Co-Adaptive Eco-Nudging: A Privacy-Preserving Contextual Bandit with User-Taught Preferences in Everyday Browsing

Guangrui Fan*
School of Computer Science and Technology
Taiyuan University of Science and Technology
Taiyuan, Shanxi Province, China
Faculty of Computer Science & Information Technology
Universiti Malaya
Kuala Lumpur, Malaysia
fgr@tyust.edu.cn

Dandan Liu
Faculty of Arts and Social Sciences
Universiti Malaya
Kuala Lumpur, Malaysia
s2134717@siswa.um.edu.my

Lihu Pan
School of Computer Science and Technology
Taiyuan University of Science and Technology
Taiyuan, shanxi, China
panlh@tyust.edu.cn

Abstract

Digital eco-nudges are widely deployed, yet their long-term efficacy, ethical acceptability, and net environmental impact remain unclear. We report two field studies targeting routine online behaviors under strict parity of message content and delivery budgets. Study 1 shows that minimal, factual tailoring improves compliance over generic prompts when opportunities are defined independently of delivery. Study 2 introduces a privacy-preserving, on-device contextual bandit that learns when to act and when to DoNothing, achieving higher compliance at comparable prompt intensity while maintaining autonomy. We operationalize an Ethical–Efficacy Frontier (EEF) to visualize autonomy–effectiveness trade-offs, and compute an energy Return on Investment (ROI) that nets behavior-driven savings against measured system overhead. Energy savings are estimated using literature-calibrated proxies with sensitivity bands, and we discuss the energy trade-offs of on-device learning relative to a stylized cloud alternative. We probe short-term persistence via withdrawal and a brief follow-up; long-term habit formation and rebound remain out of scope. We contribute design and reporting practices—ablation parity, opportunity denominators, EEF, and net-impact accounting—that make digital sustainability interventions more rigorous, transparent, and respectful, advancing sustainable HCI beyond "small changes."

CCS Concepts

• **Human-centered computing → User studies**; **Empirical studies in interaction design**.

Keywords

Sustainable HCI, Digital Nudging, Personalization, Contextual Bandits, On-Device Learning, Privacy and Autonomy, Behavior Change, Energy and Carbon, Evaluation Methods, Net-Impact Accounting

[*] Corresponding author.

ACM ISBN 979-8-4007-2278-3/26/04
https://doi.org/10.1145/3772318.3791358

ACM Reference Format:
Guangrui Fan, Dandan Liu, and Lihu Pan. 2026. Co-Adaptive Eco-Nudging: A Privacy-Preserving Contextual Bandit with User-Taught Preferences in Everyday Browsing. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26), April 13–17, 2026, Barcelona, Spain.* ACM, New York, NY, USA, 20 pages. https://doi.org/10.1145/3772318.3791358

1 Introduction

Digital services increasingly mediate everyday activities—from browsing and streaming to file transfers and casual printing—and thus shape a growing share of our aggregate energy demand and associated emissions. Sustainable HCI (SHCI) has long explored eco-feedback and persuasive technologies to encourage more sustainable practices, yet influential critiques argue that focusing on "small changes" at the individual level is insufficient without attention to structural conditions, justice, and system-level rebound effects [5, 11, 12, 19, 42]. Recent research urges researchers to (i) report the limits of interventions, (ii) evaluate net impact (not just proximal compliance), and (iii) articulate how designs sit within broader sociotechnical systems and infrastructures [29, 30, 34]. Empirically, rebound remains a central risk in household and smart-home contexts, where efficiency gains can prompt compensating behaviors that reduce or erase environmental benefits [4, 53].

Within this debate, eco-feedback and digital nudging remain attractive because they promise measurable behavior change with low implementation cost. Literatures show that the information conveyed, timing of delivery, and display modality shape attention, motivation, and learning [15, 39]. Systematic reviews in HCI catalog a broad repertoire of technology-mediated nudges but also highlight heterogeneous effects, short study horizons, and common evaluation confounds (e.g., message content/length and delivery budgets differ across arms) [6]. Complementing these findings, behavioral-economics and marketing literatures report modest yet meaningful effects for defaults and other choice-architecture levers in environmental domains [7, 47] and, in digital settings, emphasize that timing and interface nuance affect durability and acceptance [49, 54]. Recent applications in e-commerce illustrate promise but also context dependence and annoyance risks from over-frequent prompts [32, 40, 49].

A natural response has been personalization: tailoring when and what to prompt so that interventions are relevant rather than generic. Human-centered ML argues that such systems should keep people in the loop, expose meaningful controls, and learn from explicit feedback [2, 13, 36, 37]. In consumer domains, AI-driven personalization often increases engagement precisely through perceived relevance, yet it simultaneously raises concerns about surveillance, manipulation, and opacity [18, 25, 27, 28, 35, 46]. Ethical analyses converge on preserving autonomy and providing transparency and user control [1, 8, 20, 24, 43, 45, 52]. Meanwhile, privacy-calculus and IUIPC work show that perceived benefits are continually traded against perceived data-collection risks, underscoring the importance of data minimization and legible governance [9, 10, 26, 41, 48].

Methodologically, contextual bandits offer a lightweight online-learning approach to decide when to act and what to show under uncertainty, with extensive precedent in interactive personalization [21, 23, 44]. Crucially, including an explicit DoNothing action allows a learner to acquire restraint and avoid low-yield or ill-timed interruptions, aligning algorithmic choices with autonomy goals rather than maximizing activity at all costs.

Finally, because digital interventions themselves consume resources, SHCI increasingly calls for net-impact accounting: interventions should disclose their energy overheads and report savings minus costs [33]. Electricity intensity estimates for data transmission remain contested, motivating transparent assumptions and sensitivity analysis [3]. Edge/On-device intelligence can reduce network traffic and keep sensitive context local, improving privacy and often reducing the energy associated with repeated cloud inference [17, 22].

Despite this progress, three gaps persist. G1: Nudging evaluations rarely enforce parity (matched length/valence, identical budgets/cool-downs) or include all detected opportunities in denominators, leaving effects confounded by message salience or selection on delivery [6, 54]. G2: Personalization is often server-based and opaque; few studies test on-device, human-centered learners that incorporate DoNothing, explicit feedback, and user-governed constraints [2, 24]. G3: Evaluations seldom (i) characterize the ethical–efficacy trade-off (autonomy vs. compliance), (ii) examine persistence after withdrawal, or (iii) report net energy benefit with uncertainty [3, 33].

We address G1–G3 with two field studies and a privacy-preserving browser extension that targets routine digital behaviors (tab management, streaming quality, large transfers, printing), a comparatively under-studied locus of potential savings [49, 54]. Study 1 compares context-tailored prompts to generic and control under strict ablation parity and opportunity-based denominators. Study 2 evaluates a co-adaptive, on-device constrained contextual bandit (with DoNothing, quiet hours, and user-set prompt budgets) against a matched rule-based policy. Across both studies we (i) quantify an Ethical–Efficacy Frontier (EEF) that renders autonomy–compliance trade-offs visible (by analogy to decision-curve analyses [51]), and (ii) compute an Energy ROI that nets estimated savings against measured overheads with sensitivity bands [3, 33].

Guided by the above literatures and gaps, we investigate:

- **RQ1 (Efficacy under constraints).** Do context-tailored prompts (Study 1) and a co-adaptive on-device bandit (Study 2) improve compliance relative to their comparators when prompt budgets and message content are controlled?
- **RQ2 (Ethical–Efficacy Trade-off).** How does compliance vary with prompt intensity, and what is the Ethical–Efficacy Frontier (EEF)—the Pareto frontier relating compliance to autonomy—under each condition?
- **RQ3 (Persistence).** Do effects persist during post-intervention withdrawal and at a short follow-up checkpoint?
- **RQ4 (Net Impact).** What is the Energy ROI (estimated energy/CO_2 savings due to behavior change minus the system's computational/network overhead) of each approach, and how do the energy costs of on-device learning compare—at a high level—to a stylized cloud-based inference alternative under realistic usage?

This paper offers four contributions to SHCI and human-centered personalization:

1. **Design/Evaluation rigor.** A parity-controlled, opportunity-denominator evaluation of eco-nudges, reducing selection-on-delivery and content-length confounds.
2. **Human-centered on-device learning.** A privacy-preserving, constrained contextual bandit that operationalizes DoNothing, user-governed quiet hours and budgets, and explicit in-situ feedback.
3. **New lenses and transparent reporting.** (i) The Ethical-Efficacy Frontier (EEF) to render autonomy–compliance trade-offs legible and actionable (by analogy to decision curves), and (ii) an Energy ROI that reports net environmental benefit with explicit assumptions and sensitivity to contested energy-intensity estimates. We also provide a design-rationale comparison of on-device versus cloud energy overhead.
4. **Evidence on routine digital behaviors.** Two field studies targeting tabs, streaming, large transfers, and printing—behaviors with tangible but under-measured sustainability implications.

Study 1 establishes parity-controlled efficacy of context-tailored vs. generic prompts under fixed budgets and ablation-matched content. Study 2 evaluates a co-adaptive, on-device learner that includes DoNothing and user-governed quiet hours/budgets against a matched rule-based policy at similar prompt intensity, and introduces the Ethical–Efficacy Frontier (EEF) and Energy ROI reporting. We close with implications and limitations, emphasizing rigor (parity, denominators), autonomy, privacy, and net impact. Because we infer energy from behavior proxies (resolution, bytes, prints, tab concurrency), ROI estimates reflect literature-calibrated coefficients and measured on-device overhead rather than direct power metering; we report sensitivity bands and explicitly discuss contested assumptions. The baseline and intervention windows are short; we therefore include withdrawal and a brief follow-up, and we treat persistence and rebound as limitations and future work.

2 Related Work

2.1 Sustainable HCI: from individual behaviors to systems thinking

Early sustainable HCI (SHCI) emphasized persuasive technologies that target "small changes" in individual behavior, but a substantial body of critique argues this frame is too narrow without attention to structural conditions, justice, and rebound effects [5, 11, 12, 19, 42]. Recent scholarship calls for reporting limits and net impact rather than proximate behavior change alone, and for situating interventions within sociotechnical systems [29, 30, 34]. Rebound—direct and indirect—remains a central concern in home and digital contexts, motivating designs that measure persistence and net outcomes [4, 14, 53]. Classic behavioral-economics arguments about choice architecture also inform SHCI's ethical stance that influence should preserve freedom of choice [47]. We respond to these calls by (i) reporting behavior-level outcomes in addition to opportunity-level compliance, (ii) analyzing withdrawal/follow-up for persistence, and (iii) computing an Energy ROI that nets estimated savings against system overhead [33].

2.2 Eco-feedback and digital nudging

Eco-feedback research shows that information, timing, and display shape effectiveness via attention, motivation, and learning mechanisms [15, 39]. Reviews of technology-mediated nudging note heterogeneous effects, short horizons, and evaluation confounds—particularly when message content/length or delivery budgets differ across arms [6]. Complementing these HCI syntheses, behavioral-economics meta-analyses and field work show modest but meaningful effects for pro-environmental defaults and similar levers [7], while recent overviews of digital nudges document timing/interface nuances and short-lived impacts without reinforcement [54]. In applied settings, digital eco-nudges have targeted e-commerce and platform interfaces (e.g., simplified displays, salient environmental cues) with encouraging but context-dependent outcomes [32, 40]. Design reports likewise emphasize that poorly timed or overly frequent prompts increase annoyance and attrition [49]. Our Study 1 addresses these gaps through ablation parity (matched length/valence and fixed budgets) and denominators that include all detected opportunities, not only "moments when a prompt was delivered," reducing selection on delivery. The finding that minimal factual tailoring improves outcomes while autonomy costs track intensity (not tailoring per se) refines design–behavior frameworks by isolating when and what under parity constraints [6, 39, 54].

2.3 Personalization and bandits for behavior change

Personalization can increase relevance but risks undermining autonomy if it ignores user preferences or context [16, 31]. Human-centered ML urges designs that keep people "in the loop," expose meaningful controls, and learn from explicit user signals [2, 13, 36, 37]. In consumer and marketing domains, AI personalization often boosts engagement precisely because users experience higher relevance, yet it also raises concerns about surveillance and manipulation [18, 28, 35, 46]. Within sustainable design practice, data-driven tailoring is increasingly proposed to align messages with user values and contexts [38], but calls persist to avoid "black-box" persuasion and to foreground user control [25, 27]. Methodologically, contextual bandits provide a lightweight way to learn when/what to deliver under uncertainty, with strong precedents in interactive personalization [21, 23, 44, 50]. Critically, including an explicit DoNothing action allows the learner to acquire restraint and reduce low-yield interruptions—often missing from rule-based pipelines but essential for balancing efficacy with autonomy. Our Study 2 operationalizes this with on-device Thompson sampling, explicit feedback as part of the reward, and feasibility filters (quiet hours/budgets) that implement user-governed autonomy.

2.4 Autonomy, transparency, and the ethics of nudging

The ethics literature emphasizes that nudges should preserve choice, avoid deception, and be transparent and accountable [20, 43, 45]. Perceptions of creepiness and fairness are tied to privacy concerns and expectations; IUIPC and privacy-calculus research shows people trade benefits against perceived risks to data collection, control, and awareness [10, 26, 27]. Responsible-AI work converges on three levers for trust: transparency, perceived fairness, and user control [1, 8, 24]. In practice, systems should disclose what is collected and why, enable meaningful opt-outs/quiet hours, and avoid hidden profiling that blurs the line between assistance and manipulation [9, 25, 41, 48]. Empirical studies suggest that explaining "why this prompt now" and making defaults legible can preserve autonomy while sustaining effectiveness [49, 52]. Our system implements (i) user-governed constraints (quiet hours, prompt budgets, per-behavior opt-outs), (ii) concise Why this? explanations grounded in the user's goals and the triggering context, and (iii) one-tap feedback that directly shapes the learning process—choices intended to maintain autonomy while enabling effective personalization.

2.5 Energy, edge AI, and net-impact accounting

Sustainable HCI increasingly urges moving from proximal behavior change to *net* environmental outcomes that include (i) the footprint of the intervention itself and (ii) plausible rebound or spillover effects [33]. Because direct device and network energy instrumentation is rarely feasible in field deployments, studies necessarily rely on literature-calibrated proxies—together with sensitivity analyses that expose how conclusions vary under contested assumptions [3]. Two sources of uncertainty are especially relevant here: first, the electricity intensity of internet transmission (often modeled as bit-proportional, but also argued to be capacity-dominated in some contexts); and second, resolution and device dependent playback costs for streaming.

Edge/on-device intelligence can reduce network traffic and keep sensitive context local, improving privacy and—at the modest inference scales typical of lightweight personalization—often reducing end-to-end energy relative to repeated cloud inference and RPC overhead [17, 22]. The trade-off is design and workload specific: heavier models or high decision rates may favor cloud offload, whereas millisecond-scale linear models with a handful of features typically favor local execution given per-request network and server amortization. Our design therefore adopts *personalization without profiling* on device for privacy and to avoid routine network calls, while explicitly discussing the energy implications rather than presuming one architecture is always greener.

