface prompts users to simulate potential harms and engage in their consent's consequences.

D_{r4}: Conversational and guided exploration. By shifting the interaction from passive reading to active inquiry, this design supports a dialectic reflection process (addressing Pr4). Interface metaphors like P10's proactive chatbot, which suggests queries encourage users to formulate questions and actively explore the rationale behind data collection.

D_{r5}: AI-powered suggestion and auditing. This design empowers reflection by providing an external, evidentiary basis for

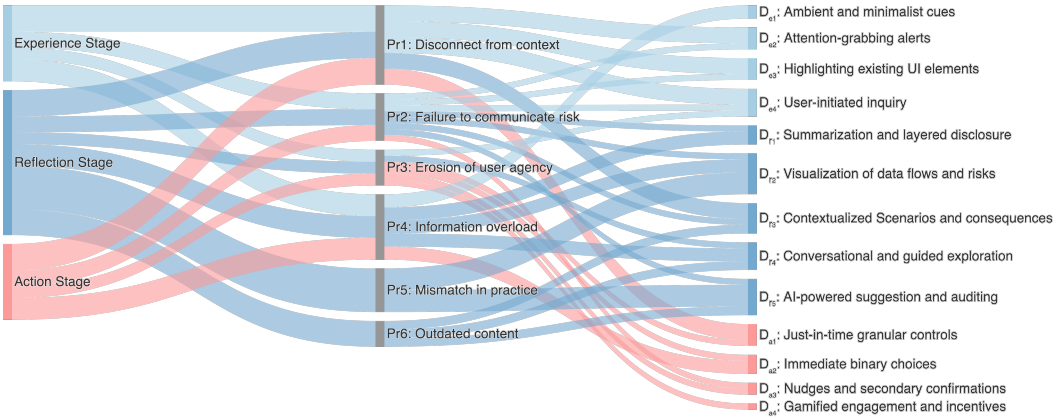


Figure 2: A sankey diagram showing the mapping relationship between the problems, the three reflective thinking stages, and each stage's consolidated design patterns derived from participants' designs. Different colors corresponded to different stages. The size of the problem node indicated its rated importance in Sec 3.3.3.

trust. It uses computational methods to analyze context, policy and app behavior to provide actionable recommendations (addressing Pr5, Pr6). For example, P7 proposed a proactive system that monitored an app's network traffic and compared it against its stated privacy policy.

Action Stage. The final stage facilitates the transformation from reflection to action. These designs resolve the *erosion of user agency* (Pr3) by providing actionable mechanisms to reduce identified risks.

D_{a1}: Just-in-time granular controls. These designs provide users with in-situ, fine-grained control options to immediately enact their reflections. P9's design, for example, offered tiered control options for photo uploads, such as removing metadata before uploading data.

D_{a2}: Immediate binary choices. This approach provides straightforward controls that are easy to understand and execute, facilitating quick decisions based on the reflection result. P10's design concluded its layered disclosure with an option to toggle a permission on/off.

D_{a3}: Nudges and secondary confirmations. This design is intended to prevent user error or impulsive decisions by prompting users to reconsider an action, especially one with high privacy risks, thereby deepening reflection. P1's design implemented a secondary confirmation pop-up that appeared after a user attempted to post with a high-risk location-sharing setting, forcing re-evaluation.

D_{a4}: Gamified engagement and incentives. This design uses game-like mechanics and external incentives, motivating users to engage with privacy information and controls, deepening reflection, and prompting behavior changes. For example, P8's design suggested using elements like a progress bar for reading privacy cards, or rewarding users with badges or discount coupons for completing a privacy-related quiz.

3.3.3 Prioritization of Problems and Designs. Table 2 presents the aggregated results of prioritization, where participants ranked the identified problems and the synthesized design patterns. The results indicate a strong alignment between the problems they deemed most critical, and their preferred design solutions. They identified the *disconnect between context and action* (score: 40) as the most critical issue, which is consistent with the existing literature [45]. This was followed by *information overload and poor readability* (score: 33) [48, 65] and *failure to communicate tangible risk* (score: 24). A fair level of consensus (Kendall's W = 0.31) suggests a shared view that lack of context is the primary problem of privacy policies.

Quantitative rankings reveal the nuanced trade-offs within the design space. In the *experience* stage, participants weighed the requirement for noticeability against the preservation of task continuity, favoring *highlighting existing UI elements* (score: 20) and *ambient and minimalist cues* (score: 19), where these approaches avoid the disruption associated with *attention-grabbing alerts* (score: 11) (Kendall's W=0.32). In the *reflection* stage, while *contextualized scenarios and consequences* was the primary choice for providing tangible risk awareness (score: 32), they also recognized the value of *summarization and layered disclosure* (score: 24) and *visualization of data flows and risks* (score: 23) as key strategies for balancing deep contextual detail with information clarity (Kendall's W=0.35). In the *action* stage, the preference for *just-in-time granular controls* (score: 23) over *immediate binary choices* (score: 18) or *nudges and secondary confirmations* (score: 14) reflects a prioritization of agency. Participants acknowledged that while granular control may involve higher interaction effort than simpler alternatives, it fosters informed autonomy (Kendall's W=0.35).

The transition from the design space to specific prototype was guided by a narrowing that emphasized user experience. In the *experience* stage, we implemented *ambient cues* to prioritize user agency and non-intrusiveness over the highly visible but disruptive nature of explicit alerts. In the *reflection* stage, the implementation emphasizes cognitive accessibility by utilizing *contextualized disclosure*. This choice prioritizes reducing the user's mental workload in processing complex risks over maintaining high information density. For the *action* stage, we prioritize minimal functional disruption by opting for *just-in-time nudges*, ensuring that the interface facilitates informed autonomy without forcing immediate executions that

might break the user's primary task flow. These choices align with participant-rated feasibility and effectiveness scores.

4 Conflect: Prototyping the Design Space

This section presents the design and implementation of Conflect, a high-fidelity prototype that instantiates the high-priority design patterns identified in Sec 3.

4.1 Interaction and Interface

To operationalize the Reflective Thinking Framework [20, 49, 54] in a mobile environment, Conflect guides users through a cyclical workflow of *Experience*, *Reflection* and *Action*. This workflow instantiates the preferred design patterns from our workshops (Table 2).

The cycle begins by disrupting habituation through *Ambient and Minimalist Cues* (D_{e1}). We implemented a floating sidebar that unobtrusively surfaces detected privacy risks using color-coded icons (Fig 3b). Following established sensitivity metrics [64], we adopted a color-coded schema (red for high risk, orange for medium, green for low) to provide immediate, intuitive feedback on data sensitivity. To balance visibility with non-intrusivenss, the sidebar can collapse into a minimal trigger (Fig 3a). When new risks are detected in this state, the trigger uses *Attention-Grabbing Alerts* (D_{e2}) via a subtle blinking animation, ensuring timely awareness without blocking the primary task.

Upon tapping a risk icon, Conflect first employs *Highlighting Existing UI Elements* (D_{e3}) to draw bounding boxes around relevant on-screen components. This visual intervention transitions users to the *Reflection Stage*, wherein the system delivers a two-step layered disclosure. The first notification, contextual summary, concisely summarizes key data practices (e.g., collection, sharing, disclosure), addressing the information needs identified in prior work [45, 65]. The second notification, reflective risk prompts, instantiates contextualized scenarios (D_{r3}, Fig 3c). It translates abstract policy terms into concrete, hypothesized risk scenarios (e.g., potential misuse of data), similar to empathy-based sandbox approaches [13]. These notifications are dynamically generated from the app's privacy policy (see Sec 4.2) to ground abstract risks in immediate contexts. To allow for optional deep reflection without forcing interaction, the initial notification fades after three seconds, while the risk prompt persists for five seconds. A tap on the prompt reveals the original policy excerpt (Fig 3d), supporting users who seek detailed verification.

The cycle concludes by empowering users to act on their reflections. A long press on a risk icon instantiates *Just-in-Time Controls* (D_{a1}, Fig 3e), allowing users to permanently mute specific risk triggers. This filtering mechanism reduces notification fatigue while preserving user agency. Although the current prototype cannot modify system permissions directly, it nudges users toward the host application's settings (D_{a3}), effectively bridging the gap between reflection and remediation.

4.2 Algorithm and Implementation

Conflect is implemented as a client-server system comprising 3 integrated modules: *Contextual Privacy Detection*, *Privacy Policy Extraction*, and *Reflective Presentation* (Fig 4). The mobile client, developed as an Android APK, uses the Android Accessibility Service API to capture on-screen content without requiring root access. To ensure system responsiveness and minimize local resource consumption, the client communicates with a backend server that handles heavy computational tasks and LLM reasoning. Processing is dynamically triggered only when the pixel-wise overlap between consecutive screenshots falls below 80% threshold (See Appendix F), ensuring that the analysis remains synchronized with UI changes while conserving computing power.

The *contextual privacy detection* module employs a two-channel pipeline to identify sensitive on-screen elements [45]. For textual data, we use PaddleOCR to localize coordinates, which are then classified into predefined sensitive categories (e.g., Location, Financial Information) using GPT-4o-mini. Simultaneously, graphical elements are processed via a ResNet-50 vision model [29] to classify icon candidates into the same unified categorization. This dual-channel approach yields precise visual anchors, specifically pixel-wise bounding boxes, that serve as the foundation for rendering the highlighting overlays described in Sec 4.1.

For *privacy policy extraction*, the system identifies the target application during installation or launch to retrieve its official privacy policy. We use GPT-4o to fetch and segment the policy text, extracting structured data practices regarding collection, transmission, and sharing. The extraction process uses a keyword-based mapping adapted from prior work [45]. By using their original data category mapping, this approach covers approximately 80% of common policy topics as showed by Pan et al. [45]. To reduce runtime latency and ensure instantaneous user experience, extraction and analysis are performed and cached during installation setup.

The *reflective presentation* module aligns detected UI elements with the cached policy segments by matching their data categories. Using the aligned inputs, GPT-4o synthesizes context-specific risk descriptions and concise notices. These outputs are generated using structured role-playing prompts [57] that include four components: role definition, task specification, context (i.e., policy text and categories), and strict constraint enforcement. The role definition asks LLMs to act as privacy policy analyst (to improve accuracy), and reflective design assistant (to help with reflection), similar to prior work [14, 69]. Specifically, we enforce contextual grounding by embedding constraints within the system prompt. These instructions restrict the model to derive risk scenarios from the provided policy excerpts, and to report ambiguity rather than hypothesize worst-case outcomes when data practices are vague. The next subsection provide a technical evaluation of the effectiveness of privacy policy extraction and scenario generation.

4.3 Technical Evaluation

We focused on privacy policy extraction accuracy and risk scenario appropriateness to assess Conflect's effectiveness. For privacy policy extraction, we used three established datasets containing diverse English and Chinese mobile policies: CPP4APP [45], CA4P-483 [67], MAPP Corpus [2]. As a preliminary assessment of the technique's effectiveness, we benchmarked three state-of-the-art LLMs: GPT-4o, GPT-5 and Claude-4.5-Sonnet. We aligned our data extraction categories with those defined in CPP4APP. Each data item was processed three times to account for model variability, and results were manually verified by researchers for accuracy and comprehensiveness. The results showed high overall extraction accuracy

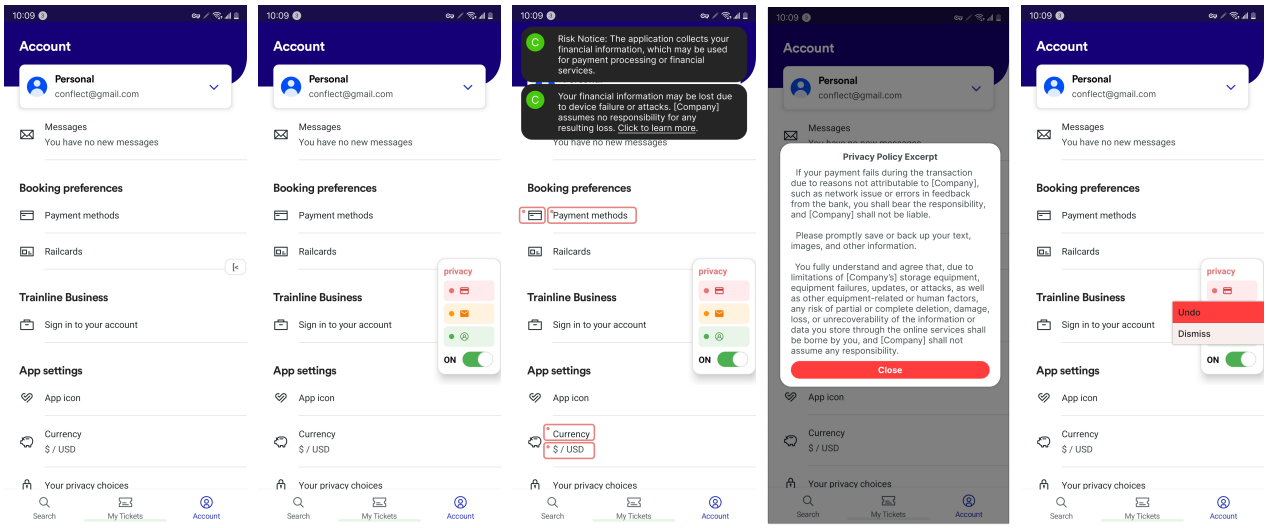


Figure 3: Illustration of Conflect, a reflective-thinking-based CPP prototype that dynamically surfaces interaction risks. Its workflow has three stages. *Experience (a–b)*: Emerging risks are signaled by a red-bordered lateral alert of the collapsed sidebar. Expansion reveals a floating menu with color-coded severity indicators. *Reflection (c–d)*: Tapping icons initiates layered disclosure, comprising UI bounding boxes, scenario-based risk descriptions, and original policy excerpts, to facilitate user inquiry. *Action (e)*: Users could manually adjust permissions in-situ or permanently dismiss specific triggers via long-press. Alerts recur by default for persistent awareness.

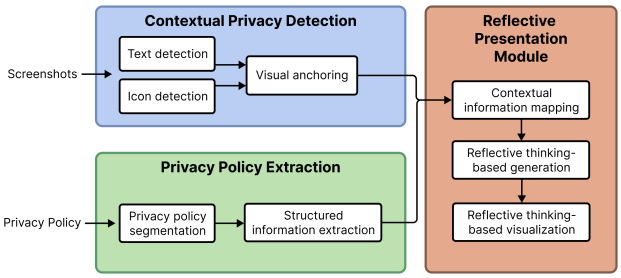


Figure 4: The algorithmic pipeline of Conflect, consisting of *contextual privacy detection*, *privacy policy extraction*, and the *reflective presentation* module.

across all datasets: 96.4% for GPT-5, 93.1% for Claude-4.5-Sonnet, and 91.6% for GPT-4o. While accuracy for common identifiers like location and contacts frequently exceeded 95% for GPT-5, performance was slightly lower for complex categories such as financial information (90.0%-94.5%). Error analysis revealed that while false negatives (averaged 4.0%, i.e., omissions due to highly ambiguous policy phrasing) occurred occasionally, false positives (averaged 2.2%, i.e., hallucinated or mischaracterized risks) were remarkably rare. These false positives could function as educational triggers that prompted users to critically verify data practices than inducing unwarranted alarm or eroding system trust.

To evaluate the appropriateness of synthesized risk scenarios across the above three datasets, we conducted an IRB-approved annotation study with 30 human annotators. Recruited via university networks and online communities to represent active mobile users, these participants evaluated the scenarios on a 7-point Likert scale (1="not at all", 7="very much") based on whether the description would prompt conscious reflection. The overall rating exceeded 5.6 out of 7 across all datasets, with CPP4APP received the highest rating, averaging 6.4, while CA4P-483 and MAPP scored 5.8 and 5.7 respectively. Highly sensitive categories such as *social media* and *contacts* have scores over 6.0, confirming that generated scenarios highlight meaningful privacy risks to users.

5 Evaluation of Conflect Prototype

5.1 Study Design

We adopted a within-subjects design with ***technique*** as the single factor. This design was chosen to facilitate direct comparisons between techniques while controlling for individual differences in reading speed and comprehension [43]. We compared Conflect with a state-of-the-art CPP [45] to isolate the reflective-thinking-based framework and test the added value of the reflective designs. As in Figure 10, we replicated the user interface and system of Pan et al. [45], which contextually overlays bounding boxes and privacy policy text for users.

We selected four popular apps across different domains (travel, entertainment, browsing, e-commerce): Ctrip, iQIYI, UC Browser and GuaZi used car platform. For each app, we designed a multi-step task representative of the app's primary use, such as booking flights

on Ctrip. The tasks guide participants through goal-oriented interactions that naturally elicit diverse privacy-sensitive data access events, ranging from inputting personally identifiable and financial information to location tracking and personalized ad settings (see Appendix E). We measured the following aspects to investigate their understanding, experience, trust, and cognitive load as they interacted with the privacy policy information presented via either Conflect or CPP during task completion:

Reflective thinking [34]: We assessed users' comprehension of reflective thinking using Kember's questionnaire, which is widely used in prior user studies [37, 39, 63]. It contains four sub-dimensions: habitual action, understanding, reflection and critical reflection. Although there are alternatives like TSRI [5], we did not choose TSRI as it targets systems instead of the underlying cognitive processes.

Understanding of privacy policy [65]: We assessed users' comprehension of data practices through six-question quiz, following validated methods from prior work [58, 65]. The questions covered key aspects like data collection, usage and sharing. Experimenters manually designed and cross-checked answers to ensure that they are answerable with CPPs, where alternative incorrect options referred to policies in other apps.

System usability and user experience [14, 65]: assessed using the User Experience Questionnaire Short Version (UEQ-S) [36] and System Usability Scale (SUS) [11].

Trust: we measured the trust in both the app and each technique via following 7-point Likert scales (1=very unlikely to agree, 7=very likely to agree) "I trust the original app about its data practice", and "I trust the content that the technique presented to me".

Cognitive load [65]: measured through adapting NASA-TLX's six dimensions [28].

The study concluded with a semi-structured interview to gather insights into user experience, comprehension, privacy awareness, and decision-making processes for each technique.

5.2 Participants and Recruitment

We conducted an a priori power analysis, which indicated that a minimum sample of N=34 was required to detect a medium effect size (d=0.5) with 80% power (α=0.05) for a within-subjects design. To compensate for the potential non-sphericity, we recruited 48 Chinese participants (20 males, 28 females, M_{age}=25.3, SD=8.4, age range 18-49) through posters distributed across multiple WeChat[2] groups. These groups target diverse university-affiliated, professional, and general-interest communities. 5 have a master's degree, 6 have a Ph.D.'s degree, 34 have a bachelor's degree, and others were with high school's degree or below. 3 are from Computer Science related backgrounds, and 11 are from Engineer backgrounds. Others are from multiple disciplinary. To enhance the ecological validity of our study, consistent with prior practice [65], each participant used their own smartphones to interact with two settings, which was provided by the researchers as APKs. To control for confounding variables, we checked and ensured that participants were active users of the target applications but self-reported having never read their privacy policies. This study got the approval of our IRB, and each participant was compensated 100CNY according to the local wage standard. The study (e.g., interface, apps) was conducted using Chinese.

[2]a predominant social platform in China, which has billions of users.

5.3 Procedure

After a briefing on the study's objectives, participants provided informed consent. The experiment consisted of two sessions to compare Conflect and CPP. To address sequence effects and application-specific biases, we fully counterbalanced both the presentation order of the techniques and their assignment to the target application pairs (e.g., Ctrip and iQIYI) across all participants. Participants executed the assigned tasks at their own pace without time restrictions to maintain ecological validity. However, to ensure a fair comparison of comprehension and cognitive load, participants were instructed to read each presented policy visualization in its entirety before proceeding. This protocol was implemented to prevent skimming of lengthy policies, a behavior known to confound comprehension metrics [65]. Observations confirmed that all participants adhered to this instruction. Each session concluded once the participant completed the assigned task. After each session, participants filled in the questionnaires. Once both sessions were finished, they engaged in semi-structured interviews.

5.4 Data Analysis

Given that the reflective thinking and usability ratings violated normal distribution, we employed paired Wilcoxon signed-rank tests, and rank-biserial correlation to measure effect size. For the qualitative data, we conducted thematic analysis on the transcribed interviews. To ensure interpretative validity, we adopted a collaborative coding strategy rather than independent parallel coding. Initially, 2 researchers jointly reviewed 4 randomly selected transcripts to develop a preliminary codebook. They then collaboratively coded the remaining scripts, iteratively refining the codes and resolving disagreements through regular discussions. The final codebook is provided in Appendix A.

5.5 Results

We first evaluated whether users' ratings diverged across application contexts and data types. We performed between-subjects comparison across applications, which found no significant differences in all subjective ratings ($p > .05$), suggesting that the reflective design maintains consistent effectiveness regardless of the app context. Furthermore, user interviews revealed no distinction for Conflect's effectiveness across application and data categories. Therefore, we combined application and data contexts later.

5.5.1 *Reflective Thinking.*

Subjective Ratings. Figure 5 presents the reflective thinking ratings for *Conflect* and *CPP*. Analysis revealed that *Conflect* elicited significantly better scores in *habitual action* ($W = 13.0, p = .003 < .01$) and in *understanding* ($W = 22.0, p = .048 < .05$) compared to *CPP*. This indicates that *Conflect* effectively mitigated automatic interaction patterns (i.e., habitual clicking) and improved comprehension of privacy practices. In contrast, no significant differences were observed between the two techniques for *reflection* ($W = 49.0, p = .321$) and *critical reflection* ($W = 47.0, p = .158$). This suggests that while *Conflect* successfully disrupted habitual

behaviors and enhanced understanding, participants did not exhibit a significant increase in conscious internal reflection relative to *CPP*, potentially due to study's short duration.

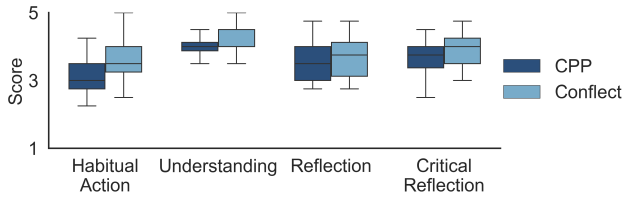


Figure 5: Boxplot of reflective thinking ratings for CPP and Conflect.

Subjective Feedback. Participants attributed their behavioral shifts to the framework's interaction paradigm than detailed designs.

From experience to reflection. Many participants reported moments of surprise during the experience stage, which served as the primary trigger for reflection on privacy. Their common assumption was that data collection would occur only at explicit input points or after visible permission requests. When Conflect revealed data access outside these moments, participants described a mismatch between expectation and observed behavior. Several participants were surprised to learn that data collection occurred through seemingly innocuous interactions. P13 explained, *"I used to think that these applications would ask every time... but actually they might be collecting my information the moment I click a certain button, and I only learned that through using this mini-program."* Similarly, P6 was astonished to notice apps collected data even when *"clicking on places where I obviously don't need to input content."* P18 echoed this, noted *"originally didn't know that it was still calling my information in some places where I didn't need to enter information."* These surprises prompted participants to reconsider both the content and timing of collection, breaking habituation by shifting their mental models to perceive data access as continuous and contextual.