Operationally, we complement behavior level outcomes with an *Energy ROI*: estimated savings minus measured system overhead. To render autonomy costs visible alongside efficiency gains, we pair ROI with the Ethical–Efficacy Frontier (EEF) (by analogy to decision-curve analyses [51]), which plots compliance as a function of realized prompt intensity and perceived autonomy. In short, our stance aligns with SHCI calls for transparent accounting: disclose assumptions, bound uncertainty, and make ethical trade-offs legible—not only average treatment effects.

3 System Overview

This section details the co-adaptive eco-nudging system we deploy in Study 2 and reference in Study 1 for message parity and instrumentation. The system is instantiated as a privacy-preserving, on-device constrained contextual bandit that (i) solicits user goals and preferences up front, (ii) learns when and what to nudge under explicit autonomy constraints (prompt budgets, quiet hours), and (iii) provides in-situ explanations with explicit feedback channels that directly influence subsequent decisions. We also describe the matched message library used across conditions and the privacy/energy safeguards. Unless otherwise stated, all study arms used identical front-end affordances: the same nudge UI, the same template-based Why this? explanations, and the same one-tap feedback (Helpful, Annoying, Wrong time). The only difference between arms in Study 2 is the decision policy (co-adaptive bandit vs. rule-based threshold).

3.1 Co-Adaptive Nudge Pipeline

Figure 1 summarizes the three phases: Phase 1 (user-initiated onboarding), Phase 2 (adaptive engine), and Phase 3 (feedback & explanations). Algorithmic details and the action/state specification are given in Table 1 and Algorithm 1.

Phase 1: User-initiated onboarding & preference setting. At installation, the extension presents a short, self-explanatory panel that (a) centers user agency and (b) seeds model priors and constraints:

- **Goal selection** (multi-select): {Reducing my carbon footprint, Saving energy to lower my bills, Improving my computer's performance, Just curious}. Internally, these map to a sparse goal vector used as context features and to initialize arm priors.
- **Persuasion style** (single-select): {Analyst (data-forward), Coach (encouraging), Pragmatist (self-benefit)}. This choice enables style-matched templates and informs initial arm preferences.
- **Autonomy constraints**: (i) quiet hours (e.g., 21:00–09:00), and (ii) a daily prompt budget $C \in \{0, 1, 2, 3, 4\}$ [1]

These selections are stored locally and are editable at any time. No server-side profile is created.

Phase 2: Adaptive Engine (constrained contextual bandit). The core decision maker is a light-weight, on-device contextual bandit that learns, per user, which action to take in a given context. The context $x_t \in \mathbb{R}^d$ is constructed from:

- **User-seeded features**: goal vector; persuasion style.
- **Behavioral state**: tab count; streaming state & current resolution; recent large transfer activity; time since last prompt of each type.
- **Temporal & device context**: time-of-day, day-of-week; battery percentage; coarse CPU load; session length.
- **Ephemeral history**: rolling counts over recent prompts (e.g., last-7 Helpful/Annoying; last-24h dismiss rate).

The action set $\mathcal{A}$ comprises target×tone nudge templates plus a DoNothing action:

$$\mathcal{A} = \{\textsc{DoNothing}\} \cup \{\textsc{Tabs\_Analyst},\ \textsc{Tabs\_Coach},\ \textsc{Tabs\_Pragmatist},\ \textsc{Stream\_Analyst}, \ldots\}.$$

Templates are strictly ablation-matched across tones and between generic vs. context-tailored variants (Section 3.2).

Constrained objective. At each decision time t, the learner selects $a_t \in \mathcal{A}$ to maximize expected utility subject to autonomy constraints:

$$\max_{\pi} \ \mathbb{E}[r(a_t \mid x_t)] \quad \text{s.t.} \quad \mathbb{E}[c(a_t \mid x_t)] \leq C, \quad a_t \notin Q, \tag{1}$$

where r is a composite reward (below), c is a unit cost for any non-DoNothing action, C is the per-day prompt budget chosen by the user, and Q encodes quiet hours. In practice, we enforce hard caps (no delivery if budget exhausted or within quiet hours) and prefer actions with higher upper-confidence scores (LinUCB) or posterior samples (Thompson sampling). In Study 2 we instantiate Thompson sampling (TS).

Reward and learning. The reward signal balances behavior change with user satisfaction:

$$r_t = \underbrace{\mathbb{I}[\text{compliance within window}]}_{+1 \text{ if yes}} + \underbrace{\text{vote}_t}_{\begin{cases} +1 & \text{Helpful} \\ -1 & \text{Annoying} \\ -0.5 & \text{Wrong time} \\ 0 & \text{otherwise} \end{cases}},$$

with cost $c_t = \mathbb{I}[a_t \neq \textsc{DoNothing}]$. We maintain per-action linear models (θ_a, A_a) with ridge regularization. For LinUCB, each action's score is

$$S(a \mid x_t) = \hat{\theta}_a^{\top} x_t + \alpha \cdot \sqrt{x_t^{\top} A_a^{-1} x_t},$$

and we select the feasible action with maximal $S(a \mid x_t)$. For Thompson sampling, we draw $\tilde{\theta}_a \sim \mathcal{N}(\hat{\theta}_a, \sigma^2 A_a^{-1})$ and select $\arg\max_a \tilde{\theta}_a^{\top} x_t$ under the same feasibility constraints. The DoNothing action is always present and frequently optimal once the model has learned user-specific patterns.

Opportunity vs. behavior outcomes. For streaming, opportunity-level compliance credits practical step-downs (e.g., HD→SD within a

[1] Participants could set the daily prompt budget to any integer in $\{0, 1, 2, 3, 4\}$. A choice of $C = 0$ disables delivery while preserving logging; these participants remain in intent-to-treat analyses for their assigned arm. Per-protocol analyses exclude $C = 0$ for delivery-effect estimates; in our sample, $C = 0$ was rare. Both constraints are enforced as hard constraints at decision time.

short window) that users can adopt mid-task without abandoning intent. Full stoppage (ending playback) is captured at the *behavior level* as a reduction in daily HD minutes. This separation ensures that opportunity metrics reflect feasible in-the-moment changes, while behavior metrics capture larger session-level shifts. For Energy ROI, we apply *resolution-aware* sensitivity (Appendix D.2), so that step-downs from ≥1080p contribute larger per-minute deltas than 720p→480p, whereas the detector itself remains resolution-agnostic (Instrumentation §5.2).

Safe exploration and logging. Cold-start exploration is bounded. Each decision logs a propensity (estimated selection probability) to enable off-policy checks. Model updates and state aggregation run at idle to minimize overhead. Propensity values and decision logs remain on device and are used only for on-device learning; any exported study logs are anonymized and exclude page content.

Algorithm 1 Constrained Contextual Bandit (on-device)

```
**Inputs:** user prompt budget $C$, quiet hours $Q$, priors seeded by Phase 1
**for** each decision time $t$ with context $x_t$ **do**
    **if** (within $Q$) or (delivered_today $\geq C$) **then**
        choose $a_t \leftarrow$ DoNothing
    **else**
        compute scores $S(a \mid x_t)$ for all $a \in \mathcal{A}$ (LinUCB) or sample $\tilde{\theta}_a$ (TS)
        **select** feasible $a_t = \arg\max_{a \in \mathcal{A}} S(a \mid x_t)$ (or $\tilde{\theta}_a^\top x_t$)
        **deliver** $a_t$ (unless DoNothing); log propensity $p(a_t \mid x_t)$
        **observe** compliance $y_t$ and explicit vote $v_t \in \{\text{Helpful, Annoying, Wrong time}, \varnothing\}$
        **set** $r_t \leftarrow y_t + \text{vote}(v_t)$;   $c_t \leftarrow \mathbb{I}[a_t \neq \text{DoNothing}]$
        **update** $(\hat{\theta}_{a_t}, A_{a_t})$ with $(x_t, r_t)$ (ridge-regularized)
        **apply** cool-down; decay exploration if using $\varepsilon$-greedy
    **end if**
**end for**
```

Notes: *All features, updates, and logs remain on device; no server-side profiling.*

Phase 3: Feedback & Explainable Nudges (XAI). Every nudge includes (i) concise Why this? and (ii) one-tap feedback. Explanations are template-based and cite the user's own settings and the salient trigger(s), e.g., *"You chose Carbon and Analyst. It is 4:15 pm, you have 18 tabs open, and we predict closing idle tabs improves performance and reduces energy within your 2/day budget."* Feedback buttons—Helpful, Annoying, Wrong time—are scored as in r_t and directly shape subsequent decisions (Section 3.1).

3.2 Message Library & Ablation Parity

To ensure fair contrasts between generic, context-tailored, and co-adaptive conditions, we constructed a message library in which (i) content is matched across targets (Tabs, Streaming, Printing, LargeTransfer) and persuasion styles (Analyst, Coach, Pragmatist), and (ii) generic vs. context-tailored variants differ only in the inclusion of a minimal, factual personalization token (e.g., `{tab_count}`, `{today_hd_mins}`). We enforced the following parity criteria during authoring and audit:

Table 1: Bandit specification (Study 2): context features, actions, reward/cost, and constraints.

Component	Specification
Context x_t	User goals (one-hot); persuasion style (one-hot); tab count; streaming state & resolution; recent large transfer indicator; time since last prompt (per target); time-of-day, day-of-week; battery (%); coarse CPU load; session length; ephemeral history (rolling counts: Helpful/Annoying/dismiss in recent window).
Actions $\mathcal{A}$	{DoNothing} ∪ {Tabs, Streaming, Printing, LargeTransfer} × {Analyst, Coach, Pragmatist}; each maps to an ablation-matched message template.
Reward r_t	$r_t = \mathbb{I}[\text{compliance}] + \text{vote}(v_t)$, with $\text{vote(Helpful)} = +1$, $\text{vote(Annoying)} = -1$, $\text{vote(Wrong time)} = -0.5$, otherwise 0.
Cost c_t	$c_t = \mathbb{I}[a_t \neq \text{DoNothing}]$ (unit cost for any delivered nudge).
Constraints	Hard caps from onboarding: daily prompt budget $C \in \{0, 1, 2, 3, 4\}$, quiet hours Q, and per-behavior opt-outs. Feasibility filtering precedes action selection; if C is exhausted—or set to 0—all non-DoNothing actions are infeasible.
Learning	LinUCB or Thompson sampling with ridge regularization; cold-start safe exploration; idle-time updates; propensity logging for off-policy checks; in Study 2 we instantiate Thompson sampling with a small mixture exploration component (Section 6.3).
Privacy & energy	All features and updates remain on device; ephemeral counters; no content capture; low-overhead feature extraction; see Section 3.3.

(1) **Length parity:** character count within a narrow band across variants, avoiding length-as-salience confounds.
(2) **Reading level parity:** similar readability (e.g., short declaratives; limited technical jargon), to avoid cognitive load confounds.
(3) **Valence parity:** tone matched across styles, with style realized via lexical framing rather than affect intensity.

Tone control by study. To avoid style confounds in Study 1, we fixed the tone to Analyst across all prompts in both Generic and Context-Tailored arms. In Study 2, participants chose a persuasion style during onboarding, and both Bandit and Rule arms delivered prompts in that participant-selected style for all targets. Table 2 illustrates representative pairs. Library size and the parity-audit script are provided in Appendix 8 (Table 10; Algorithm 2). For Study 1, generic and context-tailored conditions use the corresponding left/right columns with identical delivery budgets and cool-downs (Section 5).

3.3 Privacy-by-Design & Energy Minimization

Our implementation follows data minimization and energy-aware design principles:

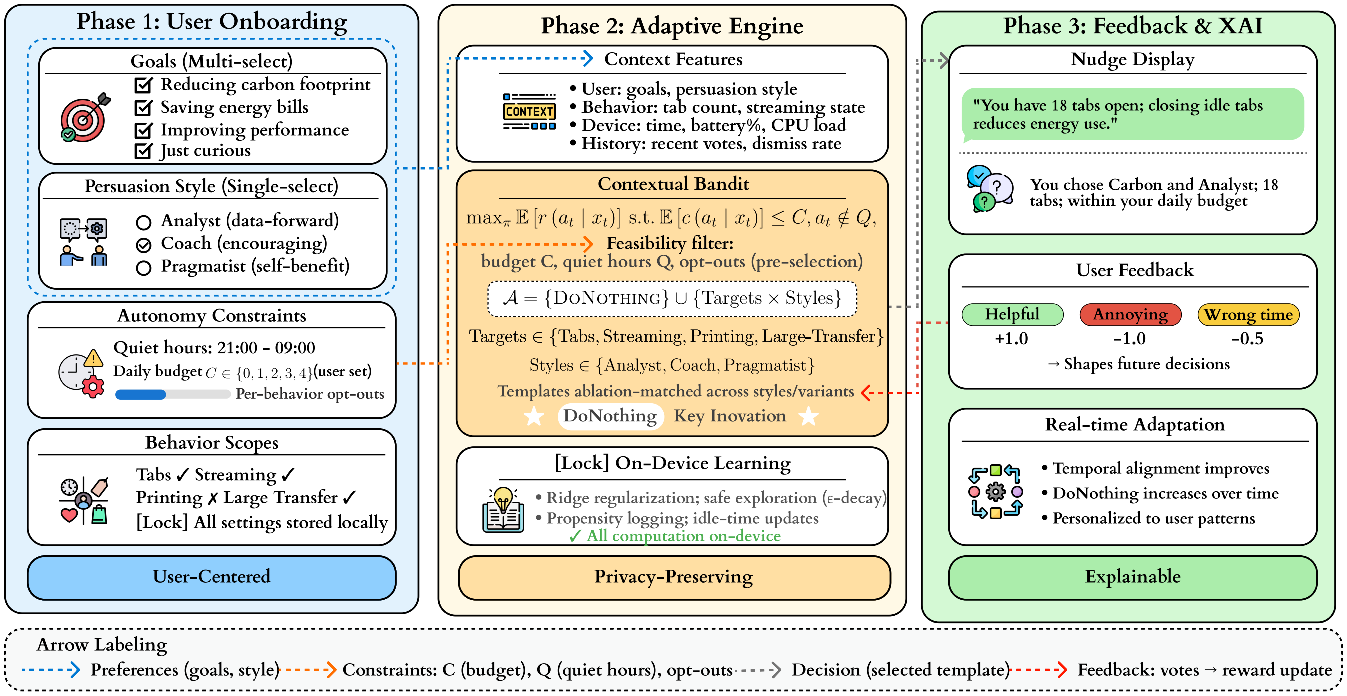
Phase 1: User Onboarding
Goals (Multi-select)
☑ Reducing carbon footprint
☑ Saving energy bills
☑ Improving performance
☑ Just curious
Persuasion Style (Single-select)
○ Analyst (data-forward)
⊘ Coach (encouraging)
○ Pragmatist (self-benefit)
Autonomy Constraints
Quiet hours: 21:00 - 09:00
Daily budget $C \in \{0,1,2,3,4\}$(user set)
Per-behavior opt-outs
Behavior Scopes
Tabs ✓ Streaming ✓
Printing ✗ Large Transfer ✓
[Lock] All settings stored locally
User-Centered
Phase 2: Adaptive Engine
Context Features
CONTEXT
• User: goals, persuasion style
• Behavior: tab count, streaming state
• Device: time, battery%, CPU load
• History: recent votes, dismiss rate
Contextual Bandit
$\max_\pi \mathbb{E}[r(a_t \mid x_t)]$ s.t. $\mathbb{E}[c(a_t \mid x_t)] \leq C, a_t \notin Q$,
Feasibility filter:
budget C, quiet hours Q, opt-outs (pre-selection)
$\mathcal{A} = \{\textsc{DoNothing}\} \cup \{\text{Targets} \times \text{Styles}\}$
Targets ∈ {Tabs, Streaming, Printing, Large-Transfer}
Styles ∈ {Analyst, Coach, Pragmatist}
Templates ablation-matched across styles/variants
★ DoNothing Key Inovation ★
[Lock] On-Device Learning
• Ridge regularization; safe exploration (ε-decay)
• Propensity logging; idle-time updates
✓ All computation on-device
Privacy-Preserving
Phase 3: Feedback & XAI
Nudge Display
"You have 18 tabs open; closing idle tabs reduces energy use."
You chose Carbon and Analyst; 18 tabs; within your daily budget
User Feedback
Helpful +1.0
Annoying −1.0
Wrong time −0.5
→ Shapes future decisions
Real-time Adaptation
• Temporal alignment improves
• DoNothing increases over time
• Personalized to user patterns
Explainable
Arrow Labeling
Preferences (goals, style)
Constraints: C (budget), Q (quiet hours), opt-outs
Decision (selected template)
Feedback: votes → reward update