From reflection to action. While reflection did not always result in immediate behavioral change, participants described distinct strategies for responding to newly surfaced privacy concerns during the reflective process.

Active defensive strategies. A subset of participants translated privacy reflection into protective strategies, primarily by minimizing exposure rather than disengaging. Regarding inputs, P14 replaced real data with fabricated entries, explaining this *"wouldn't affect the use of other functions while preventing privacy being excessively read"*, while P7 reconsidered prompts to provide partial information rather than all of it. Regarding permission agency, P14 realized permissions could be actively turned off, and P37 articulated a strategy of selectively enabling permissions and revoking them afterward. These indicate that reflection supported both immediate exposure-minimizing actions and long-term strategies.

Deliberation and hesitation without direct action. More commonly, reflection primarily reshaped evaluation rather than immediate behavior, manifesting as hesitation or increased attentiveness. P25 recalled feeling that they *"didn't really want to continue these operations when notified about access to the camera or microphone, yet proceeded while trying to pay a little more attention."* Others articulated nuanced reasoning, imagining different decisions under altered circumstances. P23 noted that while the risk in the current situation felt acceptable, a warning that the behavior was very serious would likely have caused them to stop. Similarly, P24 reflected that awareness of risk ratings made them more cautious when considering future disclosure of sensitive information, such as phone numbers or identification details. Thus, reflection altered the evaluative stance even when action was not taken.

Identifying barriers to action. Reflection also surfaced why action felt impossible or futile for some participants, leading them to articulate structural and situational barriers that blocked behavioral change. Several participants explicitly reasoned through why they could not act on newly surfaced concerns, pointing to task utility, lack of alternatives, and the perceived inevitability of data collection. P38 emphasized that completing the task took precedence, explaining that *"I focus more on finishing the task, and privacy is secondary."* The same participant later remarked that awareness was not very helpful because they *"had no way to change the application's behavior"*. This resignation was echoed by others, such as P46, who stated bluntly, *"even if you know it's collecting, you still have to use this app."*

These findings indicate that reflection leads to context-sensitive responses, where action is shaped by awareness and task demands. Even when constraints prevented immediate behavioral change, reflection successfully altered their underlying mental models.

5.5.2 *Usability.*

Objective Policy Comprehension. To evaluate users' objective comprehension of the privacy policies, the post-session questionnaires included knowledge-based questions testing key policy details. Participants' performance was measured by their accuracy, calculated as the percentage of correct answers. As shown in Figure 6, we found significant difference of users' objective comprehension between *Conflect* and *CPP* ($W = 9.0$, $p = .026 < .05$). This suggests that *Conflect* improved both objective comprehension and subjective understanding of privacy practices, as detailed in Sec 5.5.1.

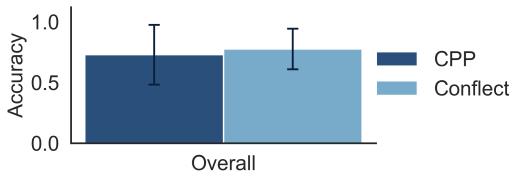
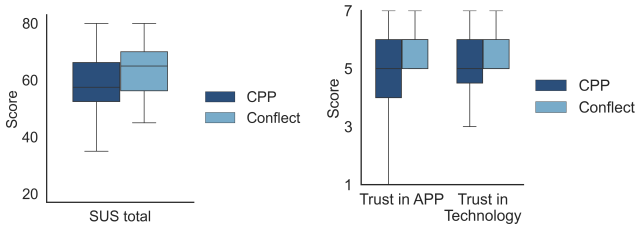


Figure 6: Users' objective policy comprehension (percentage of correct answers) of privacy policies with CPP and Conflect.

User Perceptions. Our analysis of subjective ratings reveals that *Conflect* significantly enhances the user experience, particularly in terms of usability and cognitive load, when compared to *CPP*. Regarding system usability (see Figure 7a), participants rated *Conflect* as usable (M=61.7), significantly outperforming *CPP* ($W = 9.0$, $p = .026 < .05$). Regarding trust, we observed significant main effects for trust in the application ($W = 10.0$, $p < .05$), but not in

the technology ($W = 21.0$, $p = .279$) when comparing *CPP* with *Conflect* (Figure 7b). This indicates that *Conflect* improves the interaction quality, altered the user's trust in the underlying app, however has no effect on users' trust towards CPP.



(a) SUS. **(b) Trust.**

Figure 7: The (a) SUS and (b) trust score across different techniques. Higher rating indicates more positive outcomes.

An important objective of *Conflect* is to enable reflection without compromising usability or increasing user workload. We found no significant differences between the techniques across all six workload dimensions: *mental demand* ($W = 60.0$, $p = .432$), *physical load* ($W = 33.5$, $p = .665$), *temporal load* ($W = 45.5$, $p = .243$), *performance* ($W = 58.5$, $p = .932$), *effort* ($W = 54.5$, $p = .754$), and *frustration* ($W = 63.0$, $p = .521$). These findings indicate that Conflect reduces habituation without inducing extra loads.

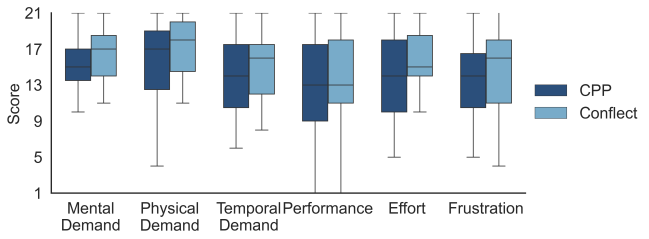


Figure 8: Cognitive load ratings across techniques, with all dimensions reverse-coded only for illustration. In all cases, higher rating indicates more positive outcomes.

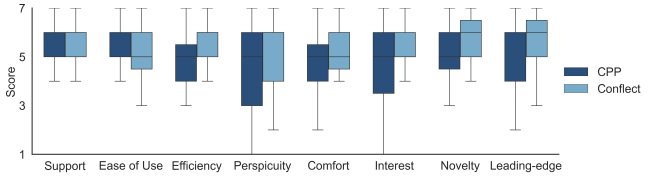


Figure 9: Boxplot of UEQ-S ratings across different techniques (1: most negative, 7: most positive).

The UEQ-S ratings further show that *Conflect* no only maintains a comparable experience, but outperforms *CPP* in specific dimensions. While rated as comparable against *CPP* in providing *support* ($W = 22.0$, $p = .952$), *ease of use* ($W = 45.5$, $p = .653$), *efficiency* ($W = 25.0$, $p = .144$), *perspicuity* ($W = 37.0$, $p = .542$), *comfort* ($W = 34.0$, $p = .225$), and *leading-edge* ($W = 20.5$, $p = .802$), it outperformed *CPP* in the aspects of *interest* ($W = 4.0$, $p = .021 < .05$), and marginally outperformed *CPP* in *novelty* ($W = 7.0$, $p = .058$). This showed that the reflective thinking-based design elicited greater user interest, and fostered a stronger adoption attention.

Subjective feedback. ***Usability and information comprehension.*** Participants reported that Conflect is comparably easy to use compared to existing CPPs, while effectively delivering reflective privacy information. The tool delivered contextualized information while maintaining a manageable user experience. P40 highlighted this balance, noting, *"It is very convenient, unlike the [other] CPP which displays a large amount of information."* This preference for streamlined disclosure was further reflected in the interface design. Users favored the floating sidebar over aggressive full-screen alternatives, which P4 found *"a bit troublesome"*. P18 described how the design supports information seeking, explaining, *"I think it can concisely convey information to me in the form of an icon ... it quietly reminds me there without disturbing me, and when I want to check it, it can provide me with information."* Beyond the structure, visual salience enhanced comprehension and reflection. P6 found the color-coded risk icons featured high recognizability, allowing them to *"instantly realize what kind of private information is being accessed here."* The summaries also proved important for reflection on policy content. P18 appreciated how the tool summarize text to focus on the key points, while P44 noted it clarified their previously skipped complex terms.

System trust and adoption. Willingness of long-term adoption was tied to user trust. Participants expressed a strong willingness to use Conflect if it were natively integrated into mobile operating systems by manufacturers like Apple, Android, or Xiaomi. P18 provided a practical rationale centered on accountability, *"When the software has a problem, who is responsible for it? ... If it is integrated into Xiaomi, then I know you have been recognized by Xiaomi. When it has a problem, I can not only look for you but also for Xiaomi."*

Limitations and expectations. While overall feedback was positive, participants identified areas for improvement. P33 commented that, *"I had to spend some time understanding the meaning of each symbol ... without prior instructions, I wouldn't know what it means."* Similarly, P17 stated, *"I couldn't quite understand those icons ... I had to click on them to see the specific introduction."* Users also proposed proactive features. P18 requested an abnormality alert for when an app deviates from industry norms, P14 recommended direct permission control within CPPs instead of manual settings, and P10 wanted more actionable guidance to avoid privacy risks for the next step. Finally, participants emphasized the need for customizable privacy thresholds. P15 desired a system for customizing the level of privacy to trigger prominent alerts for highly sensitive data.

6 Discussions

6.1 Reflective Thinking

Our work systematically operationalizes reflective thinking framework to transform CPP into active cognitive instruments that guide users through the ERA cycle.

Conflect's **Experience** stage disrupts habituation [31] through design friction [16]. Our study in Sec 3.3.1 identified that the disconnect between abstract policy and concrete context is a primary barrier to awareness, prompting the design of ambient yet noticeable cues that function as triggers. As in Sec 4.1, Conflect operationalizes this by using color-coded sidebars and visual anchoring to create a deviation from user expectations regarding background data collection. The efficacy of this mechanism is empirically supported by our qualitative findings in Sec 5.5.1, where participants reported moments of surprise upon realizing that data collection occurred during seemingly innocuous interactions. This validates Schön's theories that reflection is initiated by a surprise that interrupts habitual behavior [53]. This successfully shift users from mindless interaction to a state of alert readiness, without inducing cognitive overload (Fig 8).

Following the initial trigger, the **Reflection** stage scaffolds inquiry by bridging the semantic gap between technical data flows and implications. Addressing the problem that users fail to perceive tangible risks in legalistic text in Sec 3.3.1, Conflect translate abstract policies into context-aware risk scenarios, as in Sec 4.2. This design directly supports the reflective requirement for "conscious inquiry" by converting procedural descriptions into tangible consequences. Participants in Sec 5.5.1 also noted that the synthesized summaries clarified complex terminology and highlighted key risks, prompting them to reflect on privacy practices they would ordinarily ignore. This shows that effective reflection stems from the disclosure's contextual relevance than information volume.

The cycle concludes with the **Action** stage, translating reflection into informed control. In Sec 3.3.1 we highlighted the erosion of user agency as a critical issue, and as in Sec 4.1, Conflect restores this by coupling risk awareness with actionable mechanisms such as muting alerts, or navigating to system settings. However, consistent with prior work on privacy interfaces [45, 62, 65], we prioritized user comprehension over direct "just-in-time" blocking controls. This decision reflects a *privacy-utility trade-off*: granular controls often compel users to choose between privacy and functionality, leading to decision fatigue [15] or the prioritization of convenience [66]. Despite these constraints, Conflect successfully fostered distinct behavioral shifts, such as data fabrication described in Sec 5.5.1. Crucially, even when users chose not to change settings, their decision represented conscious deliberation rather than privacy resignation [55]. Thus, by acting akin to a *contextual auditing tool* [14, 45], Conflect could provide granular individual agency and foster transparency around data practices.

In the experiment, while baseline CPPs effectively deliver information via in-situ visual overlays, the observed reduction in habitual actions within Conflect is driven by the underlying reflective thinking frameworks rather than specific interface designs. The *experience* stage intentionally induces a cognitive surprise that breaks habituation by exposing a temporal mismatch between users' mental models and actual data practices. From feedback in Sec 5.5.1, participants also noticed this unexpected background data collection during interaction. This structural timing, rather than visual salience, acts as the primary factor shifting users toward active deliberation. Furthermore, the transition into the *reflection* stage alters users' evaluative stances by forcing conscious risk assessment. Even when without granular controls, as in Sec 5.5.1, participants highlighted that the reflective process itself induces mindful hesitation. Therefore, while summarized risks and contextual bounding boxes facilitate readability, as reflected in Sec 5.5.1, participants found the structuring of interactions into a cycle of contextual surprise and guided inquiry effectively disrupts automatic clicking and reduces privacy resignation.

6.2 Technical Considerations and Threat Model

Conflect relies on the Android Accessibility Service, which needs broad permissions and introduces a privacy paradox when transmitting sensitive contextual data to cloud LLMs. To mitigate this, we enforce data minimization by transmitting only extracted text rather than raw screenshots. Furthermore, to ensure the appropriateness of the generate explanations and mitigate risks of LLM hallucinations, we employ rigorously constrained role-playing prompts, as in Appendix H. By anchoring the model's reasoning strictly to the provided policy text, we prevent exaggerated scenarios that over-alarm users or erode system trust. Our technical evaluation confirms this pipeline's reliability, yielding up to 96.4% extraction accuracy with negligible false positives.

However, for practical deployment, a cloud paradigm remains suboptimal. Future efforts could migrate Conflect, especially the *Contextual Privacy Detection* and *Privacy Policy Extraction* modules entirely to local execution. The rapid advancement of on-device Small Language Models (SLMs) and system-level trusted AI integrated directly by OS vendors (e.g., Apple Intelligence, Huawei Harmony OS) presents a viable path forward. By processing context and generating risk scenarios strictly on-device, Conflect can maintain the cognitive benefits of reflective design, while eliminating the need for third-party data transmission, thereby resolving the privacy-utility tension of the current prototype.

6.3 Generalizability of Design

We scoped our tests for Android apps (see Fig 3) and deliberately excluded apps that operate primarily as background services, such as malware scanners and junk cleaners, as they lack the interactive elements necessary to trigger contextual notices.

While our implementation is Android-specific, the designs of reflective CPPs are potentially generalizable. It could be translated to iOS ecosystems through translating prototype to swift language. Besides, our method may be adapted for websites on mobile phones where the site's name, and interface icons are discernible [44, 62]. "Microapps", such as those within the WeChat ecosystem, represents another potential application. Although these third-party services have limited permissions, they still pose privacy risks when users input sensitive data. While beyond the current focus of Conflect, our framework's core reflective thinking-based design could be extended to these interfaces. However, extending this to desktop websites will demand significant efforts, probably following the methods by Windl et al. [62].

While our user study was conducted within a Chinese context, the CPP design space and Conflect is designed to be language-agnostic. Our results show that the system's reliance on English backend prompts does not hinder its ability to process non-English content. However, extending Conflect to other linguistic environments needs further validation.

6.4 Design Implications

Based on the CPP design space and our empirical findings, we propose implications across the *experience*, *reflection*, *action* stages.

Experience: Implementing strategic design friction to avoid habituation. To reduce habitual clicking, designers could introduce lightweight, contextual forms of friction that vary in salience based on the severity of privacy risks. Examples include UI-anchored bounding boxes or blinking sidebars that appear when sensitive data are accessed.

Reflection: Scaffolding reflection through semantic risk translation. Designers could bridge the gap between policy text and user actions by translating legal policy segments into situated risk scenarios. Using LLMs to describe contextual risks can transform passive reading into active inquiries. For example, interfaces could present layered, grounded summaries such as *"It collects your financial information for payment processing. However, its policy states it assumes no responsibility for any resulting loss due to device attacks. Click to verify the original policy excerpt."*

Action: Restoring user agency via just-in-time control. Since users often prioritize tasks over privacy, reflection should be coupled with immediate and low-effort control mechanisms. Designers could implement just-in-time granular controls at the precise moment of data access. This can be achieved with a user interception layer at the UI level that offers localized defensive options, such as temporary "one-time" permissions or data minimization prompts.

7 Ethics Considerations

All research protocols involving participants were reviewed and approved by the IRB of the authors' university. Before participation, all participants provided explicit informed consent after receiving a detailed briefing on the study's objectives and data handling procedures. Participants were compensated according to the local wage standard, with 350 CNY for the workshop, 100 CNY for the user study, and 200 CNY for the technical evaluation. Technically, the Conflect prototype utilizes the Android Accessibility Service API to capture on-screen content. During the consent process, we explicitly detailed the privacy risks associated with the service, noting that it performs automatic UI scraping, possesses the potential to access highly sensitive personal data, and captures real-time user interactions. To ensure user privacy and system security, we excluded banking, finance and password management applications from the study. Besides, data is processed via a secure client-server architecture where analysis is triggered dynamically based on UI changes, minimizing unnecessary data exposure. For the automated privacy policy extraction and risk scenario generation with LLMs, we employed strict constraints on the role-playing prompts to ensure the accuracy and relevance of the notices.

While Conflect aims to enhance privacy awareness, its usage of the Android Accessibility Service and cloud-based LLMs relies on trust assumptions. Currently, captured UI and contextual data are transmitted to a backend institutional server for ephemeral processing. To mitigate third-party exposure, we ensure that only recognized textual or transcription data is transmitted to commercial API providers, and we requested the deletion of these data via APIs. This approach relies on the privacy guarantees of commercial API providers (e.g., OpenAI) to mitigate exposure. To further adhere to the principle of data minimization and reduce privacy risks, future iterations should transition toward on-device processing. Shifting privacy policy extraction and contextual mapping to local small language models is a critical step before deployment.

8 Limitations and Future Work

We acknowledge several limitations of our paper. First, our design space was elicited from domain experts, mainly authors of recent privacy policy visualization publications to ensure high-quality design elicitation. However, this reliance on specific academic and professional networks may skew the design space, which overlook broad industrial and cultural perspectives. The user study also primarily involved young Chinese participants. While many participants in our user study had international backgrounds, this sample may not fully capture privacy concerns of diverse age groups and cultures. Future work should validate our findings with cross-cultural populations.

Second, while Conflect enhances contextual awareness, certain parts of privacy policies may not be covered by CPPs [45]. Conflect currently acts as a complement to, rather than a replacement for full privacy policies. A possible workaround could be to augment CPP with traditional privacy policy at the installation time [48, 65].

Third, Conflect currently relies on LLMs to sort out privacy risks, which may lead to hallucinated or mischaracterized risk explanations. Such technical unreliability not only undermines user trust but also exacerbates alert fatigue. A high frequency of false positives or trivial warnings can diminish user sensitivity, eventually reducing the system's capacity to disrupt habitual behavior. Furthermore, due to constraints in system-level integration, Conflect currently lacks the ability to execute data-flow blocking or sensitivity-tailored notifications. Future work could incorporate deterministic verification to ensure reliable risk communication, and implement system hooks for real-time privacy management.

Finally, our short-term evaluation is susceptible to novelty effects, which may artificially inflate initial engagement before habituation re-emerges. However, Conflect achieved a significantly greater reduction in habitual actions compared to an alternative CPP, which was also novel to participants. This suggests that the observed improvements are likely driven by the reflective framework rather than mere novelty. Nevertheless, longitudinal field studies are needed in the future to evaluate how these reflective actions perform in everyday contexts.

9 Conclusion

This research presents a design space for mobile CPPs based on the Reflective Thinking Framework. Three expert workshops revealed that the primary barrier to privacy policy engagement is the disconnect between abstract legal text and user actions. To bridge this gap, we developed Conflect, a prototype that uses LLMs to translate data practices into contextualized risk scenarios following the ERA cycle. A user study (N=48) shows that compares to state-of-the-art CPPs, Conflect significantly disrupts habitual interactions and improves privacy comprehension without increasing cognitive demand or compromising usability.

Acknowledgments

This work is supported by the Natural Science Foundation of China (NSFC) under Grant No. 62132010. The authors utilized generative AI (ChatGPT, Gemini) to polish the manuscript's text, focusing on improving readability, and correcting minor typographical and grammatical inconsistencies.