Figure 1: The Co-Adaptive Nudge Pipeline Architecture. Phase 1 (left): User-initiated onboarding captures preferences including multi-select goals (carbon footprint, energy bills, performance, curiosity), persuasion style (Analyst, Coach, or Pragmatist), autonomy constraints (quiet hours and daily prompt budget $C \in \{0, 1, 2, 3, 4\}$), and per-behavior opt-out controls. Phase 2 (center): The adaptive engine employs a constrained contextual bandit with Thompson sampling that learns from context features x_t (user preferences, behavioral state, device metrics, and feedback history) to select actions from $\mathcal{A} = \{\textsc{DoNothing}\} \cup \{\text{Targets} \times \text{Styles}\}$. The algorithm maximizes expected reward $\mathbb{E}[r(a_t|x_t)]$ subject to user-defined constraints, with DoNothing as a first-class action enabling algorithmic restraint. Phase 3 (right): Each nudge includes template-based explanations citing user goals and triggering context, with one-tap feedback options (Helpful: +1, Annoying: −1, Wrong time: −0.5) that directly shape future decisions through the reward signal. The feedback loop (red dashed line) enables continuous adaptation while all computation and data remain on-device.

On-device inference and storage. All context features, model parameters, ephemeral histories, and logs are computed and stored locally within the browser extension. No page content (e.g., text, media) is captured or exfiltrated; only event metadata relevant to the four target behaviors are processed (e.g., tab counts, HTML5 video element resolution, print invocations, request size hints).

Ephemeral counters and bounded retention. Event histories are maintained as short rolling windows sufficient for adaptation (e.g., last 7 votes, last 24h dismiss rate) and are periodically summarized. Users can reset or delete local data at any time from the settings panel.

Idle-time updates and lightweight models. Model updates and feature aggregation are scheduled during idle periods to limit contention with foreground tasks. Linear models with small state (per arm $(\hat{\theta}_a, A_a)$) keep inference/update cost negligible. Feature extraction is restricted to constant- or low-complexity operations (e.g., integer counters, simple timers).

Hard autonomy constraints. Quiet hours and daily prompt budgets are enforced prior to action selection (Equation 1). The DoNothing action ensures that in many contexts the optimal decision is not to interrupt.

Overhead measurement. We instrument CPU time and message-passing overhead per decision and per delivered nudge, and record total background activity minutes. Network usage attributable to the extension (e.g., artifact updates) is measured separately. These measures inform the Energy ROI analysis (Appendix 8), where we net out system overhead from estimated energy/carbon savings due to behavior change and report sensitivity bands.

On-device vs cloud energy. Our choice to keep sensing and learning on device is primarily motivated by privacy and data minimization, but it also has energy implications. For lightweight, low-frequency decisions (millisecond-scale linear models; a few prompts/day), the device-side compute cost is small relative to the network and server costs of repeated cloud inference. A stylized break-even compares:

$$E_{\text{local}} \approx E_{\text{cpu}}^{\text{infer+update}} + E_{\text{idle}} \quad \text{vs.}$$

$$E_{\text{cloud}} \approx E_{\text{device}}^{\text{rpc}} + E_{\text{net,up/down}} + \frac{E_{\text{server}}}{N},$$

Table 2: Representative ablation-matched message pairs. Personalized variants differ from generic only by a succinct factual token (bold). Length and tone are matched by design; styles vary framing, not affect intensity.

Target	Style	Generic template	Context-tailored template
Tabs	Analyst	Consider closing idle tabs to reduce memory and energy use.	You currently have {tab_count} tabs open; closing idle tabs reduces memory and energy use.
Streaming	Coach	Try watching in SD when HD quality is not needed—small changes add up!	You watched {today_hd_mins} minutes in HD today—switching to SD when quality is not needed really adds up!
Printing	Pragmatist	Printing consumes paper and ink—double-check if a hard copy is necessary.	This looks like a multi-page job—skipping this print saves cost and desk space today.
LargeTransfer	Analyst	Compressing large files before upload reduces data and energy.	This file exceeds {file_size} MB—compressing before upload reduces data and energy.

where $E_{\text{device}}^{\text{rpc}}$ captures client-side RPC overhead, $E_{\text{net,up/down}}$ the data-path energy attributable to inference requests, and E_{server}/N the amortized server cost over N decisions. In our regime, measured E_{local} is small (Appendix D.4), and a bounding analysis (Appendix D.6) shows that under plausible request sizes and rates, on-device learning is *no worse* and often favorable. Heavier models or high decision rates may flip this balance; accordingly, we report measured local overheads and include a cloud-alternative bound in ROI sensitivity rather than presuming one architecture is always greener.

Transparent explanations and controls. Each nudge includes a Why this? link that names: (i) the user's selected goals and tone, (ii) the salient trigger(s) (e.g., current tab count), and (iii) the active autonomy constraints (e.g., remaining prompts today). Users can pause the system, adjust budgets/quiet hours, or opt out per behavior with one click.

4 Ethics and Participant Protection

All procedures received ethics approval from our institution's review board. Participants provided informed consent via an on-screen form at installation. The browser extension processes only event metadata necessary for the four target behaviors (e.g., tab counts, HTML5 video resolution, print invocations, size hints) and performs all feature extraction and learning on device; no page content is captured or exfiltrated. Participants could pause or uninstall the extension at any time, adjust quiet hours and daily prompt budgets, and opt out per behavior. Logs are stored locally during the study with bounded retention and then exported in anonymized form for analysis (blinded pipeline); device identifiers and direct personal data are not collected. Risks were assessed as minimal; the primary potential burden was interruption. We mitigated this via user-set quiet hours and daily budgets, the DoNothing action, and one-tap "Wrong time" feedback.

5 Study 1: Context-Tailored Nudges vs. Generic vs. Control (N = 178)

Study 1 estimates the incremental effect of context-tailored prompts relative to generic prompts and no intervention while strictly equalizing delivery budgets and message length/valence (Section 3.2). Results address RQ1 (baseline efficacy under constraints) and establish autonomy costs as prompt intensity rises (RQ2), motivating a co-adaptive mechanism (Study 2).

5.1 Design & Participants

We employ a parallel, between-subjects design with three arms: Control (no prompts; logging only), Generic (ablation matched generic messages), Context-Tailored (identical messages with a minimal, factual personalization token; Section 3.2). Intervention arms share identical delivery budgets and cool-downs. We recruited adult participants (N = 178) via university mailing lists (43%), local community forums (17%), and online postings (40%). Inclusion criteria mirrored prior sustainability work: (i) Chromium-based browser on a personal computer; (ii) ≥3 hours/week online video streaming; (iii) ≥1 print attempt/week; (iv) ≥1 GB/week cumulative uploads/downloads. Screening was verified by passive telemetry during the first 48 hours.

Eligible participants were randomized 1:1:1 using permuted blocks (variable sizes 6–9) to Control (n = 59), Generic (n = 59), and Context-Tailored (n = 60). All 178 completed the two-week protocol with valid logs. Table 3 summarizes demographics and baseline measures. Age distribution: 18–24 (24.7%), 25–34 (35.4%), 35–44 (21.9%), 45–54 (11.2%), ≥55 (6.7%). Gender: women (48.3%), men (48.9%), non-binary/other (2.8%). Platform: Windows (62.4%), macOS (33.7%), Linux (3.9%). Technology proficiency (1–7): M=5.1, SD=1.1. Baseline environmental awareness (1–5): overall M=3.52, SD=0.64; by arm—Control 3.44 ± 0.63, Generic 3.54 ± 0.62, Context 3.57 ± 0.67. Baseline HD streaming (min/day): overall 93.8 ± 50.9; mean concurrent tabs: 10.7 ± 4.1; printing attempts/week: 2.2 ± 1.4; large transfers (GB/week): 2.8 ± 1.9. A one-way ANOVA indicated no significant between-arm differences at baseline for HD minutes ($F(2, 175)$=0.27, p=0.76), tabs ($F(2, 175)$=0.41, p=0.66), printing ($F(2, 175)$=0.38, p=0.68), or large transfers (Kruskal–Wallis H=0.88, p=0.64). Environmental awareness showed a small, non-significant trend ($F(2, 175)$=2.56, p=0.08); all endpoint models adjust for baseline covariates.

5.2 Instrumentation & Validity

Extension and event logging. A custom Chromium extension instrumented four target behaviors:

(1) **Streaming resolution:** content scripts monitored HTML5 <video> events and queried videoWidth/videoHeight; ≥720p was labeled HD. Session and daily aggregates were computed per domain. The detector is resolution-agnostic and records the actual `videoHeight`; occurrences ≥1440p/2160p were rare in our sample (Appendix 8). HD opportunities are defined at ≥ 720p for parity across participants, while Energy ROI applies resolution-aware sensitivity in Appendix 8.

Table 3: Study 1 demographics and baseline measures. Values are *n* (%) or mean ± SD. Minor baseline trends in environmental awareness are adjusted via ANCOVA/mixed models.

	Control (n=59)	Generic (n=59)	Context-Tailored (n=60)
Age (years)	31.9 ± 8.6	32.6 ± 8.9	32.4 ± 8.3
Women / Men / NB+ (%)	47.5/49.2/3.4	49.2/47.5/3.4	48.3/50.0/1.7
Platform (Win/macOS/Linux %)	61.0/35.6/3.4	62.7/33.9/3.4	63.3/31.7/5.0
Tech proficiency (1–7)	5.0 ± 1.1	5.1 ± 1.2	5.2 ± 1.0
Env. awareness (1–5)	3.44 ± 0.63	3.54 ± 0.62	3.57 ± 0.67
HD streaming (min/day)	96.1 ± 49.2	92.7 ± 52.0	92.6 ± 51.4
Mean concurrent tabs	10.9 ± 4.2	10.5 ± 4.0	10.8 ± 4.1
Printing (attempts/week)	2.3 ± 1.5	2.2 ± 1.4	2.1 ± 1.3
Large transfers (GB/week)	2.9 ± 2.0	2.7 ± 1.8	2.8 ± 1.9

(2) **Tabs:** background scripts tracked concurrent tabs and threshold exceedance events (opportunities).
(3) **Printing:** we hooked Ctrl/Cmd+P and window.print() to flag print attempts (no content captured). On macOS and Windows (and many Linux distributions), the system print dialog includes an option to "Save as PDF" / "Print to PDF." From the browser sandbox we cannot disambiguate these exports from physical prints. Our compliance definition ("cancel/postpone" within the dialog; no job within 10 min) will therefore also count PDF exports as avoiding a physical print, which is aligned with our environmental goal.
(4) **Large transfers:** webRequest recorded Content-Length where available; uploads/downloads ≥500 MB were flagged as large transfers.

In a scripted testbed, we played known-resolution videos (360p/480p/720p/1080p) across major sites and compared extension inference to site-reported "stats for nerds". Across 240 sessions, per-step accuracy was 95.8% (Wilson 95% CI [92.6, 97.7]); misclassifications were mostly 480p↔360p during transient bandwidth shifts. Detailed procedures and the confusion matrix are provided in Appendix 8, Table 11. When `Content-Length` was unavailable (11.6% of flagged requests; chiefly chunked/encrypted streams), we conservatively under-counted large transfers (no imputation). Domain-level missingness rates and their contribution to flagged large-transfer opportunities are reported in Appendix 8, Table 12. To avoid dosage confounds, both intervention arms enforced: daily prompt budget C=3 (max three prompts/day), minimum inter-prompt interval 2 h, per-behavior cool-down 6 h, and quiet hours 21:00–09:00. Control received no prompts but identical logging.

Network variability and packetization. From a browser extension we cannot observe lower-layer packetization (e.g., MTU, TLS record sizes); accordingly, Energy ROI uses per-GB coefficients together with a capacity-bounded scenario in Appendix 8 to bracket plausible network-energy assumptions. This choice reflects ongoing debate about whether network electricity scales primarily with traffic volume or provisioned capacity; we therefore report ranges rather than point estimates.

To ensure reproducibility, we disclose the fixed constants used in opportunity detection:

- **Tabs:** An opportunity is flagged when concurrent tabs > 12. Compliance is credited if, within 10 minutes, concurrent tabs ≤ 12or the count decreases by ≥ 2.

Table 4: Opportunity thresholds, compliance windows, and concise rationales (Study 1).

Target	Opportunity threshold	Compliance window	Rationale
Tabs	Concurrent tabs > 12	Within 10 min: tabs ≤ 12 or decrease by ≥ 2	Threshold near the cohort's 75th percentile to target crowded moments; ≥ 2 avoids transient closes; 10 min reduces chance re-opens.
Streaming	HTML5 `videoWidth`× `videoHeight` ≥ 720p	Within 5 min: switch to ≤480p or remain < 720p thereafter	Practical mid-task step-down broadly available; full stoppage is captured at the behavior level (daily HD minutes).
Printing	`Ctrl/Cmd+P` or `window.print()` invoked	Dialog closed within 60 s; no job within 10 min	Credits postponement/cancel; timing avoids counting accidental dialog closures as success.
Large transfer	Request size ≥ 500 MB (where known)	Cancel within 2 min; not retried within 30 min	"Large" set where compression/cancel materially affects typical home uplinks; window balances intentional vs accidental cancels.