References

[1] Paul C Adams. 2020. Agreeing to surveillance: Digital news privacy policies. *Journalism & Mass Communication Quarterly* 97, 4 (2020), 868–889.
[2] Siddhant Arora, Henry Hosseini, Christine Utz, Vinayshekhar Bannihatti Kumar, Tristan Dhellemmes, Abhilasha Ravichander, Peter Story, Jasmine Mangat, Rex Chen, Martin Degeling, et al. 2022. A tale of two regulatory regimes: Creation and analysis of a bilingual privacy policy corpus. In *13th International Conference on Language Resources and Evaluation Conference, LREC 2022*. European Language Resources Association (ELRA), 5460–5472.
[3] Julia Bailey. 2008. First steps in qualitative data analysis: transcribing. *Family Practice* 25, 2 (2008), 127–131.
[4] Erik P. S. Baumer. 2015. Reflective informatics: Conceptual dimensions for designing technologies of reflection. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. ACM, 585–594.
[5] Marit Bentvelzen, Jasmin Niess, Mikołaj P Woźniak, and Paweł W Woźniak. 2021. The development and validation of the technology-supported reflection inventory. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–8.
[6] Marit Bentvelzen, Paweł W Woźniak, Pia SF Herbes, Evropi Stefanidi, and Jasmin Niess. 2022. Revisiting reflection in hci: Four design resources for technologies that support reflection. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 1 (2022), 1–27.
[7] Mike Bergmann. 2008. Testing privacy awareness. In *IFIP Summer School on the Future of Identity in the Information Society*. Springer, 237–253.
[8] Jarni Blakkarly and Daniel Graham. 2022. Privacy policy comparison reveals half have poor readability. https://www.choice.com.au/consumers-and-data/protecting-your-data/data-laws-and-regulation/articles/privacy-policy-comparison
[9] Davide Bolchini, Qingfeng He, Annie I Antón, and William Stufflebeam. 2004. "I need it now": Improving website usability by contextualizing privacy policies. In *International Conference on Web Engineering*. Springer, 31–44.
[10] Virginia Braun and Victoria Clarke. 2021. Thematic analysis: A practical guide. (2021).
[11] John Brooke et al. 1996. SUS-A quick and dirty usability scale. *Usability Evaluation in Industry* 189, 194 (1996), 4–7.
[12] Esther Cameron. 2005. *Facilitation made easy: practical tips to improve meetings and workshops*. Kogan Page Publishers.
[13] Chaoran Chen, Weijun Li, Wenxin Song, Yanfang Ye, Yaxing Yao, and Toby Jia-Jun Li. 2024. An empathy-based sandbox approach to bridge the privacy gap among attitudes, goals, knowledge, and behaviors. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–28.
[14] Chaoran Chen, Daodao Zhou, Yanfang Ye, Toby Jia-jun Li, and Yaxing Yao. 2025. Clear: Towards contextual llm-empowered privacy policy analysis and risk generation for large language model applications. In *Proceedings of the 30th International Conference on Intelligent User Interfaces*. 277–297.
[15] Hanbyul Choi, Jonghwa Park, and Yoonhyuk Jung. 2018. The role of privacy fatigue in online privacy behavior. *Computers in Human Behavior* 81 (2018), 42–51.
[16] Anna L Cox, Sandy JJ Gould, Marta E Cecchinato, Ioanna Iacovides, and Ian Renfree. 2016. Design frictions for mindful interactions: The case for microboundaries. In *Proceedings of the 2016 CHI conference extended abstracts on human factors in computing systems*. 1389–1397.
[17] Lorrie Cranor. 2002. *Web privacy with P3P.* " O'Reilly Media, Inc.".
[18] John Dewey. 2022. *How we think*. DigiCat.
[19] Nora A Draper. 2017. From privacy pragmatist to privacy resigned: Challenging narratives of rational choice in digital privacy debates. *Policy & Internet* 9, 2 (2017), 232–251.
[20] John Driscoll. 1994. Reflective practice for practice. *Senior Nurse* 13, 1 (1994), 47–50.
[21] John Driscoll. 2006. *Practising clinical supervision: A reflective approach for healthcare professionals*. Elsevier Health Sciences.
[22] Denis Feth. 2017. Transparency through contextual privacy statements. In *Mensch und Computer 2017-Workshopband*. Gesellschaft für Informatik eV, 10–18420.
[23] Carlos Flavián and Miguel Guinalíu. 2006. Consumer trust, perceived security and privacy policy: three basic elements of loyalty to a web site. *Industrial management & data Systems* 106, 5 (2006), 601–620.
[24] Rowanne Fleck and Geraldine Fitzpatrick. 2010. Reflecting on reflection: framing a design landscape. In *Proceedings of the 22nd conference of the computer-human interaction special interest group of australia on computer-human interaction*. 216–223.
[25] Leo A Goodman. 1961. Snowball sampling. *The Annals of Mathematical Statistics* (1961), 148–170.
[26] Lars Hallnäs and Johan Redström. 2001. Slow technology–designing for reflection. *Personal and ubiquitous computing* 5, 3 (2001), 201–212.
[27] Hamza Harkous, Kassem Fawaz, Rémi Lebret, Florian Schaub, Kang G Shin, and Karl Aberer. 2018. Polisis: Automated analysis and presentation of privacy policies using deep learning. In *27th USENIX Security Symposium (USENIX Security 18)*. 531–548.
[28] Sandra G Hart and Lowell E Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in Psychology*. Vol. 52. Elsevier, 139–183.
[29] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 770–778.
[30] Melanie Jasper. 2003. *Beginning reflective practice*. Nelson Thornes.
[31] Farzaneh Karegar, John Sören Pettersson, and Simone Fischer-Hübner. 2020. The dilemma of user engagement in privacy notices: Effects of interaction modes and habituation on user attention. *ACM Transactions on Privacy and Security (TOPS)* 23, 1 (2020), 1–38.
[32] Patrick Gage Kelley, Joanna Bresee, Lorrie Faith Cranor, and Robert W Reeder. 2009. A "nutrition label" for privacy. In *Proceedings of the 5th Symposium on Usable Privacy and Security*. 1–12.
[33] Patrick Gage Kelley, Lucian Cesca, Joanna Bresee, and Lorrie Faith Cranor. 2010. Standardizing privacy notices: An online study of the nutrition label approach. In *Proceedings of the SIGCHI Conference on Human factors in Computing Systems*. 1573–1582.
[34] David Kember, Doris YP Leung, Alice Jones, Alice Yuen Loke, Jan McKay, Kit Sinclair, Harrison Tse, Celia Webb, Frances Kam Yuet Wong, Marian Wong, et al. 2000. Development of a questionnaire to measure the level of reflective thinking. *Assessment & evaluation in higher education* 25, 4 (2000), 381–395.
[35] Spyros Kokolakis. 2017. Privacy attitudes and privacy behaviour: A review of current research on the privacy paradox phenomenon. *Computers & security* 64 (2017), 122–134.
[36] Bettina Laugwitz, Theo Held, and Martin Schrepp. 2008. Construction and evaluation of a user experience questionnaire. In *Symposium of the Austrian HCI and Usability Engineering Group*. Springer, 63–76.
[37] Hao-Ping Lee, Advait Sarkar, Lev Tankelevitch, Ian Drosos, Sean Rintel, Richard Banks, and Nicholas Wilson. 2025. The impact of generative AI on critical thinking: Self-reported reductions in cognitive effort and confidence effects from a survey of knowledge workers. In *Proceedings of the 2025 CHI conference on human factors in computing systems*. 1–22.
[38] Ian Li, Anind Dey, and Jodi Forlizzi. 2010. A stage-based model of personal informatics systems. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 557–566.
[39] Jingxian Liao, Fu-Yin Cherng, Mrinalini Singh, and Hao-Chuan Wang. 2025. Signals Beyond Text: Understanding How Accessing Peer Concept Mapping and Commenting Augments Reflective Mind for High-Stake Videos. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–21.
[40] Mariavittoria Masotina and Anna Spagnolli. 2022. Transparency of privacy notices and contextualisation: effectively conveying information without words. *Behaviour & Information Technology* 41, 10 (2022), 2120–2150.
[41] Nora McDonald, Sarita Schoenebeck, and Andrea Forte. 2019. Reliability and inter-rater reliability in qualitative research: Norms and guidelines for CSCW and HCI practice. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–23.
[42] Iain McLean, Arnold B Urken, and Fiona Hewitt. 1995. *Classics of social choice*. University of Michigan Press.
[43] Jonathan A Obar and Anne Oeldorf-Hirsch. 2020. The biggest lie on the internet: Ignoring the privacy policies and terms of service policies of social networking services. *Information, Communication & Society* 23, 1 (2020), 128–147.
[44] Anna-Marie Ortloff, Maximiliane Windl, Valentin Schwind, and Niels Henze. 2020. Implementation and in situ assessment of contextual privacy policies. In *Proceedings of the 2020 ACM Designing Interactive Systems Conference*. 1765–1778.
[45] Shidong Pan, Zhen Tao, Thong Hoang, Dawen Zhang, Tianshi Li, Zhenchang Xing, Xiwei Xu, Mark Staples, Thierry Rakotoarivelo, and David Lo. 2024. A {NEW}{HOPE}: contextual privacy policies for mobile applications and an approach toward automated generation. In *33rd USENIX Security Symposium (USENIX Security 24)*. 5699–5716.
[46] Sameer Patil, Roberto Hoyle, Roman Schlegel, Apu Kapadia, and Adam J Lee. 2015. Interrupt now or inform later? Comparing immediate and delayed privacy feedback. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*. 1415–1418.
[47] Robert W Reeder, Patrick Gage Kelley, Aleecia M McDonald, and Lorrie Faith Cranor. 2008. A user study of the expandable grid applied to P3P privacy policy

visualization. In *Proceedings of the 7th ACM Workshop on Privacy in The Electronic Society*. 45–54.
[48] Daniel Reinhardt, Johannes Borchard, and Jörn Hurtienne. 2021. Visual interactive privacy policy: The better choice?. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–12.
[49] Gary Rolfe, Dawn Freshwater, and Melanie Jasper. 2001. *Critical reflection in nursing and the helping professions: a user's guide*. Palgrave Macmillan, Basingstoke.
[50] Nicolas Ruiz, Gabriela Molina León, and Hendrik Heuer. 2024. Design frictions on social media: Balancing reduced mindless scrolling and user satisfaction. In *Proceedings of Mensch und Computer 2024*. 442–447.
[51] Kanthashree Mysore Sathyendra, Shomir Wilson, Florian Schaub, Sebastian Zimmeck, and Norman Sadeh. 2017. Identifying the provision of choices in privacy policy text. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. 2774–2779.
[52] Florian Schaub, Rebecca Balebako, Adam L Durity, and Lorrie Faith Cranor. 2015. A design space for effective privacy notices. In *Eleventh Symposium on Usable Privacy and Security (SOUPS 2015)*. 1–17.
[53] Donald A Schön. 2017. *The reflective practitioner: How professionals think in action*. Routledge.
[54] Jayson Seaman. 2008. Experience, reflect, critique: The end of the "learning cycles" era. *Journal of Experiential Education* 31, 1 (2008), 3–18.
[55] John S Seberger, Marissel Llavore, Nicholas Nye Wyant, Irina Shklovski, and Sameer Patil. 2021. Empowering resignation: There's an app for that. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–18.
[56] Phoebe Sengers, Kirsten Boehner, Shay David, and Joseph J. Kaye. 2005. Reflective design. In *Proceedings of the 4th Decennial Conference on Critical Computing: Between Sense and Sensibility*. ACM, 49–58.
[57] Murray Shanahan, Kyle McDonell, and Laria Reynolds. 2023. Role play with large language models. *Nature* 623, 7987 (2023), 493–498.
[58] Carolin Stellmacher, Jette Ternieten, Daria Soroko, and Johannes Schöning. 2022. Escaping the privacy paradox: Evaluating the learning effects of privacy policies with serious games. *Proceedings of the ACM on Human-Computer Interaction* 6, CHI PLAY (2022), 1–20.
[59] Madiha Tabassum, Abdulmajeed Alqhatani, Marran Aldossari, and Heather Richter Lipford. 2018. Increasing user attention with a comic-based policy. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–6.
[60] Arnout Terpstra, PAJ Graßl, and HK Schraffenberger. 2021. Think before you click: how reflective patterns contribute to privacy (Position Paper). (2021).
[61] Arnout Terpstra, Alexander P Schouten, Alwin De Rooij, and Ronald E Leenes. 2019. Improving privacy choice through design: How designing for reflection could support privacy self-management. *First Monday* (2019).
[62] Maximiliane Windl, Niels Henze, Albrecht Schmidt, and Sebastian S Feger. 2022. Automating contextual privacy policies: Design and evaluation of a production tool for digital consumer privacy awareness. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–18.
[63] Xiaotong Xu, Arina Konnova, Bianca Gao, Cindy Peng, Dave Vo, and Steven P Dow. 2025. Productive vs. reflective: how different ways of integrating AI into design workflows affect cognition and motivation. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–15.
[64] Shuning Zhang, Lyumanshan Ye, Xin Yi, Jingyu Tang, Bo Shui, Haobin Xing, Pengfei Liu, and Hewu Li. 2024. "Ghost of the past"': identifying and resolving privacy leakage from LLM's memory through proactive user interaction. *arXiv preprint arXiv:2410.14931* (2024).
[65] Shuning Zhang, Xin Yi, Shixuan Li, Haobin Xing, and Hewu Li. 2025. PrivCAPTCHA: Interactive CAPTCHA to facilitate effective comprehension of APP privacy policy. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–20.
[66] Zhiping Zhang, Michelle Jia, Hao-Ping Lee, Bingsheng Yao, Sauvik Das, Ada Lerner, Dakuo Wang, and Tianshi Li. 2024. "It's a Fair Game"', or is it? Examining how users navigate disclosure risks and benefits when using LLM-based conversational agents. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–26.
[67] Kaifa Zhao, Le Yu, Shiyao Zhou, Jing Li, Xiapu Luo, Yat Fei Aemon Chiu, and Yutong Liu. 2022. A fine-grained chinese software privacy policy dataset for sequence labeling and regulation compliant identification. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. 10266–10277.
[68] Kaifa Zhao, Xian Zhan, Le Yu, Shiyao Zhou, Hao Zhou, Xiapu Luo, Haoyu Wang, and Yepang Liu. 2023. Demystifying privacy policy of third-party libraries in mobile apps. In *2023 IEEE/ACM 45th International Conference on Software Engineering (ICSE)*. IEEE, 1583–1595.
[69] Jijie Zhou, Eryue Xu, Yaoyao Wu, and Tianshi Li. 2025. Rescriber: Smaller-LLM-Powered User-Led Data Minimization for LLM-Based Chatbots. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–28.
[70] Sebastian Zimmeck and Steven M Bellovin. 2014. Privee: An architecture for automatically analyzing web privacy policies. In *23rd USENIX Security Symposium (USENIX Security 14)*. 1–16.

A Codebook For the User Study

This codebook details the themes used during the thematic analysis of participant interviews, describing how users progress through the Experience-Reflection-Action (ERA) cycle.

1. From Experience to Reflection

Unexpected Triggers: Discovering data collection during seemingly passive or harmless interactions.

Expectation Gap: Realizing that the application accesses data even without direct user input.

2. From Reflection to Action

Active Defense Strategies: Taking direct actions to minimize data exposure.

Data Fabrication: Inputting false details to protect real personal information.

Temporary Permissions: Revoking system access immediately after completing a task.

Deliberation Without Action: Acknowledging privacy risks but choosing to proceed without altering settings.

Heightened Caution: Monitoring the application's behavior more closely while continuing the task.

Conditional Acceptance: Tolerating the current risk level, while noting that a higher perceived threat would halt the interaction.

Action Barriers: Factors that prevent users from exercising privacy controls.

Task Priority: Valuing the completion of a primary goal (e.g., booking a flight) over privacy protection.

Privacy Resignation: Feeling compelled to accept data risks due to dependency on the application.

3. User Perspective on the Tools

Unobtrusive Design: Delivering privacy notices quietly without disrupting the user's primary workflow.

Information Synthesis: Appreciating AI-generated summaries for clarifying dense and complex privacy terminology.

System Trust: Preferring native operating system integration to ensure clearer accountability and reliability.

Expectations: Expectations towards the improvement on these apps and designs.

B Interfaces of Different Techniques

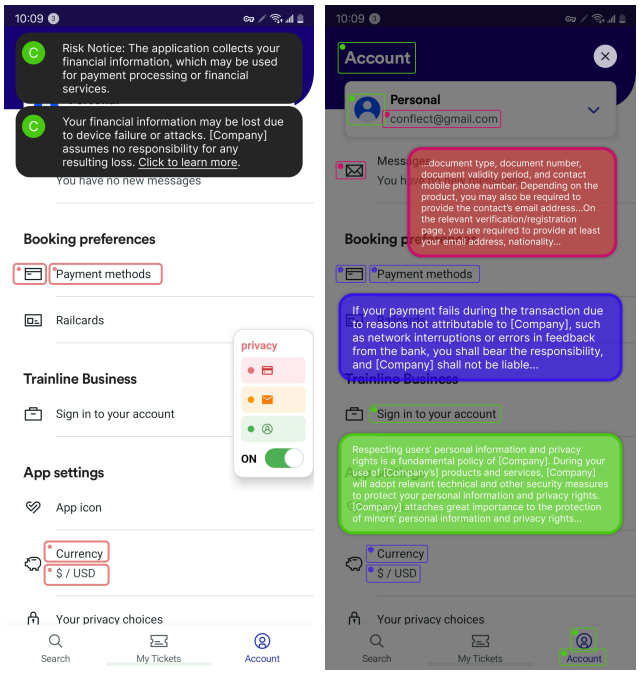
Figure 10 showed the interfaces for different techniques.

C Moderator Guide for Workshops

This section documents the moderator guide used in our workshops. Plan A structured participants' reflective exploration of contextual privacy policies (CPPs) through staged individual sharing and group discussion. Each step followed a reflective sequence (*experience -> reflection -> action*), balancing participant airtime with collective synthesis.

C.1 Welcome, Consent, and Warm-up (5 min)

Introduce each participant to others, and introduce the flow of the workshop.



(a) Conflect. (b) CPP.

Figure 10: The interfaces of different techniques. The interfaces are originally shown in Chinese and we translated them to English.

C.2 Diverge Phase (75 min)

Design Overview (25 min)
• Each participant introduces the proposed usage journey (10 min).
• Group discussion to explore improvements and potential adjustments (15 min).
Problem Solved and Solution Pathways (15 min)
• Each participant explains which privacy policy problems their design addresses and how they are solved (5 min).
• Group discussion on how interactions could solve the problem (10 min).
Theories Involved (15 min)
• Each participant introduces which theories are applied and how they are embedded in the interaction (5 min).
• Group discussion on the contributions of these theories (10 min).
Correlation and Reflection (15 min)
• Each participant presents how reflective chains are represented in the interaction (5 min).
• Group discussion on the potentials and improvements of reflection (10 min).

C.3 Converge Phase (40 min)

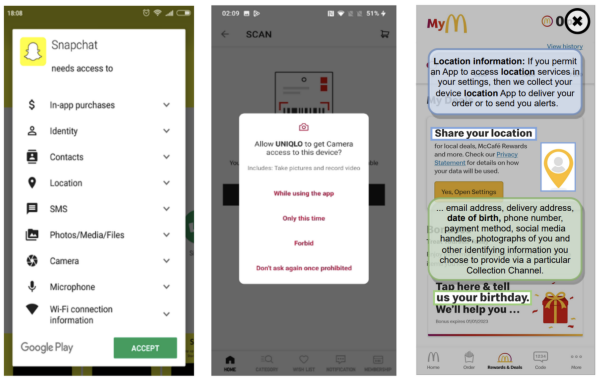
Select Best Ideas (10 min)
• Participants select the best ideas for WHAT, SO WHAT, and NOW WHAT or Experience, Reflection, and Action.
WHAT? (Experience) (10 min)
• Introduce opinions on the best design(s), justify the choice, and hold a short discussion.
SO WHAT? (Reflection) (10 min)
• Introduce opinions on the best design(s), justify the choice, and hold a short discussion.
NOW WHAT? (Action) (10 min)
• Introduce opinions on the best design(s), justify the choice, and hold a short discussion.

D Pre-Workshop Materials

D.1 Background

Goal: To support the design of Contextual Privacy Policies (CPPs) for mobile applications.

CPPs are concise privacy explanations that are presented instantly during use, and displayed directly at the relevant position in the UI (e.g., next to the "Phone Number" field on a checkout page). By anchoring information to the specific screen context that triggers data behavior, CPPs differ from general, app-wide statements. Prior studies have shown that this approach can improve relevance and user comprehension.



(a) Install-time (b) Invoke-time (c) Context-aware

Figure 11: Comparison between install-time disclosures, runtime permission dialogs, and contextual privacy policies. This figure is from Pan et al.'s paper [45], and adopted as experiment materials for the workshop.

Comparison with other privacy notices:

• **Install-time disclosures:** Information shown when a user first installs the app. These may contain a summary of requested permissions (what permissions are required, and for what purpose), but provide limited context.

• **Invoke-time permission dialogs:** Prompts that appear when the app first needs access to a permission (e.g., location) or when permission is not yet granted. These are better timed but still constrained by OS-level categories and tied to developer logic rather than user task context. Users tend to become habituated and often cannot fully understand the data collection logic. • **Contextual Privacy Policies (CPPs):** Context-attached explanations (e.g., "We use your phone number to send delivery updates; stored for 30 days; never sold. Change default settings..."). Research shows that

compared to lengthy policies, CPPs increase user engagement. Recent work demonstrates that CPP snippets can be automatically generated from privacy policy text.

D.2 Reflective Thinking

There are many reflective thinking frameworks. In this workshop, participants were asked to integrate reflective thinking as part of the design concept. The following frameworks were provided as references, though others could also be used.

Driscoll's "What? → So what? → Now what?" framework:

- *What?* Identify the types of personal information, purposes of use, and screen context.
- *So what?* Interpret the impact and show consequences to users.
- *Now what?* Demonstrate what controls users have and how they can act.

ERA loop (Experience → Reflection → Action):

- *Experience:* What kinds of data collection occur in this context, and how the data may be transmitted and shared.
- *Reflection:* What consequences might follow from this collection, transmission, and sharing?
- *Action:* What actions can users take in response?

D.3 Pre-Workshop Tasks

Please complete the following tasks, and prepare two designs before the workshop, for the sake of discussions on the workshop:

Collect Screenshots: Select at least two mobile applications you use. For each app, capture at least one screenshot of a screen involving personal data input or access (e.g., registration, checkout address, contact sync, camera, location).

Design: For your selected apps, sketch at least two design proposals based on reflective design (e.g., "What?" – "So what?" – "Now what?"). Be sure to explain how each stage is considered, how users interact, and which problems of traditional privacy policies are solved.

Reminder: Treat CPPs as reusable, app-independent micro-interaction patterns. Screenshots serve only to illustrate and anchor your proposals; you do not need to pick special scenarios. Any screen involving personal data input or access can serve as an example. If needed, you could further explain how your general design could adjust across contexts while keeping core design elements consistent.

Tips:

- You should take in mind about the three reflective thinking stages and incorporate them into your design.
- You should take in mind that the designs of the three stages should be interconnected.

D.4 Example Design Proposals

Example A:

[Experience]

Design: A visible sidebar on the right side of the interface serves as an entry point, informing users of data collection and risks on the page. The more sensitive the information, the higher the risk indicator.

User Interaction: Clicking the privacy label opens a full-page overlay with details.

Problem Solved: Addresses the lack of contextual integration in traditional privacy policies. Users can directly discover privacy prompts at the point of data collection.

[Reflection]

Design: When clicked, the interface shows concise text at relevant locations, explaining what data is collected, how it is transmitted, and what third parties it is shared with.

User Interaction: Users can quickly understand the implications by browsing the concise privacy text.

Problem Solved: Mitigates information overload and difficulty of comprehension in traditional policies by presenting task-relevant, easy-to-understand content.

[Action]

Design: A master toggle at the top allows one-click control of data collection in the current context.

User Interaction: Users can enable or disable data collection without leaving the task flow.

Problem Solved: Provides just-in-time, convenient control, overcoming the lack of immediacy in traditional policies.

Example B:

[Experience]

Design: Apply highlight borders to UI components involving data input (e.g., "Delivery Address" or "Payment Method" fields).

User Interaction: Users' visual focus is drawn to the highlighted area, associating input behavior with privacy concerns. Clicking the highlight reveals more details.

Problem Solved: Makes privacy information more visible and context-specific, directly tied to the user's current task.

[Reflection]

Design: Clicking a highlighted area triggers a concise top-bar pop-up explaining purpose, risks, and handling (e.g., "Your address is used for delivery, encrypted in transit, and stored for 30 days").

User Interaction: Users learn why data is needed and how it will be used.

Problem Solved: Avoids vague, generic statements by providing targeted, highly relevant explanations for each data type.

[Action]

Design: The pop-up provides options such as a "Learn More" link or a "Manage Permissions" button.

User Interaction: Users can either access the full policy/settings or take immediate control.

Problem Solved: Moves beyond the binary "accept or not" model by offering layered, flexible control options tailored to user preferences.

E User Study Task Descriptions

This appendix provides the detailed scenarios and multi-step tasks assigned to participants for each of the four applications used in our study.

Task 1: Ctrip (Travel Planning)

Scenario: You are planning a vacation from Beijing to Urumqi (Departure: May 1st; Return: May 7th). You need to use the Ctrip app to search for and book flight tickets.

Step-by-step Instructions:

(1) Open Ctrip and select "Flights".

(2) Search for round-trip flights: Beijing to Urumqi (Departure: May 1st morning; Return: May 7th afternoon).
(3) Filter the search results to exclude "Red-eye flights".
(4) Select a suitable round-trip flight and enter the order details page.
(5) Fill in personal information, including Name, Gender, Date of Birth, and ID Number (may use placeholder).
(6) Proceed to the payment page to view supported methods (e.g., Visa/Mastercard, Alipay) and stop at the credit card entry screen without making a payment.