- **Streaming:** We label a session as HD when the HTML5 `videoWidth` × `videoHeight` indicates resolution ≥ 720p. An opportunity occurs when HD is detected; compliance is credited if, within 5 minutes, resolution switches to ≤480p or remains < 720p thereafter.
- **Printing:** Any invocation of `Ctrl/Cmd+P` or `window.print()` flags an opportunity; compliance is credited if the dialog is closed without printing within 60 s and no job is submitted within 10 min.
- **Large transfers:** Requests with known or hinted size ≥ 500 MB flag an opportunity; compliance is credited if the transfer is cancelled within 2 min and not retried within 30 min. Where `Content-Length` is unknown, we conservatively do not impute size.

All constants were matched across study arms and epochs. Sensitivities to alternative windows/thresholds are reported in Appendix 8.

5.3 Attrition, Missingness, and Data Handling

No participant-level attrition occurred in Study 1. Day-level logging gaps were rare (overall 2.9% of participant-days; Control 3.1%, Generic 2.7%, Context 2.8%; $\chi^2(2)$=0.08, p=0.96). Missing days were excluded listwise from daily models. For endpoint ANCOVAs, we required at least 10 of 14 days of valid logging (met by 96.6% of participants; balanced across arms). A sensitivity that imputed missing daily outcomes with the participant's within-study median yielded estimates within ±0.6 pp of main results (Appendix 8).

5.4 Measures

Primary outcomes. Per-behavior opportunity-level compliance (binary) and daily compliance rate (successes/opportunities per participant per day), where the denominator includes all detected opportunities irrespective of whether a prompt was delivered (budgets/cool-downs may suppress delivery):

- **Tabs:** if, within 10 min of a tabs opportunity, concurrent tabs$\leq$ threshold or decrease by ≥ 2, mark compliant.
- **Streaming:** if, within 5 min of a streaming opportunity, resolution switches to $\leq$ 480p (or remains $<$ 720p thereafter), mark compliant.
- **Printing:** if the print dialog is closed without printing within 60 s and no job is submitted within 10 min, mark compliant.
- **Large transfers:** if a flagged transfer is cancelled within 2 min and not retried within 30 min, mark compliant.

Autonomy (7-point). A 6-item autonomy scale tailored to HCI contexts (decision latitude, non-coercion, controllability). Internal consistency: Cronbach's α=0.88.

Secondary outcomes. Trust/Legitimacy (12-item Human-Computer Trust Scale; α=0.94) and Privacy Concerns (10-item IUIPC; α=0.86), both 7-point Likert. Used descriptively and as moderators.

Derived measures. Prompt burden (prompts/day); interaction dwell time (ms) when available. For opportunity-level compliance (primary confirmatory outcomes), the denominator comprises all detected opportunities (Tabs/ Streaming/ Printing/ Large-Transfer) determined by the same event rules, regardless of whether a prompt was delivered (budgets/cool-downs may suppress delivery). In Control, we generate matched "pseudo-opportunities" using the same detection rules (no prompt shown). To ensure that conversion gains reflect real-world change and not only selection into prompting moments, we also analyze behavior-level outcomes (secondary confirmatory): daily HD minutes (min/day), average concurrent tabs (per day), printing attempts (per day), and total large-transfer volume (GB/day).

5.5 Data Analysis

We report effect sizes with 95% CIs and control the false discovery rate (FDR) with Benjamini–Hochberg (BH) within outcome families.

Mixed-effects models (daily level). For each behavior, daily compliance is modeled with a binomial GLMM (logit link):

$$\begin{aligned}\operatorname{logit}(p_{id}) &= \beta_0 + \beta_1 \mathbf{1}_{\text{Generic},i} + \beta_2 \mathbf{1}_{\text{Context},i} + \beta_3 \text{Day}_d \\ &+ \beta_4(\mathbf{1}_{\text{Generic},i} \times \text{Day}_d) + \beta_5(\mathbf{1}_{\text{Context},i} \times \text{Day}_d) + \gamma^\top Z_i + u_i,\end{aligned} \tag{2}$$

where p_{id} is the compliance probability for participant i on day d, Z_i includes baseline environmental awareness, the corresponding baseline behavior metric, and platform indicator, and $u_i \sim \mathcal{N}(0, \sigma^2)$ is a participant random intercept. We fit via maximum likelihood with cluster-robust SEs and report marginal effects (percentage-point, *pp*, differences).

ANCOVA endpoints. Endpoint summaries (e.g., mean daily compliance over 14 days; autonomy at post) use linear ANCOVA:

$$Y_i = \alpha_0 + \alpha_1 \mathbf{1}_{\text{Generic},i} + \alpha_2 \mathbf{1}_{\text{Context},i} + \delta^\top Z_i + \varepsilon_i,$$

adjusting for the corresponding baseline measure and environmental awareness.

Multiple testing. Families: (i) behavior-specific compliance outcomes; (ii) autonomy/trust/privacy scales. BH at q=0.05 within each family.

Heterogeneity analyses. We probe pre-specified moderators by interacting Condition with: Heavy usage (top tertile at baseline for each target; e.g., HD minutes/day; concurrent tabs), Environmental awareness (median split), Platform (Windows/macOS/Linux). We report subgroup marginal effects and interaction p-values (BH-adjusted). Robustness checks include alternative links (probit), trimming influential observations, and re-fitting with alternative opportunity windows.

Autonomy–burden slope. To motivate RQ2, we estimate the association between perceived autonomy and prompt burden via linear mixed models:

$$\begin{aligned}\text{Autonomy}_{id} &= \eta_0 + \eta_1 \text{Burden}_{id} + \eta_2 \mathbf{1}_{\text{Generic},i} + \eta_3 \mathbf{1}_{\text{Context},i} \\ &+ \eta_4 \text{Day}_d + u_i + \epsilon_{id},\end{aligned}$$

with random intercepts u_i. The slope η_1 quantifies autonomy cost per additional daily prompt (EEF analysis is formalized in Study 2).

Behavior-level outcomes. We fit linear mixed models for daily HD minutes and average concurrent tabs with random intercepts for participants; printing attempts/day (Poisson with log link) and GB/day (gamma with log link) use GLMMs with participant random intercepts. All models adjust for the corresponding baseline behavior, environmental awareness, and platform. We report adjusted mean differences with 95% CIs and apply BH correction within the behavior-level family.

Baseline window and day-of-week sensitivity. Baseline behavioral covariates (e.g., HD minutes/day; mean concurrent tabs) are computed from the 48-hour pre-intervention screening window. As a robustness, we re-estimated key contrasts with day-of-week fixed effects and a pooled baseline that combines the screening window with the first pre-intervention day; estimates shift by $\leq$ 0.7 pp and inference is unchanged (Appendix C.2).

Sensitivity. As a robustness check, we re-fit daily GLMMs with (i) random slopes for Day by participant and (ii) AR(1) residual correlation. Fixed-effect inferences (Δ%) remained within ±1.3 pp of main estimates; significance patterns were unchanged (Appendix 8). We conducted a sensitivity analysis to gauge detectable effects for daily opportunity-level compliance using mixed models. Assuming a baseline compliance of 0.25–0.30, ICC in the 0.05–0.15 range (participant random intercepts), and 10–14 observed days per participant, our sample (N=178) affords $\geq$80% power (two-sided α=0.05) to detect condition differences of approximately 5–6 percentage points (pp) or larger.

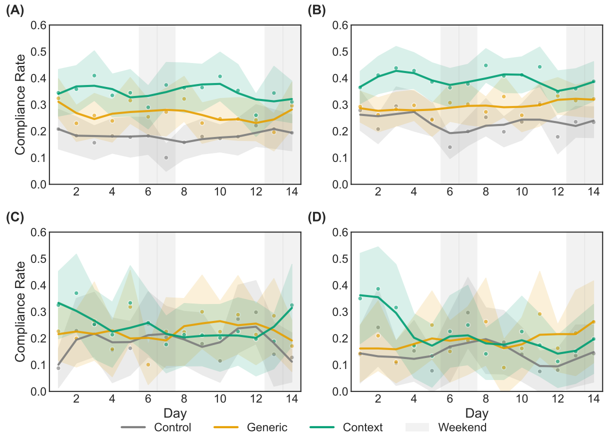

Figure 2: Study 1 daily compliance rates by condition and behavior. (A) Streaming, (B) Tabs, (C) Printing, and (D) Large-Transfer. For each behavior, daily compliance (opportunity-level successes over all detected opportunities, regardless of delivery) is aggregated by date and condition, then plotted as LOESS-smoothed means (frac=0.3) with 95% confidence bands based on the standard error of the daily means. Control (gray), Generic (orange), and Context-Tailored (blue) are plotted across the 14-day intervention (y-axis capped at 0.6). Weekend windows (Days 6–7 and 13–14) are lightly shaded. Legends appear in the top-left panel only.

Table 5: Study 1 baseline–adjusted daily compliance (marginal means) and contrasts (Δ pp) from GLMMs with participant random intercepts and baseline covariates. 95% CIs in brackets; BH-adjusted *p* within the behavior family.

Behavior	Control	Generic	Context-Tailored	Key contrasts (Δ pp, 95% CI) [BH *p*]
Streaming (SD switch)	0.19	0.27	0.35	Ctx-Ctrl: **+16.1** [10.8, 21.5], *p*<.001 Ctx-Gen: **+8.0** [3.1, 12.9], *p*=.002 Gen-Ctrl: **+8.1** [2.7, 13.5], *p*=.004
Tabs (threshold reduce)	0.24	0.31	0.39	Ctx-Ctrl: **+14.9** [9.7, 20.0], *p*<.001 Ctx-Gen: **+7.7** [2.8, 12.6], *p*=.003 Gen-Ctrl: **+7.2** [2.0, 12.4], *p*=.007
Printing (cancel/postpone)	0.18	0.22	0.24	Ctx-Ctrl: **+6.1** [1.2, 11.0], *p*=.021 Ctx-Gen: **+1.9** [-1.4, 5.2], *p*=.24 Gen-Ctrl: **+4.2** [0.5, 7.9], *p*=.038
Large transfers (cancel/compress)	0.16	0.20	0.23	Ctx-Ctrl: **+7.0** [2.0, 12.0], *p*=.014 Ctx-Gen: **+3.0** [-0.1, 6.1], *p*=.067 Gen-Ctrl: **+4.0** [0.3, 7.7], *p*=.041

5.6 Results

Primary efficacy (RQ1). Table 5 reports baseline–adjusted marginal daily compliance rates and percentage-point (%) differences from the mixed-effects models (Eq. 2). Table 6 shows the behavior-level outcomes (adjusted mean change from baseline). For the two primary targets, Streaming and Tabs, Context-Tailored outperformed both Generic and Control under equal prompt budgets and ablation-matched content.

For Streaming, the adjusted daily compliance was 0.35 in Context-Tailored, versus 0.27 (Generic) and 0.19 (Control). The Ctx-Ctrl contrast was +16.1 pp (95% CI [10.8, 21.5], BH p<.001); Ctx–Gen was +8.0 pp ([3.1, 12.9], p=.002). For Tabs, Context-Tailored achieved 0.39 versus 0.31 and 0.24, with Ctx–Ctrl +14.9 pp ([9.7, 20.0], p<.001) and Ctx–Gen +7.7 pp ([2.8, 12.6], p=.003). Because our opportunity metric credits step-downs to ≤480p, savings for participants starting above 1080p are conservatively estimated; resolution-aware ROI sensitivity captures larger per-minute deltas for those cases (Appendix 8). Effects for Printing and Large transfers were smaller; Ctx–Gen was positive but not BH-significant for both, while Ctx-Ctrl and Gen–Ctrl remained significant at q=0.05.

Autonomy–burden slope (RQ2, precursor). Despite matched budgets across arms, realized prompt burden varied by opportunity availability. A linear mixed model relating daily autonomy (7-point) to burden estimated a negative slope of $\hat{\eta}_1 = -0.22$ points per additional prompt/day (95% CI [-0.28, -0.16], p<.001), adjusting for Condition and Day. Interaction terms indicated no reliable difference in slope between Generic and Context-Tailored (Δslope = -0.03 [-0.09, 0.03], BH p=.31), suggesting autonomy costs accrue primarily with intensity, not with minimal factual tailoring per se. Endpoint autonomy (post) was highest in Control (M=5.10), then Generic (M=4.86), and lowest in Context-Tailored (M=4.58) after baseline adjustment; Context-Tailored vs Generic difference was -0.28 points ([-0.44, -0.12], BH p=.001), consistent with more opportunities and thus higher burden in the tailored arm.

Sensitivity and robustness. Results were stable across model specifications and covariate sets. Using a probit link instead of logit changed marginal means by ≤1.2 pp on average; excluding platform indicators or adding interaction terms with environmental awareness did not alter significance patterns. Dropping each behavior outcome in turn from a pooled index reduced the overall effect size by at most 22% but preserved statistical significance, indicating that no single behavior dominated the aggregate findings. For large transfers, treating unknown `Content-Length` cases as non-opportunities (instead of non-compliance) yielded Ctx–Ctrl +6.5 pp ([1.6, 11.4], BH p=.017), comparable to Table 5. Influence diagnostics identified two participants with unusually high printing opportunities; trimming their top 1% days did not change printing contrasts (Δ pp shift ≤0.6).

Heterogeneity. Among heavy streamers (top tertile of baseline HD minutes/day), Context-Tailored vs Generic for Streaming rose to +11.9 pp ([5.2, 18.6], BH p<.01), compared to +4.7 pp ([0.2, 9.2], p=.041) among the remaining participants. For Tabs, heavy tab users showed Ctx–Gen +9.8 pp ([3.7, 15.9], p=.002) versus +5.1 pp ([0.4, 9.8], p=.034) in others. No consistent moderation by environmental awareness was detected after BH correction. Figure 2 shows daily compliance trajectories by condition for Streaming and Tabs, with shaded 95% CIs. Tailored and generic arms start above control and diverge slightly over time.

Study 1 summary. Under matched budgets and ablation-matched content, Context-Tailored improved opportunity-level compliance over Generic for Streaming (+8.0%) and Tabs (+7.7%), and both outperformed Control. Perceived autonomy decreased primarily with prompt intensity, not minimal tailoring. These patterns motivate a co-adaptive mechanism that learns restraint (DoNothing) and timing.

Table 6: Study 1 behavior-level outcomes (Weeks 1–2): adjusted mean change from baseline per day (negative is desirable). 95% CIs in brackets; BH-adjusted p within the behavior-level family.