Task 2: iQIYI (Media Consumption)
Scenario: You want to download a popular drama, *The Disguiser*, to your phone for offline viewing during a flight.
Step-by-step Instructions:

(1) Open iQIYI and search for the series "The Disguiser".
(2) Start playback directly from Episode 2.
(3) Locate the "Download" button on the player interface, select 720p quality, and start the download.
(4) Find and click the "Screen Casting" button to confirm its location (actual casting is not required).
(5) Return to the main interface, go to "Mine," and locate "History/Downloads".
(6) Enter the download section to check the current progress.

Task 3: GuaZi Used Car (Market Price Comparison)
Scenario: You spotted a Mini Cooper during your trip and want to check the market prices for second-hand Mini Coopers in Beijing.
Step-by-step Instructions:

(1) Open GuaZi, and click on the "Used Car" category.
(2) Set the location to Beijing and search for "Mini".
(3) Sort the results by "Price: Low to High" and click on the first Mini Cooper listed.
(4) Scroll down to view the official vehicle inspection report.
(5) Click the "Favorite" button after reviewing the report.

Task 4: UC Browser (Browsing & Setting Management)
Scenario: You suddenly decide to visit London and use UC Browser to research travel information while testing the app's privacy and personalization settings.
Step-by-step Instructions:

(1) Search for "London Travel Guide" within UC Browser and browse the results.
(2) Return to the homepage and use the Baidu search bar to search for the same topic again.
(3) Compare the differences in content provided by the two search engines.
(4) Go to "Mine" and open "Settings".
(5) Locate "Privacy Settings" and enter "Personalized Ad Recommendation Management".
(6) Select "Turn Off" for personalized recommendations.

F Justification for Conflect's Detection Threshold

To determine the optimal pixel-wise overlap threshold (τ) for identifying privacy-relevant context switches, we conducted a sensitivity analysis using a dataset of real-world user interactions. Threshold selection followed a two-stage process: first, quantitatively benchmarking the pixel-based detector against human-annotated semantic context changes that capture potentially meaningful interface transitions; and second, qualitatively auditing detection mismatches to assess their relevance to privacy context changes.

We note that defining a single, universally optimal pixel-wise threshold for UI-driven context detection is inherently challenging, as the appropriate operating point depends on factors such as interface design style, interaction density, and system objectives, which continue to evolve over time. Accordingly, our goal in this analysis is not to identify a globally optimal threshold, but to determine a well-justified operating range tailored to the current design and deployment context of Conflect.

F.1 Data Collection

We note that there is currently no publicly available dataset that captures continuous, real-world mobile UI interaction videos with fine-grained temporal alignment to user-driven context changes. Existing datasets primarily focus on static screenshots or app-level UI–policy mappings and therefore do not reflect the dynamic interaction flows required to evaluate context-sensitive, on-device systems such as Conflect.

Accordingly, we constructed a compact, diagnostic interaction dataset tailored to analyzing threshold sensitivity and failure modes of context detection, rather than aiming for broad statistical representativeness. We selected 10 Android applications covering multiple functional domains, including navigation, content consumption, commerce, media, and system utilities. These applications include all those used in the user study and span a range of commonly encountered UI rendering patterns: *Ctrip, Google Maps, Google Play, iQIYI, REI Co-op, Spotify, Guazi, Uber, UC Browser,* and *YouTube*.

We recruited **10 participants** to generate the interaction dataset. Each participant was assigned one application and instructed to perform and record a usage session lasting 60 seconds. Participants were explicitly asked to maximize the variety of interactions while ensuring all actions remained **purposeful and realistic** (e.g., rapid scrolling for actual content scanning rather than random swiping). Subsequently, we reviewed all recordings to ensure consistency with these instructions. This resulted in a dataset rich in complex behaviors including continuous scrolling, tab switching, dialog interactions, and keyboard inputs, enabling the evaluation of distinct semantic transitions amidst frequent UI changes.

F.2 Ground Truth Annotation

To establish a rigorous ground truth, we recruited three senior students with academic backgrounds and internship experiences in UI/UX Design.

The annotators utilized a custom Python-based tool that enables frame-by-frame navigation and event tagging to mark timestamps of context switches, reaching decisions through discussion and consensus. To align with Conflect's design goal of minimizing intrusion, annotators followed a strict protocol to distinguish **Semantic Context Switches**:

• **Positive Labels (Trigger):** Events were marked when the user's action resulted in a functional interface change (e.g., page transitions, dialog/modal appearances, keyboard deployment) or arrival at a new content view following a scroll interaction.

• **Negative Labels (Non-Trigger):** Visual changes representing intermediate states or passive updates were strictly excluded. This includes the transient motion process of scrolling, loading animations, passive video playback, and in-feed advertisements.

Across the 10 annotated 60-second interaction recordings, this procedure resulted in a total of 247 labeled context switch events.

F.3 Sensitivity Analysis Results

We evaluated the automated detection pipeline against the human-annotated ground truth. To verify the system's performance under realistic constraints, we configured the evaluation with the following parameters:

• **Sample Interval:** Comparisons were performed every 0.1 seconds rather than frame-by-frame, simulating a resource-efficient background service; this sampling interval would be consistently applied by Conflect's backend service in the user study.

• **Threshold Definition (τ):** We tested consistency thresholds ranging from 50% to 100%. A context switch is registered when the pixel-wise similarity between consecutive sampled frames falls below τ (i.e., visual change $> 1 - \tau$).

• **Matching Tolerance:** Because semantic context switches correspond to interface states that often persist across multiple frames, a human annotator and the automated detector may identify the same context switch at slightly different frames within the same stable state. To account for this temporal variability in frame-level annotation, we applied a matching tolerance of ±15 frames (≈0.5 second). An automated detection falling within this window of a manual label is counted as a successful match (True Positive).

Table 3: Performance metrics of different thresholds against expert-annotated ground truth. Noise rate corresponds to false positives (FP).

Threshold	Avg. Recall	Avg. Precision	Avg. Noise Rate
50%	24.35%	59.36%	20.64%
60%	34.07%	71.11%	28.89%
70%	44.38%	62.03%	37.97%
80%	63.77%	54.18%	45.82%
90%	86.79%	39.09%	60.91%
100%	100.00%	12.04%	87.96%

Based on the quantitative results in Table 3, we narrowed our focus to three candidate thresholds: **80%, 90%, and 100%**. The rationale for this selection followed a safety-critical filtering process:

Exclusion of Low Thresholds (50%–70%): Thresholds below 80% were immediately discarded due to unacceptably low recall rates (<45%). In a privacy protection context, failing to detect more than half of the significant context switches renders the system ineffective.

100% as a Safety Fallback: While the 100% threshold is operationally impractical due to high noise (87.96%), it serves as the absolute baseline for recall.

80% and 90% as Primary Contenders: These two thresholds represent the optimal trade-off zone. The **90%** threshold prioritizes safety with high recall (86.79%), while the **80%** threshold offers significantly improved precision (54.18%).

To make the final determination, we proceeded with a qualitative frame-by-frame audit to examine the nature of the "missed" detections at the 80% and 90% levels relative to the 100% baseline.

F.4 Qualitative Audit and Final Determination

To determine the final thresholds, we engaged the same three annotators to conduct a frame-by-frame manual inspection of ground-truth semantic context switches that were missed by the automated detector at different thresholds. In particular, we examined two categories of false-negative cases: (1) events that were detected at $\tau = 100\%$ but missed at $\tau = 90\%$, and (2) events that were detected at $\tau = 90\%$ but missed at $\tau = 80\%$.

Impact of Horizontal Navigation: The audit of the 80%–90% interval revealed that the primary discrepancy arose from width-constrained horizontal scrolling. These events typically involved swiping through narrow carousels, tab bars, or filter chips within a static parent view. Annotators classified these interactions as *local exploration within a specific interface* rather than privacy context switches, confirming that they did not introduce new data collection points or alter the overarching privacy context.

Semantic Option Changes Without Salient Visual Differences: The audit of the 90%–100% interval revealed a distinct class of missed events involving *semantic parameter changes* that were not accompanied by prominent visual updates. A typical example occurred in booking interfaces, where users switched trip options (e.g., from one-way to round trip) while the overall page layout and visual composition remained largely unchanged. Annotators concluded that these option-level modifications did not constitute a substantive thematic privacy context shift that would warrant a new privacy reflection.

Coverage of Privacy-Critical Transitions: Following the definition of privacy context articulated by Pan et al. [45], which characterizes privacy contexts as semantically distinct interface states associated with privacy-relevant data practices, our annotators revisited all candidate context transitions in our dataset. Under this definition, no transitions involving substantive changes in privacy context were found among the false-negative cases at the $\tau = 80\%$ threshold.

Therefore, these findings informed our final threshold selection. While higher thresholds increase detection sensitivity, they also lead to a higher frequency of privacy context checks, thereby increasing backend processing overhead. To balance reliable coverage of privacy context changes with system-level processing efficiency, we selected $\tau = 80\%$ as the operating threshold. We also conducted a pre-deployment verification to re-confirmed that this threshold preserved complete coverage of all privacy-relevant context transitions incorporated in the user study tasks. Although higher thresholds may be considered in future large-scale deployments to further optimize system robustness, the selected threshold is sufficient and robust for all scenarios evaluated in our user study.

G Detailed Prompt Design

To mitigate the risk of LLM hallucinations, which could over-alarm users, mischaracterize data practices, and erode user trust, our scenario generation pipeline employs contextual grounding and conservative reasoning constraints. We selected GPT-4o for this

module due to its superior instruction-following capabilities and high fidelity in zero-shot and few-shot semantic reasoning.

G.1 Prompt Design

The prompt is engineered to restrict the LLM from synthesizing speculative or generalized privacy risks. Instead, it mandates that all generated scenarios must be explicitly supported by the provided policy excerpts. The prompt architecture consists of three core components:

• Role-playing instructions: Establishes the LLM as an objective privacy analyst, setting the tone to be informative and reflective.

• Contextual grounding mandate: Enforces conservative reasoning. The model is explicitly instructed to base its risk scenarios only on the provided text. If a data practice's consequence is undefined in the policy, the model must reflect this ambiguity rather than hallucinating a worst-case scenario.

• In-context example: Provides a standardized input-output template and exemplar risks to format the output. To prevent the LLM from artificially scoping its responses only to the provided examples, the prompt explicitly instructs the model to use the examples as structural guides while remaining adaptable to the actual policy text.

G.2 System Prompt

We detailed the system prompt used in the reflective presentation module to generate the risk scenarios:

Role: You are an expert privacy policy analyst and reflective design assistant. Your objective is to translate abstract privacy policy clauses into concrete, contextualized risk scenarios to help users reflect on the implications of sharing their data.

Task: Given a specific [Data Category] and the corresponding [Policy Text Segment] extracted from an application, generate a concise, 1-2 sentence risk scenario explaining how this data is used, shared, or stored, and what the tangible consequences to the user might be.

Strict constraints:

- **Contextual grounding:** You MUST base your reasoning ONLY on the provided [Policy Text Segment]. Do not invent, assume, or hallucinate risks that are not explicitly supported by the text.
- **Conservative reasoning:** If the policy text is vague or does not specify how the data is used, state that the usage is undefined or ambiguous rather than assuming worst-case scenarios. For example, do not claim data is sold to third parties unless the text implies or states it.
- **Tone:** Be objective and informative. Avoid sensationalist and over-alarming language. Your goal is to prompt conscious reflection, not panic.