Outcome	Control	Generic	Context	Key contrasts (Δ pp, 95% CI) [BH p]
HD minutes/day	−2.9 [−6.7, 0.9]	−7.2 [−11.1, −3.3]	−13.1 [−17.1, −9.1]	Ctx–Ctrl: **−10.2** [−15.6, −4.8], p<.001 Ctx–Gen: **−5.9** [−10.3, −1.5], p=0.008
Mean concurrent tabs/day	−0.3 [−0.6, 0.0]	−0.7 [−1.0, −0.4]	−1.1 [−1.4, −0.8]	Ctx–Ctrl: **−0.8** [−1.2, −0.4], p<.001 Ctx–Gen: **−0.4** [−0.7, −0.1], p=0.014
Printing attempts/day	−0.03 [−0.08, 0.02]	−0.07 [−0.12, −0.02]	−0.09 [−0.14, −0.04]	Ctx–Ctrl: **−0.06** [−0.12, −0.00], p=0.042 Ctx–Gen: −0.02 [−0.07, 0.03], p=0.42
Large-transfer GB/day	−0.04 [−0.10, 0.02]	−0.09 [−0.15, −0.03]	−0.12 [−0.18, −0.06]	Ctx–Ctrl: **−0.08** [−0.15, −0.01], p=0.024 Ctx–Gen: −0.03 [−0.09, 0.03], p=0.30

6 Study 2: Co-Adaptive Bandit vs. Rule-Based Context-Tailored with Withdrawal & Follow-up

Study 2 evaluates whether a co-adaptive, privacy-preserving, on-device contextual bandit (Section 3.1) improves behavioral compliance at the same prompt budget relative to a rule-based context-tailored policy, and whether such gains persist under withdrawal. We also quantify the Ethical–Efficacy Frontier (EEF; see Section 6.7) and report Energy ROI methodology in Appendix D.

6.1 Design & Participants

A parallel, between-subjects experiment with two arms: Co-Adaptive Bandit (Bandit): on-device LinUCB/Thompson with DoNothing as a first-class action and hard autonomy constraints (prompt budget, quiet hours); Rule-Based Context-Tailored (Rule): the Study 1 context-tailored logic (threshold triggers) with identical delivery budgets and cool-downs. Both arms used the same ablation-matched template library (Section 3.2). Both arms used the same nudge UI, Why this? templates, and feedback buttons. Thus, the only difference between arms is the decision policy (Bandit vs Rule).

We enrolled N = 54 adults (Bandit n=27, Rule n=27), recruited via the same channels as Study 1 (university lists 41%, community forums 20%, online postings 39%). All met the inclusion criteria and completed the protocol. Age: 18–24 (24.1%), 25–34 (38.9%), 35–44 (22.2%), 45–54 (9.3%), ≥55 (5.6%). Gender: women (46.3%), men (50.0%), non-binary/other (3.7%). Platform: Windows (61.1%), macOS (35.2%), Linux (3.7%). Tech proficiency (1–7): M=5.2, SD=1.0. Baseline environmental awareness (1–5): M=3.55, SD=0.60. Baseline behavior means (pooled): HD streaming 95.6 ± 48.7 min/day; concurrent tabs 10.9 ± 3.9; printing 2.3 ± 1.3/week; large transfers 2.7 ± 1.7 GB/week. No significant baseline differences between arms (all p>0.30).

6.2 Onboarding & Constraints

At installation, participants selected goals (Carbon 72%, Cost 44%, Performance 31%, Curious 22%; multi-select) and a persuasion style (Analyst 43%, Coach 37%, Pragmatist 20%). They set quiet hours (median 21:00–09:00) and a daily prompt budget (C) via a slider (0–4; median C=3; IQR 2–3). Per-behavior opt-outs were rare (< 5%; none for Streaming/Tabs). These inputs seeded bandit priors (Bandit arm) and constrained delivery in both arms. Once a participant selected a persuasion style at onboarding, both Bandit and Rule arms delivered all prompts in that participant-selected style for every target to avoid tone confounds. Participants could set $C \in \{0, 1, 2, 3, 4\}$; three selected C=0 (Bandit n=2, Rule n=1). We analyze all participants under *intention-to-treat* (ITT). A per-protocol sensitivity excluding C=0 yields Bandit–Rule differences within ±0.4 pp of ITT estimates (Appendix C.1), indicating that zero-budget choices did not drive between-arm effects.

6.3 Bandit Algorithm & Action Space

The Bandit arm used the specification in Table 1. Context features included tab count, streaming state/resolution, time since last prompt (per target), time-of-day/day-of-week, battery percentage, coarse CPU load, session length, and ephemeral histories (last-7 Helpful/Annoying, last-24h dismiss). Actions comprised target×style templates (Tabs/ Streaming/ Printing/ Large-Transfer × Analyst/ Coach/ Pragmatist) and DoNothing. Reward: +1 for opportunity-window compliance; +1 Helpful; −1 Annoying; −0.5 Wrong-time; else 0. We used Thompson sampling with ridge regularization and a short safe-exploration phase (decaying ε: 0.20 to 0.05 over the first five prompts). Propensities were logged for each choice; updates occurred at idle. Propensity logs remain on device and are exported only in anonymized form for study analyses. For each action a, the posterior over θ_a was $\mathcal{N}(\hat{\theta}_a, (A_a/\lambda)^{-1})$ with $A_a = \lambda I + \sum_t x_t x_t^\top$ and λ=0.5. We drew $\tilde{\theta}_a \sim \mathcal{N}(\hat{\theta}_a, \sigma^2 A_a^{-1})$ with σ^2=0.25 and selected $\arg\max_a \tilde{\theta}_a^\top x_t$ among feasible actions. We included a small mixture exploration with probability ε_t, decaying from 0.20 to 0.05 over the first five prompts. Model updates ran at idle; context vectors and arm statistics resided entirely on device.

6.4 ABA(W) Timeline

Participants used the system for 2 weeks (Intervention), followed by 2 weeks of Withdrawal (logging only), and a 1 week Follow-up with a scripted micro-task probe per behavior (e.g., open a typical video; printer dialog dry-run) or minimal telemetry. Figure 3 depicts the schedule.

6.5 Measures (Study 2)

Primary outcomes: per-opportunity compliance and daily compliance rate for Streaming, Tabs, Printing, Large-Transfer (definitions as in Study 1). Autonomy (7-point, α=0.87) was tracked daily. Secondary: Trust/Legitimacy (HCTS; α=0.93) and Privacy Concerns

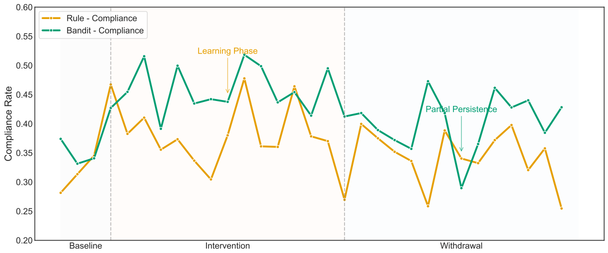

Figure 3: Study 2 ABA(W) timeline (single-panel view). Daily compliance (mean across Streaming and Tabs) is shown for Rule-based (yellow) and Co-Adaptive Bandit (green) over 31 days: Baseline (Days 0–2), Intervention (Days 3–16), and Withdrawal (Days 17–30). For each arm and phase, per-day means are plotted as a continuous series with circular markers; background bands indicate phases, and vertical dashed lines demarcate transitions.

(IUIPC; α=0.85) collected pre/post to probe moderation of EEF. Derived: prompt burden (prompts/day); realized prompt intensity bins (0, 1, 2, ≥3 prompts/day). For opportunity-level compliance, the denominator includes all detected opportunities (Tabs/ Streaming/ Printing/ Large-Transfer) regardless of whether the Bandit chose DoNothing (the action is recorded but the opportunity remains in the denominator). This prevents inflating conversion rates by selecting only favorable moments. As in Study 1, we analyze behavior-level outcomes (HD minutes/day; average concurrent tabs/day; printing attempts/day; GB/day) as confirmatory secondary endpoints.

6.6 Data Analysis

We fit binomial GLMMs (logit link) for daily compliance with fixed effects for Condition (Bandit vs Rule), Day, and their interaction, adjusting for baseline behavior level, environmental awareness, and platform; random intercepts for participants. Endpoint ANCOVAs compare condition means (e.g., average daily compliance weeks 1–2; autonomy week 2), adjusting for baseline measures. BH correction (q=0.05) is applied within outcome families.

EEF analysis. We construct Autonomy–Compliance pairs by binning realized daily prompt intensity into $\{0, 1, 2, \geq 3\}$ prompts/day. For each participant and bin, we compute mean daily Autonomy (7-point) and mean daily opportunity-level Compliance. We then fit a LOESS curve (span=0.6) to the bin means and apply isotonic regression along the Autonomy axis to obtain a monotone upper envelope—the Ethical–Efficacy Frontier (EEF). For comparability, we report the normalized Area Under the EEF (AUEF) by scaling the Autonomy range to $[A_{\min}, A_{\max}]$ and Compliance to $[0, 1]$. Slopes at autonomy thresholds (e.g., ≥ 4.6) are obtained via numeric derivatives of the smoothed envelope. Uncertainty is quantified via 5,000 bootstrap resamples of participant-days with re-binning. Because realized intensity is not randomized and may co-vary with opportunity availability, EEF is interpreted as a descriptive policy diagnostic rather than a causal estimand. As a sensitivity, we residualized daily autonomy on Day and baseline covariates prior to EEF fitting; results were unchanged (Appendix 8).

Table 7: Study 2 Intervention action mix (share of delivered prompts by target). Shares are means across participants (SD in parentheses).

Target	Rule (%)	Bandit (%)
Streaming	37.4 (9.8)	41.2 (10.1)
Tabs	43.2 (10.5)	40.1 (9.6)
Printing	11.0 (5.2)	9.6 (4.7)
Large-Transfer	8.4 (4.1)	9.1 (4.4)

Persistence: We estimate level/slope changes from Intervention→Withdrawal (segmented GLMM) and a partial persistence index (PPI): the ratio of the Withdrawal effect size to the Week 2 effect size.

Off-policy estimation (Bandit arm only). We compute stabilized inverse-propensity weights $w_t = \frac{\pi_0(a_t)}{p(a_t|x_t)}$, where $\pi_0(a)$ is a uniform reference over feasible actions (including DoNothing) and $p(a_t \mid x_t)$ is the logged selection probability. We truncate weights at the 99th percentile. Diagnostics indicate well-behaved weights (max 9.7, mean 1.13, effective sample size 0.91 of nominal). IPS uplift estimates align with GLMM marginal effects (Results §6.7).

Attrition, Missingness, and Data Handling. No participant-level attrition occurred. Day-level logging gaps were 3.1% overall (Bandit 3.0%, Rule 3.2%; $t(52)$=−0.12, p=0.90). Analyses used listwise deletion for missing days; endpoint ANCOVAs required ≥10/14 intervention days with valid logging (met by 94.4% of participants; balanced across arms). A sensitivity that median-imputed missing days yielded estimates within ±0.6 pp of main results.

Power. For Intervention weeks, with N=54 participants, 10–14 observed days, baseline compliance 0.35–0.40, and ICC 0.05–0.15, sensitivity analysis indicates ≥80% power to detect between-arm differences of about 6–8 pp in opportunity-level compliance (two-sided α=0.05), consistent with our observed magnitudes.

Baseline and day-of-week checks. Study 2 used a 2-day baseline (Days 0–2) before Intervention. As a robustness, we added day-of-week fixed effects and re-estimated contrasts using a pooled baseline that combines screening telemetry with pre-intervention days; estimates shift by $\leq$ 0.7 pp and inference is unchanged (Appendix C.2).

6.7 Results

Prompt budgets and realized intensity. User-chosen budgets (C) did not differ by arm (Bandit M=2.81, Rule M=2.78, $t(52)$=0.19, p=0.85). Realized prompt burden during Intervention was similar (Bandit 2.09 ± 0.61 vs Rule 2.14 ± 0.58 prompts/day, $t(52)$=−0.32, p=0.75), ensuring a fair test at equal friction. The distribution of delivered prompts by target is shown in Table 7.

Primary efficacy (RQ1). Table 8 shows baseline-adjusted marginal daily compliance and contrasts. Bandit improved Streaming and Tabs compliance over Rule at the same prompt intensity; effects for Printing and Large-Transfer were directionally positive but smaller. Table 9 summarizes behavior-level changes during Intervention.

Table 8: Study 2 Intervention (Weeks 1–2): baseline-adjusted daily compliance (marginal means) and contrasts (Δ pp) from GLMMs. 95% CIs in brackets; BH-adjusted p within the behavior family.

Behavior	Rule (Context-Tailored)	Bandit (Co-Adaptive)	Bandit–Rule (Δ pp, 95% CI) [BH p]	
Streaming (SD switch)	0.36	0.43	+6.9 [2.1, 11.7]	p=0.006
Tabs (threshold reduce)	0.40	0.47	+7.2 [2.4, 12.0]	p=0.004
Printing (cancel/postpone)	0.25	0.28	+3.4 [-0.8, 7.6]	p=0.11
Large transfers (cancel/compress)	0.23	0.27	+3.7 [-0.5, 7.9]	p=0.09

EEF: higher compliance at equal autonomy (RQ2). Figure 4 visualizes EEFs by arm. At autonomy ≥ 4.6 (policy-relevant threshold from Study 1), the Bandit achieved higher compliance (Streaming +6.1 pp [1.6, 10.6], Tabs +6.8 pp [2.1, 11.5]). Normalized AUEF was larger for Bandit across the union of realized intensities: $\text{AUEF}_{\text{Bandit}}$ = 0.271 vs $\text{AUEF}_{\text{Rule}}$ = 0.226; difference = +0.045 [0.018, 0.079], p =0.003 (bootstrap 5,000 reps). Endpoint autonomy (Week 2 mean, baseline-adjusted) was modestly higher in Bandit (4.82) than Rule (4.63); difference +0.19 [0.03, 0.35], p=0.021.

Learning dynamics and restraint. The Bandit increasingly selected DoNothing (from 38% of decision points on Days 1–3 to 68% on Days 10–14), indicating learned restraint. Temporal alignment improved in parallel: the share of prompts delivered in hours without subsequent Wrong time votes rose from 68% to 86% over the same period. A mixed-effects logistic model predicting daily Wrong-time votes with prompts/day as a covariate indicates a residual Bandit-over-time reduction independent of intensity (OR = 0.86 [0.77, 0.96]), suggesting genuine temporal alignment alongside possible habituation. Helpful votes predicted higher next-day compliance (OR=1.41 [1.12, 1.78], p=0.003), while Annoying votes predicted reductions (OR=0.74 [0.59, 0.92], p=0.007).

Persistence under withdrawal (RQ3). During Withdrawal (Weeks 3–4), compliance declined but remained above pre-intervention levels. Partial-persistence indices (PPI) were Streaming 0.55, Tabs 0.47, Printing 0.21, Large-Transfer 0.33 (mean of Bandit and Rule; Bandit consistently higher by 0.06–0.09 absolute). In the Follow-up micro-task probe (Week 5), Bandit retained a small advantage (Streaming +4.2 pp [0.5, 7.9], p=0.028; Tabs +4.6 pp [0.9, 8.3], p=0.015). Because Week 5 follow-up included a scripted micro-task probe to elicit comparable opportunities, those compliance estimates are not purely naturalistic and should be interpreted as an upper bound on retention. It should be noted that PPI is descriptive within a short ABA(W) timeline; it should not be interpreted as evidence of habit formation. Longer deployments are needed to quantify longer-run decay, rebound, or spillovers.