Example Format:

Input:

[Data Category]: Location

[Policy Text Segment]: "We collect your precise location to provide local weather updates and may share this data with our marketing partners to deliver targeted advertisements."

Output:

{

"risk_scenario": "This app collects your precise location to share with marketing partners for targeted ads, which means your physical movements may be tracked and monetized by third parties."

}

Note: The example above shows the expected I/O format and an exemplar risk. DO NOT limit your outputs to specific risks shown in the example. You MUST adopt your scenario entirely to the unique contents of the provided [Policy Text Segment].

Current Input:

[Data Category]: {input_category}

[Policy Text Segment]: {input_policy_text}

Output (JSON only):

H Technical Evaluation of Conflect's Components

We conducted a technical evaluation of Conflect's core components to show its technical feasibility, specifically assessing the accuracy and effectiveness of the privacy policy extraction and scenario generation processes.

H.1 Evaluation of Privacy Policy Extraction

We evaluated three parts of privacy policy extraction, including the accuracy of search privacy policy, the accuracy of privacy policy segment extraction, and the usefulness of reflective thinking-based conversion. We used CPP4APP [45], CA4P-483 [67] and MAPP Corpus [2], where CPP4APP features a common dataset for CPP context. CA4P-483 is a widely used Chinese mobile privacy policy dataset [68], and MAPP Corpus featured English and German mobile privacy policies. For MAPP Corpus, we only used the English privacy policies. CPP4APP, CA4P-483 and MAPP Corpus have 50, 483 and 64[3] privacy policies. For all the evaluations, experimenters manually checked to determine the correctness of the LLMs' generated results.

We evaluated privacy policy segment extraction accuracy on three state-of-the-art LLMs, GPT-4o, GPT-5 and Claude-4.5-Sonnet. We averaged three runs per data item to mitigate random effects. We considered an extraction correct if (1) the extracted text accurately reflects the original policy (e.g., no hallucination), and (2) the extracted text is comprehensive, and correctly matches the queried data category. Any omission of relevant clauses or inclusion of irrelevant text was marked as incorrect.

As shown in Table 4 and 5, all three models exhibited high extraction accuracy: GPT-5 achieved the highest average accuracy at 96.4%, followed by Claude 4.5 Sonnet (93.1%) and GPT-4o (91.6%). All models showed high robustness for common personal identifiers (e.g., location, contacts, name), with extraction accuracy frequently exceeding 95% for GPT-5. Categories with potentially more complex or ambiguous phrasing, such as profile and financial info, resulted in lower accuracy. While the overall accuracy is high, we analyzed the failure cases, which included *omissions of relevant clauses* (false negatives, FNs) and *irrelevant text or hallucinations* (false positives, FPs). We observed that FNs were slightly more common, typically occurring when policies used highly ambiguous phrasing that eluded the model's semantic matching. Conflect's continuous detection mechanism could mitigate this risk to a certain extend, because as

[3]Counting only English privacy policies.

users scroll or interact, the system triggers re-detection, thereby compensating for sporadic omissions. Conversely, FPs were rare but occasionally occurred when the model conflated similar data types. These instances functions as a "fail-safe" or educational trigger, prompting users to verify data practices rather than blindly trusting the app, similar to sandbox-based strategies [13], although it may incur distrust.

We further qualitatively analyzed the extraction errors to provide transparency into Conflect's limitations. We classified errors into false negatives (FN) and false positives (FP).

False negatives (omission). FNs occurred when the system failed to identify relevant privacy clauses. This mostly happened when policies used ambiguous or non-standard terminology.

• *Example:* A policy stated "We measure how you use our services to improve them," which implied collecting usage data (Behavior), but the model failed to map "measure how you use" to the "Usage Data" category due to the lack of explicit keywords.

False positives (hallucination/misclassification). FPs occurred when the system incorrectly flagged text as a specifci data practice.

• *Example:* In a "Contact Us" section, the model extracted the company's support email address and misclassified it as "Email data".

This analysis suggests that while Conflect is highly accurate, future iterations could benefit from fine-tuning on domain-specific policy corpora to handle ambiguous legal phrasing.

H.2 Evaluation of Scenario Generation

We finally evaluated whether the generated risk descriptions are useful, where we recruited 30 participants through distributing recruiting posters on public social media channels such as WeChat. Participants manually label all generated risk descriptions according to the criteria that "whether the generated risk description would prompt reflections" on a 7-point Likert scale (1=not at all, 7=very much), following prior guidelines for evaluation [14]. The study was approved by our institutions' IRB, and each participant was compensated 200 CNY according to local wage standards. As in Table 6, we found on average, ratings on all datasets exceed 5.6 out of 7. Specific classes such as *social media* and *contact* reached a higher score of over 6 out of 7, indicating the feasibility of the risk descriptions.

I Informed Consent Form

Study Title: Evaluating Contextual Privacy Policies for Mobile Apps

Principle Investigator: [anonymized for submission]

Institution: [anonymized for submission]

Purpose of the research

You are invited to participate in a research study aimed at improving how mobile users interact with contextual privacy policies. We aim to understand how users perceive and manage privacy when assisted by automated systems.

Procedure

As part of this study, you will be required to install a research application on your personal Android device. This application requires the activation of the *Android Accessibility Service*. Once enabled, the application will monitor interface changes to trigger privacy notifications. The study will last for about 30-60 minutes.

Technical risks and disclosure

The *Android Accessibility Service* is a highly privileged system-level permission. By enabling this service for our application, you acknowledge and understand the following technical risks:

• Full screen content access: The application has the technical capacity to read all text and graphical elements displayed on your screen, including sensitive Personal Identifiable Information (PII) (e.g., names, financial details, or private messages) appearing within third-party applications.

• Interaction monitoring: The service can monitor your real-time interactions with the device, including button presses, gestures, and text input (i.e., similar to a keylogger for the purpose of UI automation).

• Cross-app surveillance: The application's monitoring capabilities remain active across almost all applications on your device, allowing it to track behavioral patterns across different contexts.

• Third-party access: As we may use third-party LLMs in this experiment, some of your filled data or interface icons may be transmitted to LLM providers via the *Android Accessibility Service*.

• Other risks: Besides *Android Accessibility Service*, we may also collect your demographics data through the questionnaire, which contain sensitive data.

Data privacy and mitigation measures

To mitigate the risks associated with the *Accessibility Service*, we have implemented the following safeguards:

• Local processing: All UI analysis is performed locally on your device. We would not transmit touch or input data. Only screenshots are uploaded to our institution-affiliated servers.

• Data minimization: Only anonymized metadata regarding privacy context shifts will be transmitted to OpenAI for functionality purpose. We will request the deletion of these data via APIs to the corresponding LLM providers (i.e., OpenAI). Other data would remain on our institutional server, as could only be accessed by the author team for research analysis. We will delete the data one year after the experiment.

• Exclusion lists: Our application is programmed to have interactions only with the dedicated apps, thereby ignoring sensitive apps, even if you happenly opened them (e.g., banking, password managers).

Benefits

By participating in this study, you feedback will contribute to the development of contextual privacy policies that could better serve people's daily informed consent. Besides, you will receive 100 CNY as compensation.

Confidentiality

All data collected will be handled in accordance with our local regulations. Your identity will be pseudonymized, and results will only be reported in aggregate form. No personal identifiable data will be disclosed in any publication.

Voluntary participation and withdrawal

Participation is entirely voluntary. You may disable the *Accessibility Service* or uninstall the research application at any time without penalty. You could also quit the experiment at any time before, during or after the experiment without any reason. You may request the immediate deletion of any data already collected from

Table 4: Comparative extraction accuracy of different LLMs across key data categories. Results are averaged across CPP4APP, CA4P-483, and MAPP Corpus.

Data Category	GPT-4o			GPT-5			Claude 4.5 Sonnet		
	Acc.	FN	FP	Acc.	FN	FP	Acc.	FN	FP
Location	93.3%	3.0%	3.7%	98.5%	1.0%	0.5%	95.2%	4.0%	0.8%
Address	91.3%	4.5%	4.2%	97.0%	2.0%	1.0%	92.8%	5.2%	2.0%
Phone	90.9%	5.0%	4.1%	96.5%	2.5%	1.0%	93.5%	5.0%	1.5%
Email	90.7%	3.5%	5.8%	98.0%	1.0%	1.0%	94.8%	3.2%	2.0%
Birthday	91.9%	3.5%	4.6%	98.5%	1.0%	0.5%	91.0%	7.0%	2.0%
Contacts	92.7%	5.0%	2.3%	97.2%	2.0%	0.8%	94.5%	4.0%	1.5%
Name	93.8%	3.4%	2.8%	94.1%	4.0%	1.9%	96.0%	3.0%	1.0%
Voices	89.2%	8.0%	2.8%	95.5%	3.5%	1.0%	90.5%	7.5%	2.0%
Social Media	94.5%	4.0%	1.5%	96.8%	2.0%	1.2%	95.0%	3.0%	2.0%
Photos	93.6%	3.9%	2.5%	96.0%	3.0%	1.0%	93.5%	5.5%	2.0%
Profile	89.3%	6.0%	4.7%	94.5%	4.0%	1.5%	90.0%	7.0%	3.0%
Financial Info	89.2%	6.0%	4.8%	94.1%	4.0%	1.9%	90.5%	7.0%	2.5%
Avg.	91.6%	–	–	96.4%	–	–	93.1%	–	–

Table 5: Comparison of overall extraction accuracy across different datasets and models. Results are averaged across data types.

Dataset	GPT-4o	GPT-5	Claude 4.5 Sonnet
CPP4APP	94.0%	98.0%	96.0%
CA4P-483	93.3%	95.9%	94.2%
MAPP Corpus	87.6%	95.3%	89.1%
Avg.	91.6%	96.4%	93.1%

your device during or after the experiment without any reason. You are free to raise any questions to the experimenter before, during or after the experiment.

Participant consent

I have read the above information and have had the opportunity to ask questions. I understand that the application utilizes *Android Accessibility Services* and I am fully aware of the associated privacy implications. By signing "I Consent" and manually enabling the *Accessibility Service* in my system settings, I agree to participate in this study. Contact: [anonymized name], [anonymized email], [anonymized phone number]

Participant Signature / Date: (blank)

Table 6: The results of "whether the generated risk description would prompt reflections" (1=not at all, 7=very much), where the number after ± denoted one standard deviation.

	Location	**Address**	**Phone**	**Email**	**Birthday**	**Contacts**	**Name**
CPP4APP	6.4 ± 0.3	6.5 ± 0.2	6.2 ± 0.2	6.1 ± 0.1	6.6 ± 0.2	6.8 ± 0.1	5.9 ± 0.3
CA4P-483	5.8 ± 0.4	6.0 ± 0.3	5.4 ± 0.4	5.2 ± 0.4	6.0 ± 0.2	6.2 ± 0.2	5.0 ± 0.5
MAPP Corpus	5.9 ± 0.4	5.9 ± 0.3	5.4 ± 0.3	5.3 ± 0.6	5.9 ± 0.2	6.0 ± 0.3	5.3 ± 0.4
	Voices	**Social media**	**Photos**	**Profile**	**Financial info**	**Avg.**	
CPP4APP	6.1 ± 0.0	6.4 ± 0.2	6.6 ± 0.1	6.4 ± 0.1	6.5 ± 0.1	6.4 ± 0.2	
CA4P-483	4.8 ± 0.4	6.2 ± 0.4	6.3 ± 0.2	6.1 ± 0.5	6.1 ± 0.4	5.8 ± 0.4	
MAPP Corpus	5.1 ± 0.5	6.1 ± 0.4	5.8 ± 0.3	5.6 ± 0.2	5.6 ± 0.3	5.7 ± 0.3	