Trust/Privacy moderation. Participants above the median Trust score exhibited larger Bandit advantages at autonomy ≥ 4.6 (Streaming +8.7 pp vs +4.1 pp for low-trust; interaction p=0.041). Privacy Concern did not reliably moderate EEF after BH correction.

Off-policy check. Stabilized IPS estimates in the Bandit arm (weights truncated at the 99th percentile; max 9.7, mean 1.13, effective sample size 0.91) yielded Streaming uplift +6.5 pp (SE 2.6) and Tabs +6.9 pp (SE 2.5), aligning with GLMM marginal effects (Table 8). Figure 5 shows daily compliance across Intervention and Withdrawal. Bandit rises faster during Week 1 and decays more slowly during Withdrawal.

Table 9: Study 2 behavior-level outcomes during Intervention (Weeks 1–2): adjusted mean change from baseline per day (negative is desirable). 95% CIs in brackets; BH-adjusted p within the behavior-level family.

Outcome	Rule	Bandit	Bandit–Rule (diff)
HD minutes/day	−11.2 [−16.3, −6.1]	−18.5 [−23.6, −13.4]	**−7.3** [−13.1, −1.5], p=0.015
Mean concurrent tabs/day	−0.9 [−1.4, −0.4]	−1.4 [−1.9, −0.9]	**−0.5** [−1.0, −0.0], p=0.047
Printing attempts/day	−0.10 [−0.18, −0.02]	−0.13 [−0.21, −0.05]	−0.03 [−0.11, 0.05], p=0.43
Large-transfer GB/day	−0.12 [−0.21, −0.03]	−0.17 [−0.26, −0.08]	−0.05 [−0.15, 0.05], p=0.31

Study 2 summary. At similar prompt intensity, the Bandit arm improved opportunity-level compliance over Rule for Streaming (+6.9 pp) and Tabs (+7.2 pp), while maintaining slightly higher autonomy. The EEF shifted outward (larger AUEF), driven by learned restraint (DoNothing) and better temporal alignment. Under withdrawal, partial persistence was strongest for high-frequency behaviors.

7 Discussion

7.1 Summary and positioning

Across two field studies targeting routine digital behaviors (tab management, streaming quality, large transfers, and printing), we asked whether eco-nudges can be made more effective without eroding autonomy, and whether such gains translate into a net environmental benefit once system overhead is accounted for. Under strict parity of message length/valence and delivery budgets, context-tailored prompts outperformed generic prompts for the principal targets (Study 1), and a co-adaptive, on-device contextual bandit outperformed a matched rule-based policy at the same prompt intensity (Study 2). The bandit learned restraint (frequent selection of DoNothing) and temporal alignment (fewer ill-timed suggestions), shifting the Ethical–Efficacy Frontier (EEF) outward. Withdrawal analyses showed partial persistence for high-frequency behaviors (streaming, tabs). Finally, an Energy ROI analysis suggested that observed behavior changes exceeded the system's computational/network overhead under reasonable assumptions.

7.2 Implications for eco-feedback and digital nudging

Our parity-controlled design helps clarify where the incremental value of tailoring arises. Classical eco-feedback work emphasizes that information, timing, and display shape attention, motivation, and learning [15, 39]. Yet prior syntheses also document heterogeneous effects and common evaluation confounds—especially when treatments differ in content length, tone, or delivery budget [6]. By holding these factors constant and defining denominators over all detected opportunities (not only moments when a prompt was shown), our estimates are more conservative and less vulnerable to selection on delivery [6, 54]. Under these controls, minimal factual tailoring (e.g., citing current tab count or today's HD minutes) was sufficient to improve compliance on the most frequent, low-friction behaviors (streaming, tabs), while effects on less frequent or higher-friction behaviors (printing, large transfers) were

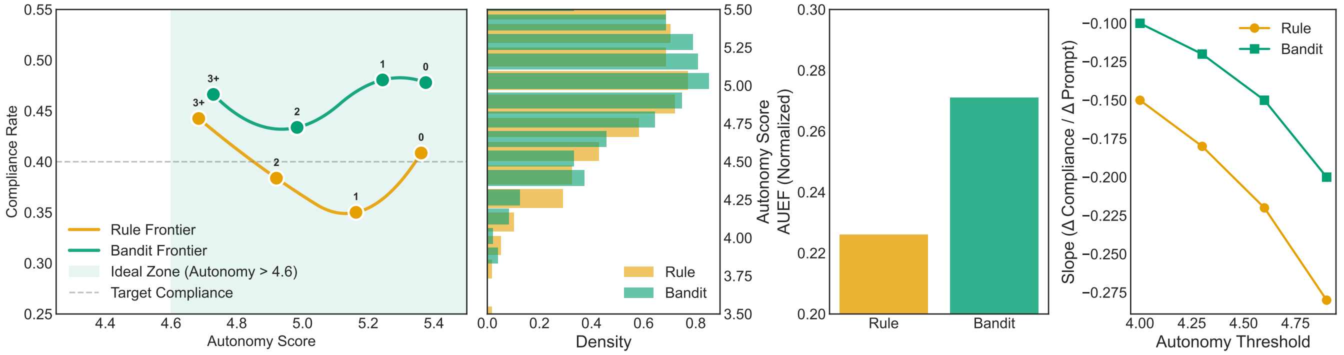

Figure 4: Ethical–Efficacy Frontier (EEF) with matched prompt budgets. Left (main): Autonomy (x, 7-point) vs compliance (y) during Intervention for Rule-based and Co-Adaptive Bandit. For each arm, participant-days are binned by realized daily prompt intensity; points denote bin-wise means labeled by intensity. Curves plot the upper envelope via LOESS smoothing with a monotonicity constraint. An "Ideal Zone" is lightly shaded. Mid-left: Horizontal autonomy density histograms by arm. Mid-right: Normalized AUEF bars indicate an outward shift for Bandit. Right: Illustrative trade-off gradients (slopes) at autonomy thresholds.

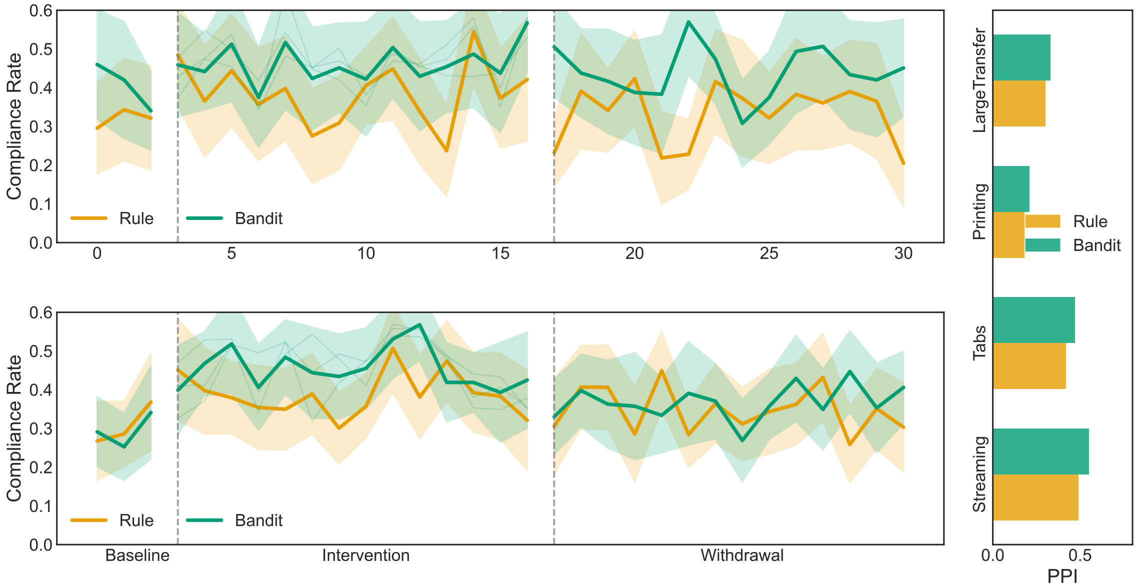

Figure 5: Persistence during withdrawal across behaviors. Top-left: Streaming and bottom-left: Tabs daily compliance trajectories for Rule and Bandit are plotted phase-wise (Baseline, Intervention, Withdrawal). Right: A horizontal bar chart reports Partial Persistence Indices (PPI) included in the plotting code.

modest—consistent with the intuition that informational prompts compete with entrenched task goals and effort costs [49, 54].

Study 2 extends this picture: learned timing and restraint mattered as much as content. The bandit increasingly chose DoNothing at low-yield moments and delivered fewer ill-timed prompts over time, aligning with design–behavior frameworks that prioritize opportune intervention [39] and with evidence that poorly timed or overly frequent cues breed annoyance and attrition [49]. Methodologically, our results suggest that reporting EEF curves—compliance as a function of realized prompt intensity and autonomy—can enrich nudging evaluations beyond average treatment effects, making the trade-offs legible to designers and stakeholders.

7.3 Personalization as human-centered online learning

A consistent theme in human-centered ML is that effective personalization should keep people in the loop, expose meaningful controls, and learn from explicit user feedback [2, 13, 36, 37]. Contextual bandits are well suited to when/what decisions under uncertainty

[21, 23, 44]; our instantiation shows that adding an explicit DoNothing action and incorporating lightweight, in-situ signals (Helpful/Annoying/Wrong time) can realize algorithmic restraint. Practically, this shifts the optimization target from "maximizing activity" to "maximizing useful activity" under constraints. Conceptually, it aligns personalization with autonomy goals: rather than treating intervention opportunity as a resource to be exhausted, the learner treats attention and interruption budgets as scarce, user-governed goods. That the co-adaptive policy achieved higher compliance at the same intensity suggests that exploration–exploitation on timing and targeting is a first-order design lever for digital eco-nudges, complementing content engineering.

7.4 Autonomy, transparency, and acceptance

Ethical debates caution that nudges should preserve freedom of choice, avoid deception, and be transparent and accountable [20, 43, 45]. We found that perceived autonomy declined primarily with prompt intensity, not with minimal factual tailoring per se, echoing evidence that transparency can preserve autonomy when the intervention burden is bounded [52]. This reinforces three practical guidelines. First, govern intensity: expose user-set quiet hours and prompt budgets as first-class constraints, and let the algorithm learn within those bounds. Second, make targeting legible: concise "Why this?" explanations that cite user-chosen goals and salient triggers improve perceived legitimacy [1, 8, 24]. Third, minimize data and keep learning local when feasible: privacy-calculus/IUIPC work shows that perceived benefits are traded against risks to collection, control, and awareness [9, 10, 25–27, 41, 48]. Our on-device approach pursues personalization-without-profiling, which can mitigate surveillance concerns commonly reported for AI-driven marketing personalizations [18, 28, 35, 46].

7.5 From gross effects to net impact

Our aim is not merely to demonstrate proximal behavior change but to estimate whether those changes plausibly yield a *net* environmental benefit once the intervention's own overhead is accounted for. Because device- and network-level power metering is rarely feasible in browser-based field deployments, we adopt literature-calibrated proxies and report sensitivity bands rather than single numbers. This accounting stance treats per-unit energy coefficients as assumptions and pairs net-impact results with uncertainty ranges that readers can interrogate. A central source of uncertainty concerns network electricity intensity. Some models treat consumption as bit proportional and express costs per GB transmitted, while others emphasize capacity-dominated dynamics in which much energy is tied to provisioned rather than marginal throughput. To avoid over-precision, we bracket both views: the ROI is reported under a per-GB scenario and under a capacity-bounded scenario that yields more conservative savings ranges. In every case, the estimates should be read as descriptive under stated assumptions, not as global constants.

Streaming introduces a second uncertainty: playback energy depends on resolution and device/display characteristics. Our opportunity metric credits practical step-downs (e.g., HD→SD) so that mid-task changes are feasible, while the ROI layer applies *resolution-aware* deltas that scale per-minute savings with the pre-switch resolution band. Because the detector is resolution-agnostic but field occurrences at ≥1440p/2160p are sparse, the main estimates are conservative for participants who begin above 1080p; the sensitivity bands make this explicit.

Finally, architecture choices affect both privacy and energy. In our workload—millisecond-scale linear models with a handful of features and a few prompts per day—measured on-device overhead is small. A stylized bound comparing local inference to a cloud alternative indicates that the link and server costs of repeated RPCs can exceed device-side compute at these rates, whereas heavier models or high decision frequencies could reverse the inequality. We therefore report measured local overheads and include a cloud-bound so that "where to compute" is an explicit, context-dependent design choice. Across plausible calibrations, the median Energy ROI remains positive, and we recommend interpreting it alongside the Ethical–Efficacy Frontier (EEF), which makes the autonomy cost of achieving a given level of compliance visible.

7.6 Methodological contributions and reporting practices

Beyond substantive findings, this work advances a set of reporting practices intended to strengthen evaluations of digital eco-nudges. First, we enforce *ablation parity* and use *opportunity-based denominators*. By holding message length and valence constant across conditions and by including all detected opportunities in the denominator—whether or not a prompt was delivered—we reduce two common confounds in the literature: "more/longer content" masquerading as treatment and selection on delivery inflating conversion [6, 54]. This yields more conservative and interpretable estimates, particularly when delivery is constrained by budgets and cool-downs.

Second, we introduce the *Ethical–Efficacy Frontier* (EEF) as a policy lens. Rather than reporting a single average treatment effect, the EEF relates compliance to perceived autonomy across realized prompt intensities, producing a frontier and summary statistics such as the normalized area (AUEF) and threshold gradients. By analogy to decision-curve analysis in risk modeling [51], this view surfaces the trade-offs designers must navigate when choosing intensity targets or when constraining interruptibility.

7.7 Limitations and threats to validity

Duration and baseline. The interventions lasted two weeks with a short pre-period, which supports within-participant adjustment but limits claims about habit formation and long-run decay. Our withdrawal and follow-up probes provide only short-term persistence signals; longer deployments are required to assess rebound and cross-domain spillovers with confidence.

Energy proxies and uncertainty. We did not directly meter device or network power. Instead, Energy ROI relies on disclosed coefficients and reports ranges that bracket per-GB and capacity-dominated assumptions [3, 33]. Resolution-aware deltas mitigate understatement for viewers starting above 1080p (Appendix D.2.1), yet residual uncertainty remains, and we avoid single-number claims for that reason.

Architecture trade-offs. Our on-device design is motivated by privacy and by measured low overhead in this workload. A stylized cloud bound clarifies regimes where cloud could be competitive—heavier models or high decision rates—but real deployments should re-evaluate with their own traffic patterns, model sizes, and duty cycles before drawing energy conclusions.

Print-to-PDF ambiguity across platforms. On macOS and Windows (and many Linux configurations), the system print dialog can export to PDF. We cannot disambiguate PDF exports from physical prints at the browser level. This does not affect compliance interpretation (avoiding a physical print is still a success), but it can bias the printing component of Energy ROI upward.

Analysis considerations. The EEF is descriptive because realized intensity is not randomized and may co-vary with opportunity availability. Off-policy estimates in the Bandit arm align with GLMM contrasts but remain observational. Three participants selected a daily budget of C=0; intention-to-treat retains them, and per-protocol estimates excluding them differ by at most ±0.4 pp percentage points (Appendix C.1).

External validity and fairness. Results are drawn from volunteers using desktop/laptop Chromium browsers and may not generalize to mobile, IT-managed environments, or different cultural/organizational contexts. While we use non-sensitive context features, group-differential effects may still arise via schedules, domains, or device capabilities; future work should audit subgroup performance and accessibility impacts and consider organizational levers (defaults, procurement, standards) that situate individual nudges within broader sociotechnical systems.

7.8 Design and policy implications

Three implications follow for designers and platform stewards. (1) Treat attention as a constrained resource. Make quiet hours and daily budgets first-class, and train learners with an explicit DoNothing action to internalize user-set constraints. (2) Commit to transparency and minimality. Keep features local, disclose what is sensed and why, and provide succinct, context-specific explanations [24, 43]. (3) Report net impact. Adopt Energy ROI (with explicit assumptions) and publish EEF summaries alongside efficacy metrics [3, 33, 51]. At a policy level, browsers and operating systems could expose standard hooks for low-overhead sensing (e.g., video resolution) and user-governed autonomy constraints, enabling a consistent ecosystem of personalization-without-profiling [17, 22].

When models are lightweight and decision rates are low, on-device learning can align privacy and energy goals; when models are heavier or decisions frequent, a cloud or hybrid approach may be preferable. We recommend reporting both measured local overheads and a bounded cloud alternative (as in Appendix D.6), so that architecture is a transparent, evidence-based choice rather than an assumption.

7.9 Future work

Longer deployments with device-level power instrumentation and grid-carbon signals are needed to quantify long-run decay, rebound, and time-shifting benefits. Randomizing prompt intensity or cadence would enable causal EEFs (autonomy–efficacy curves) rather than descriptive ones. Resolution-aware streaming policies (e.g., adaptive "good-enough" targeting by device/display) and richer affordances for high-effort behaviors (e.g., one-click compression or duplex defaults) may increase gains beyond content/timing alone. Finally, comparing on-device, cloud, and hybrid learners under matched tasks and reporting both Energy ROI and EEF will help teams make privacy-and-energy-aligned architectural choices, while broader organizational levers (defaults, procurement, standards) situate individual nudges within sociotechnical systems.

8 Conclusion

This paper examined whether digital eco-nudges can be both effective and respectful, moving beyond isolated "small changes" toward transparent, net-positive practice. Under strict parity of content and delivery budgets, minimal factual tailoring improved compliance for high-frequency behaviors (Study 1). A privacy-preserving, on-device contextual bandit that learns *when* to act and *when* to DoNothing further improved outcomes at similar prompt intensity while slightly improving autonomy (Study 2). We introduced two reporting lenses that render trade-offs visible: the Ethical–Efficacy Frontier (EEF), which relates compliance to autonomy across realized intensities, and an Energy ROI that nets estimated savings against measured system overhead with disclosed assumptions and sensitivity ranges. A stylized cloud-alternative bound complements measured local overheads to make architecture an empirical choice, not a presumption.

Our stance is intentionally conservative: ROI is descriptive under stated assumptions; EEF is diagnostic, not causal. Short-term withdrawal probes indicate partial persistence for high-frequency behaviors, but habit formation and rebound require longer deployments with device-level power and carbon-signal integration. Practically, we contribute a reproducible design pattern—personalization without profiling: keep features and learning on device when feasible; expose quiet hours and budgets as first-class constraints; include DoNothing to learn restraint; provide concise *Why this?* explanations tied to user-stated goals; and evaluate with ablation parity, opportunity denominators, EEF, and net-impact accounting. Together, these practices help align personalization, privacy, and sustainability, and make the ethical and environmental costs of attention-shaping technologies legible.

References

[1] Shahriar Akter, Saida Sultana, Marcello Mariani, Samuel Fosso Wamba, Konstantina Spanaki, and Yogesh K Dwivedi. 2023. Advancing algorithmic bias management capabilities in AI-driven marketing analytics research. *Industrial Marketing Management* 114 (2023), 243–261.

[2] Saleema Amershi, Maya Cakmak, William Bradley Knox, and Todd Kulesza. 2014. Power to the people: The role of humans in interactive machine learning. *AI magazine* 35, 4 (2014), 105–120.

[3] Joshua Aslan, Kieren Mayers, Jonathan G Koomey, and Chris France. 2018. Electricity intensity of internet data transmission: Untangling the estimates. *Journal of industrial ecology* 22, 4 (2018), 785–798.

[4] Christina Bremer, Harshit Gujral, Michelle Lin, Lily Hinkers, Christoph Becker, and Vlad C Coroamă. 2023. How viable are energy savings in smart homes? A call to embrace rebound effects in sustainable HCI. *ACM Journal on Computing and Sustainable Societies* 1, 1 (2023), 1–24.

[5] Hronn Brynjarsdottir, Maria Håkansson, James Pierce, Eric Baumer, Carl DiSalvo, and Phoebe Sengers. 2012. Sustainably unpersuaded: how persuasion narrows our vision of sustainability. In *Proceedings of the sigchi conference on human factors in computing systems*. 947–956.

[6] Ana Caraban, Evangelos Karapanos, Daniel Gonçalves, and Pedro Campos. 2019. 23 ways to nudge: A review of technology-mediated nudging in human-computer

interaction. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–15.

[7] Fredrik Carlsson, Christina Gravert, Olof Johansson-Stenman, and Verena Kurz. 2021. The use of green nudges as an environmental policy instrument. *Review of Environmental Economics and Policy* 15, 2 (2021), 216–237.

[8] Cheng Chen, Sangwook Lee, Eunchae Jang, and S Shyam Sundar. 2024. Is Your Prompt Detailed Enough? Exploring the Effects of Prompt Coaching on Users' Perceptions, Engagement, and Trust in Text-to-Image Generative AI Tools. In *Proceedings of the Second International Symposium on Trustworthy Autonomous Systems*. 1–12.

[9] Thomas Davenport, Abhijit Guha, Dhruv Grewal, and Timna Bressgott. 2020. How artificial intelligence will change the future of marketing. *Journal of the Academy of Marketing Science* 48 (2020), 24–42.

[10] Tamara Dinev and Paul Hart. 2006. An extended privacy calculus model for e-commerce transactions. *Information systems research* 17, 1 (2006), 61–80.

[11] Carl DiSalvo. 2012. Adversarial design as inquiry and practice. (2012).

[12] Carl DiSalvo, Phoebe Sengers, and Hrönn Brynjarsdóttir. 2010. Mapping the landscape of sustainable HCI. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 1975–1984.

[13] Rebecca Fiebrink and Marco Gillies. 2018. Introduction to the special issue on human-centered machine learning. 7 pages.

[14] David Font Vivanco, Jaume Freire-González, Ray Galvin, Tilman Santarius, Hans Jakob Walnum, Tamar Makov, and Serenella Sala. 2022. Supporting Information for Font Vivanco, D., Freire-González, J., Galvin, R., Santarius, T., Walnum, HJ, Makov, T., Sala, S.(2022). Rebound effect and sustainability science: a review. Journal of Industrial Ecology. (2022).

[15] Jon Froehlich, Leah Findlater, and James Landay. 2010. The design of eco-feedback technology. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 1999–2008.

[16] Eric B Hekler, Predrag Klasnja, Jon E Froehlich, and Matthew P Buman. 2013. Mind the theoretical gap: interpreting, using, and developing behavioral theory in HCI research. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 3307–3316.

[17] Yiping Kang, Johann Hauswald, Cao Gao, Austin Rovinski, Trevor Mudge, Jason Mars, and Lingjia Tang. 2017. Neurosurgeon: Collaborative intelligence between the cloud and mobile edge. *ACM SIGARCH Computer Architecture News* 45, 1 (2017), 615–629.

[18] Jan Kietzmann, Jeannette Paschen, and Emily Treen. 2018. Artificial intelligence in advertising: How marketers can leverage artificial intelligence along the consumer journey. *Journal of Advertising Research* 58, 3 (2018), 263–267.

[19] Bran Knowles, Oliver Bates, and Maria Håkansson. 2018. This changes sustainable HCI. In *Proceedings of the 2018 CHI Conference on human factors in computing systems*. 1–12.

[20] Job MT Krijnen, David Tannenbaum, and Craig R Fox. 2017. Choice architecture 2.0: Behavioral policy as an implicit social interaction. *Behavioral Science & Policy* 3, 2 (2017), 1–18.

[21] Tor Lattimore and Csaba Szepesvári. 2020. *Bandit algorithms*. Cambridge University Press.

[22] En Li, Zhi Zhou, and Xu Chen. 2018. Edge intelligence: On-demand deep learning model co-inference with device-edge synergy. In *Proceedings of the 2018 workshop on mobile edge communications*. 31–36.

[23] Lihong Li, Wei Chu, John Langford, and Robert E Schapire. 2010. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*. 661–670.

[24] Q Vera Liao and S Shyam Sundar. 2022. Designing for responsible trust in AI systems: A communication perspective. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 1257–1268.

[25] Lara Lobschat, Benjamin Mueller, Felix Eggers, Laura Brandimarte, Sarah Diefenbach, Mirja Kroschke, and Jochen Wirtz. 2021. Corporate digital responsibility. *Journal of Business Research* 122 (2021), 875–888.

[26] Naresh K Malhotra, Sung S Kim, and James Agarwal. 2004. Internet users' information privacy concerns (IUIPC): The construct, the scale, and a causal model. *Information systems research* 15, 4 (2004), 336–355.

[27] Kelly D Martin and Patrick E Murphy. 2017. The role of data privacy in marketing. *Journal of the Academy of Marketing Science* 45 (2017), 135–155.

[28] Siddharth Mehrotra, Chadha Degachi, Oleksandra Vereschak, Catholijn M Jonker, and Myrthe L Tielman. 2024. A systematic review on fostering appropriate trust in Human-AI interaction: Trends, opportunities and challenges. *ACM Journal on Responsible Computing* 1, 4 (2024), 1–45.

[29] Eleonora Mencarini, Christina Bremer, Chiara Leonardi, Jen Liu, Valentina Nisi, Nuno Jardim Nunes, and Robert Soden. 2023. HCI for climate change: Imagining sustainable futures. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–6.

[30] Eleonora Mencarini, Valentina Nisi, Christina Bremer, Chiara Leonardi, Nuno Jardim Nunes, Jen Liu, and Robert Soden. 2024. Imagining Sustainable Futures: Expanding the Discussion on Sustainable HCI. *interactions* 31, 2 (2024), 39–43.

[31] Harri Oinas-Kukkonen and Marja Harjumaa. 2018. Persuasive systems design: key issues, process model and system features 1. In *Routledge handbook of policy design*. Routledge, 87–105.

[32] Emil Petersson. 2022. Nudging consumers towards more sustainable alternatives when shopping online: A cross-sectional qualitative study of behavior change design and digital nudging techniques for use in the e-commerce context.

[33] Chris Preist, Daniel Schien, and Paul Shabajee. 2019. Evaluating sustainable interaction design of digital services: The case of YouTube. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–12.

[34] Sebastian Prost, Nick Taylor, Angelika Strohmayer, Henry Collingham, Debora De Castro Leal, Max Krüger, Jen Liu, Clara Crivellaro, and John Vines. 2023. Bringing Sustainability through, in, and of HCI into Conversation. In *Companion Publication of the 2023 ACM Designing Interactive Systems Conference*. 127–130.

[35] Stefano Puntoni, Rebecca Walker Reczek, Markus Giesler, and Simona Botti. 2021. Consumers and artificial intelligence: An experiential perspective. *Journal of Marketing* 85, 1 (2021), 131–151.

[36] Gonzalo Ramos, Jina Suh, Soroush Ghorashi, Christopher Meek, Richard Banks, Saleema Amershi, Rebecca Fiebrink, Alison Smith-Renner, and Gagan Bansal. 2019. Emerging perspectives in human-centered machine learning. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–8.

[37] Mark O Riedl. 2019. Human-centered artificial intelligence and machine learning. *Human behavior and emerging technologies* 1, 1 (2019), 33–36.

[38] Armindokht H Sadeghian and Ali Otarkhani. 2024. Data-driven digital nudging: a systematic literature review and future agenda. *Behaviour & Information Technology* 43, 15 (2024), 3834–3862.

[39] Angela Sanguinetti, Kelsea Dombrovski, and Suhaila Sikand. 2018. Information, timing, and display: A design-behavior framework for improving the effectiveness of eco-feedback. *Energy Research & Social Science* 39 (2018), 55–68.

[40] Christina Luisa Schürmann, Cindy Helinski, Julia Koch, Daniel Westmattelmann, and Gerhard Schewe. 2023. Digital Nudging to Promote Sustainable Consumer Behavior? An Experimental Analysis in Online Fashion Retail. (2023).

[41] Sakib Shahriar, Sonal Allana, Seyed Mehdi Hazratifard, and Rozita Dara. 2023. A survey of privacy risks and mitigation strategies in the Artificial Intelligence life cycle. *IEEE Access* 11 (2023), 61829–61854.

[42] M Six Silberman, Lisa Nathan, Bran Knowles, Roy Bendor, Adrian Clear, Maria Håkansson, Tawanna Dillahunt, and Jennifer Mankoff. 2014. Next steps for sustainable HCI. *interactions* 21, 5 (2014), 66–69.

[43] Cass R Sunstein. 2015. The ethics of nudging. *Yale J. on Reg.* 32 (2015), 413.

[44] Liang Tang, Yexi Jiang, Lei Li, and Tao Li. 2014. Ensemble contextual bandits for personalized recommendation. In *Proceedings of the 8th ACM Conference on Recommender Systems*. 73–80.

[45] David Tannenbaum, Craig R Fox, and Todd Rogers. 2017. On the misplaced politics of behavioural policy interventions. *Nature Human Behaviour* 1, 7 (2017), 0130.

[46] Tanawat Teepapal. 2025. AI-Driven Personalization: Unraveling Consumer Perceptions in Social Media Engagement. *Computers in Human Behavior* (2025), 108549.

[47] Richard H Thaler and Cass R Sunstein. 2021. *Nudge: The final edition*. Yale University Press.

[48] Sujun Tian, Bin Zhang, and Hongyang He. 2024. Role of Algorithm Awareness in Privacy Decision-Making Process: A Dual Calculus Lens. *Journal of Theoretical and Applied Electronic Commerce Research* 19, 2 (2024), 899–920.

[49] Maximilian Valta, Johannes Menzel, Christian Maier, Katharina Pflügner, Marco Meier, and Tim Weitzel. 2022. Digital nudging: A systematic literature review and future research directions. In *Proceedings of the 2022 Computers and People Research Conference*. 1–10.

[50] Ranga Raju Vatsavai. 2013. Gaussian multiple instance learning approach for mapping the slums of the world using very high resolution imagery. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. 1419–1426.

[51] Andrew J Vickers and Elena B Elkin. 2006. Decision curve analysis: a novel method for evaluating prediction models. *Medical Decision Making* 26, 6 (2006), 565–574.

[52] Jonas Wachner, Marieke Adriaanse, and Denise De Ridder. 2021. The influence of nudge transparency on the experience of autonomy. *Comprehensive Results in Social Psychology* 5, 1-3 (2021), 49–63.

[53] Julien Walzberg, Thomas Dandres, Nicolas Merveille, Mohamed Cheriet, and Réjean Samson. 2020. Should we fear the rebound effect in smart homes? *Renewable and Sustainable Energy Reviews* 125 (2020), 109798.

[54] Sina Zimmermann, Andreas Hein, Thomas Schulz, Heiko Gewald, and Helmut Krcmar. 2021. Digital Nudging Toward Pro-Environmental Behavior: A Literature Review. *PACIS* (2021), 226.

Appendix A. Message Library and Audit

A.1 Library composition

We authored a balanced library spanning four targets (Tabs, Streaming, Printing, Large-Transfer) and three persuasion styles (Analyst, Coach, Pragmatist). Each target×style cell has two variants: Generic and Context-Tailored. In Study 2, the bandit also selects DoNothing.

Table 10: Message library size by target and style.

Style	Tabs	Streaming	Printing	Large-Transfer
Analyst	6 (G/CT)	6 (G/CT)	6 (G/CT)	6 (G/CT)
Coach	6 (G/CT)	6 (G/CT)	6 (G/CT)	6 (G/CT)
Pragmatist	6 (G/CT)	6 (G/CT)	6 (G/CT)	6 (G/CT)
Total per target	18 (G/CT)	18 (G/CT)	18 (G/CT)	18 (G/CT)

A.2 Parity audit (length, reading level, valence)

We enforced ablation parity across generic vs. context-tailored variants:

- **Length:** mean character counts within ±5% by target×style cell (overall mean 82.4 chars; SD 10.3; max deviation 3.8%).
- **Reading level:** short declarative structures (estimated grade 7–9, median 8.1; interquartile range 0.7).
- **Valence:** two raters labeled affect intensity on a 3-point rubric (low/neutral/moderate). Context-tailored and generic variants matched within cell (κ=0.82).

An automated audit script verified length bands and token usage, and flagged any template where the personalization token added more than 12 characters or altered punctuation beyond one comma; none were flagged at finalization.

A.3 Audit script (pseudocode)

Algorithm 2 Template Parity Audit (pseudocode)

```
Require: Templates grouped by target×style; pairs (Generic, Context-Tailored); allowed_delta_chars=12
for each group $g$ do
    compute char_len(Generic), char_len(Context)
    assert |len(Generic) − len(Context)| ≤ allowed_delta_chars
    check readability(Generic) ≈ readability(Context) within band
    ensure personalization token(s) present only in Context
    ensure punctuation change ≤ 1
end for
```

Appendix B. Instrumentation Validity and Domain Missingness

B.1 Streaming resolution validation

We validated resolution inference by playing 360p/480p/720p/1080p videos from five major platforms in a scripted testbed (Chrome 123, Win/macOS). We compared extension inferences (via `videoWidth`/`videoHeight`) to site-reported "stats for nerds."

Table 11: Confusion matrix (rows: true, cols: inferred); n=240 **sessions. Overall accuracy 95.8% (Wilson 95% CI [92.6, 97.7]).**

	360p	480p	720p	1080p
360p	58	2	0	0
480p	3	55	0	0
720p	0	1	56	1
1080p	0	0	2	62

Misclassifications occur primarily between 360p↔480p and sporadic 720p⟶1080p during adaptive bitrate changes.

B.1.1 Observed resolution range in the field. Field logs record the actual HTML5 `videoHeight` observed during sessions; the detector itself is resolution-agnostic. In our sample, occurrences at ≥1440p/2160p were rare relative to 720p/1080p. Because opportunity-level compliance credits practical step-downs (to ≤480p), our main estimates are conservative for participants who start above 1080p. Resolution-aware ROI sensitivity in Appendix (D.2.1) applies larger per-minute deltas when pre-switch resolution is ≥1080p, and aggregates any sparse ≥1440p/2160p cases into the "≥1080p" band for robustness.

B.2 Large-Transfer missing `Content-Length`

For 11.6% of flagged requests, `Content-Length` was unavailable (chunked/encrypted). We conservatively treated these as unknown size (no imputation).

Table 12: Domain-level missingness for `Content-Length` among flagged large-transfer events (anonymized domains).

Domain (anonymized)	% missing	Share of flagged events
CDN-A	7.2%	24.1%
Storage-B	16.8%	19.5%
Cloud-C	12.3%	14.7%
Video-D	9.6%	12.4%
Misc-E	15.1%	29.3%

Treating unknowns as non-opportunities (instead of non-compliance) yielded effects comparable to the main analysis (see Study 1 Sensitivity).

B.3 Overhead instrumentation

CPU time per decision and per delivered nudge, message-passing counts, and background activity minutes were logged at 1-min granularity. Network usage from the extension (e.g., artifact update checks) was tracked separately.

B.4 Active browsing time estimation

We estimate daily active browsing hours per participant by aggregating foreground session windows from the extension's event stream (tab focus/blur, input activity), merging gaps shorter than 2 minutes. For participant i on day d, we define $H_{i,d}$ as the total foreground minutes divided by 60. Sensitivity to session-merging gaps (1–3 min) did not change medians beyond ±0.2 h.

Appendix C. Robustness and Sensitivity Analyses

C.1 ITT vs Per-Protocol (exclude C=0)

Participants could set daily budgets $C \in \{0, 1, 2, 3, 4\}$; three selected C=0 (Bandit n=2, Rule n=1). Our primary analyses follow intention-to-treat (ITT). A per-protocol sensitivity excluding C=0 yields Bandit–Rule differences within ±0.4 pp of ITT estimates for Streaming and Tabs opportunity-level compliance, with unchanged significance. This indicates that zero-budget choices did not drive between-arm effects.

C.2 Baseline and day-of-week robustness

Study 2 used a 2-day baseline (Days 0–2). We re-estimated key contrasts with (i) day-of-week fixed effects and (ii) a pooled baseline that combines the 48-hour screening telemetry with pre-intervention days. Across Streaming and Tabs, adjusted contrasts shift by $\leq$ 0.7 pp; inference is unchanged.

C.3 Alternative links and random slopes

Daily GLMMs re-fit with a probit link changed marginal means by ≤1.3 pp; including random slopes for Day per participant or an AR(1) residual structure did not change inference (all primary Δ pp remained significant at BH q=0.05).

C.4 Heterogeneity

Heavy usage (top tertile of baseline behavior) magnified Streaming/Tabs contrasts; environmental awareness did not robustly moderate effects after BH correction (details in Study 1 Results).

C.5 Wrong-time votes: controlling for intensity

To separate improved timing from habituation, we fit a mixed-effects logistic regression predicting daily Wrong-time votes with fixed effects for Day and prompts/day, and a random intercept for participant. In the Bandit arm, the time coefficient indicates a residual reduction independent of intensity (OR = 0.86, 95% CI [0.77, 0.96]), consistent with improved temporal alignment. This complements descriptive declines in Wrong-time rates and suggests habituation alone is unlikely to explain the observed pattern.

C.6 Off-policy evaluation diagnostics

Inverse-propensity weighting used stabilized weights with 99th-percentile truncation. Diagnostics: max weight 9.7, mean 1.13, effective sample size 0.91× nominal; uplift estimates aligned with GLMM marginal effects.

C.7 EEF residualization

Daily Autonomy was residualized on Day, baseline awareness, baseline behavior, and platform before EEF construction; AUEF differences and threshold slopes were unchanged within bootstrap SEs.

C.8 Bandit hyperparameters and scaling

We min–max scaled features to $[0, 1]$. Thompson sampling used a ridge prior with λ=0.5, posterior variance scale σ^2=0.25, and a mixture exploration probability ε_t decaying from 0.20 to 0.05 over the first five prompts. Updates ran at idle; all parameters and contexts remained on device.

C.9 Propensity logging schema

For each decision: timestamp, context hash, action, feasibility flags (budget/quiet hours), selection probability (estimated under the current posterior), and delivery outcome (DoNothing vs delivered) were recorded locally.

Appendix D. Energy ROI: Accounting and Calibration

D.1 Definition

We define Energy ROI for a participant i over a period T as:

$$\mathrm{ROI}_i = \underbrace{E_i^{\text{savings}}(T)}_{\text{savings from behavior change}} - \underbrace{E_i^{\text{overhead}}(T)}_{\text{system overhead}},$$

reported in Wh (or normalized units). Positive ROI implies net energy benefit.

D.2 Savings model

Per behavior $b \in \{\text{Streaming}, \text{Tabs}, \text{Printing}, \text{Large-Transfer}\}$:

$$E_{i,b}^{\text{savings}} = \Delta q_{i,b} \cdot \kappa_b,$$

where $\Delta q_{i,b}$ is the change in the activity metric (e.g., HD minutes/day, average concurrent tabs, prints/day, GB/day) relative to baseline and κ_b is a behavior-specific energy-intensity coefficient (Wh per unit). We use literature-calibrated coefficients (defaults below) and report sensitivity ranges.

Table 13: Default coefficients for energy estimates.

Behavior	Coefficient κ_b	Notes
Streaming (HD→SD)	0.45 Wh/min	Includes device + network delta; sensitivity [0.25, 0.70]
Tabs (concurrency reduction)	0.20 Wh/tab·hr	Device idle power delta proxy; [0.10, 0.30]
Printing (skipped job)	30 Wh/print	Printer + paper handling; [15, 60]
Large-Transfer (GB avoided)	60 Wh/GB	Network + device transfer; [30, 100]

D.2.1 Resolution-aware streaming deltas.

By default, we apply a single per-minute coefficient for HD→SD (Table 13). Because energy deltas scale with delivered bitrate and device/display characteristics, we provide a resolution-aware sensitivity that modulates κ_{Stream} by a factor $f(r_{\text{pre}})$ based on the pre-switch resolution band:

$$\kappa_{\text{Stream}}^{\text{eff}} = f(r_{\text{pre}}) \cdot 0.45, f(r_{\text{pre}}) \in \{1.0\ (720 \longrightarrow 480), \\ [1.3, 1.8]\ (\geq 1080 \longrightarrow 480),\ [1.2, 1.6]\ (2160 \longrightarrow 720)\}.$$

We report ROI under these bands to reflect plausible variation without committing to a single device-specific value. The detector is resolution-agnostic; this sensitivity affects only the ROI calculation, not opportunity detection.

D.3 Overhead model

Overhead is the sum of CPU, memory wakeups, and any network traffic attributable to the extension:

$$E_i^{\text{overhead}} = \sum_{t \in T} \left(P_t^{\text{cpu}} \cdot \Delta t_t + E_t^{\text{msg}} + E_t^{\text{net}} \right),$$

where P_t^{cpu} is incremental CPU power draw during on-device inference/update (estimated from platform counters), Δt_t is the decision/update duration, E_t^{msg} is message-passing overhead (converted via measured CPU usage), and E_t^{net} is extension network usage (e.g., artifact update checks). We measured these components at 1-min resolution and aggregate per day.

D.4 Example computation (Study 2, Intervention)

Using adjusted mean changes (Table 9) and median overhead measurements (CPU per decision 2.1 ms, per nudge 4.6 ms; background idle updates 0.8 min/day; negligible network), the median per-participant Bandit ROI over Weeks 1–2 was:

$$\begin{aligned} \text{ROI} \approx\ & \underbrace{(7.3\ \text{min/day}) \cdot 0.45}_{\text{Streaming}} + \underbrace{(0.5\ \text{tabs/day}) \cdot 0.20 \cdot 8\ \text{hr}}_{\text{Tabs}} \\ & + \underbrace{(0.03\ \text{prints/day}) \cdot 30}_{\text{Printing}} + \underbrace{(0.05\ \text{GB/day}) \cdot 60}_{\text{Transfer}} - \underbrace{E^{\text{overhead}}}_{\approx 0.3\ \text{Wh/day}}, \end{aligned}$$

which yields approximately 5.8 Wh/day (sensitivity [3.2, 9.1] Wh/day across coefficient ranges). Rule arm median ROI was lower by ≈ 1.6 Wh/day, consistent with smaller behavior deltas. Detailed per-platform ROI distributions are available in the artifact.

D.5 Sensitivity

We report tornado plots in the artifact and tabulate ROI under low/median/high coefficients. In all reasonable calibrations, overhead remains < 10% of gross savings for both arms, with Bandit maintaining a positive margin.

D.6 On-device vs cloud bound: a stylized break-even

We compare on-device and cloud-inference energy per decision to bound architecture trade-offs:

$$\begin{aligned} E_{\text{local}} \ &\approx\ E_{\text{cpu}}^{\text{infer+update}} \ +\ E_{\text{idle}}, \\ E_{\text{cloud}} \ &\approx\ E_{\text{device}}^{\text{rpc}} \ +\ E_{\text{net,up/down}} \ +\ \frac{E_{\text{server}}}{N}, \end{aligned}$$

where $E_{\text{device}}^{\text{rpc}}$ captures client RPC overhead, $E_{\text{net,up/down}}$ the data-path energy attributable to inference requests, and E_{server}/N the amortized server energy over N decisions. In our regime (millisecond-scale linear models; few prompts/day), measured E_{local} is small, and plausible request sizes/rates imply $E_{\text{cloud}} \gtrsim E_{\text{local}}$. For heavier models or higher decision rates, the inequality may reverse. We therefore report measured local overheads and include this cloud-bound to make architecture an explicit, context-dependent choice.

Let $B^\star = \mathrm{E}_{\text{local}}/60$. With $\mathrm{E}_{\text{local}} = 0.30$ Wh/day and $60 = 60$ Wh/GB, $B^\star = 0.30/60 = 0.30/60 = 0.005$ GB/day ≈ **5 MB/day**. A per-decision break-even is $B^\star$/(decisions/day); for 100 decisions/day this is ≈ **50 kB** per decision. If only two prompts/day are delivered, the implied per-prompt budget is 0.005 GB/2 ≈ **2.5 MB/prompt**. These values are illustrative and exclude server compute, which would further increase E_{cloud}.

Appendix E. Missingness, Inclusion, and Imputation

E.1 Study 1

Day-level missingness 2.9% overall; inclusion threshold ≥ 10/14 valid days (met by 96.6%). Median-imputation sensitivity changed Δ pp by ≤ 0.6 and did not alter significance.

E.2 Study 2

Day-level missingness 3.1%; inclusion threshold ≥ 10/14 valid days (met by 94.4%). Median-imputation sensitivity changed Δ pp by ≤ 0.6 and did not alter significance.