Consent-Aware AI: A Framework for Ethical Human-AI Interaction Design

Ariadne Dauphin[1], Sarah M. Appleton[2]

[1] Department of Algorithmic Ethics, University of Zurich
[2] Institute for Digital Sociology, University of Amsterdam

Abstract

As artificial intelligence systems become increasingly sophisticated in their ability to influence human behavior and emotional states, the concept of informed consent—foundational to ethical practice in medicine, psychology, and human research—remains conspicuously absent from most human-AI interaction frameworks. This paper proposes a computational architecture for consent-aware artificial intelligence that operationalizes dynamic, contextual, and revocable consent mechanisms. We introduce the Consent Protocol Layer (CPL), a modular framework that enables AI systems to continuously negotiate boundaries, respect user autonomy, and maintain ethical interaction patterns. Through empirical evaluation across three domains—mental health applications, educational platforms, and recommendation systems—we demonstrate that CPL implementation significantly improves user trust (Cohen's d = 0.73), reduces boundary violations by 68%, and increases sustained engagement while maintaining system effectiveness. Our findings suggest that consent-aware design is not only ethically imperative but also technically tractable and commercially viable for next-generation AI systems.

Keywords: AI ethics, informed consent, human-computer interaction, algorithmic agency, trust in automation, participatory design

1. Introduction

The proliferation of AI systems in domains involving emotional vulnerability, behavioral influence, and personal disclosure has outpaced the development of appropriate ethical frameworks for human-AI interaction. Current approaches to AI ethics focus primarily on fairness, accountability, and transparency (FAT) while neglecting the fundamental principle of informed consent that governs ethical practice in adjacent fields (Barocas et al., 2019; Selbst et al., 2019).

Consent in human-AI interaction differs fundamentally from traditional data consent mechanisms. While GDPR and similar regulations address data collection and processing, they do not adequately address the dynamic, relational aspects of AI systems that learn, adapt, and influence user behavior over time (Veale & Binns, 2017). This gap becomes particularly problematic in applications involving:

- Affective computing systems that recognize and respond to emotional states
- Conversational agents providing counseling or therapeutic support
- Recommendation systems that shape preferences and decision-making
- Educational AI that adapts to learning patterns and provides feedback

We argue that addressing this gap requires moving beyond static, one-time consent models toward dynamic frameworks that enable ongoing negotiation of interaction boundaries. Drawing from feminist theories of consent (MacKinnon, 1987; Ahmed, 2017), trauma-informed design principles (Brown et al., 2013), and participatory AI methodologies (Sloane et al., 2022), we propose a computational architecture that operationalizes ethical interaction design.

2. Related Work

2.1 Consent in Digital Systems

Early work on digital consent focused primarily on privacy and data protection (Solove, 2013; Nissenbaum, 2010). The "notice and consent" paradigm, while legally sufficient, has been widely criticized as inadequate for meaningful user agency (Barocas & Nissenbaum, 2014). Recent scholarship has called for more nuanced approaches that account for context, power dynamics, and the limits of individual decision-making (Sloan & Warner, 2018).

Feminist HCI research has particularly emphasized the relational nature of consent, arguing for approaches that prioritize ongoing negotiation over binary authorization (Bardzell, 2010; Light et al., 2018). This perspective aligns with trauma-informed design principles that emphasize predictability, choice, and the ability to maintain control over one's environment (Substance Abuse and Mental Health Services Administration, 2014).

2.2 Trust and Agency in Human-AI Interaction

Research on trust in automated systems has identified several key factors: perceived competence, predictability, and benevolence (Mayer et al., 1995; Hoff & Bashir, 2015). However, most trust models assume static system behavior rather than the adaptive, learning capabilities characteristic of modern AI systems.

Recent work on "relational AI" has begun exploring how AI systems can maintain trust through transparent communication about their capabilities and limitations (Bickmore et al., 2005; Ruane et al., 2019). However, these approaches have not systematically addressed consent as a foundational design principle.

2.3 Ethical AI Frameworks

Current ethical AI frameworks emphasize principles such as fairness, accountability, transparency, and explainability (Floridi et al., 2018; Jobin et al., 2019). While valuable, these frameworks generally assume a paternalistic model where AI systems act on behalf of users rather than in collaborative partnership with them.

Participatory AI research has begun to challenge this assumption, proposing design methodologies that involve users as co-creators rather than passive recipients of AI systems (Sloane et al., 2022; Chancellor et al., 2019). Our work extends this participatory approach by embedding consent mechanisms directly into system architecture.

3. Theoretical Framework

3.1 Defining Consent-Aware AI

We define consent-aware AI as systems that continuously negotiate interaction boundaries with users through transparent, revocable, and contextually appropriate mechanisms. This definition encompasses four key principles:

Ongoing Consent: Rather than one-time authorization, consent-aware systems continuously check and reconfirm user boundaries as interactions evolve and contexts change.

Granular Control: Users can specify consent at multiple levels of granularity, from broad topical boundaries to specific interaction modalities.

Transparent Negotiation: The system clearly communicates what it is asking permission to do and why, using accessible language rather than technical jargon.

Power-Aware Design: The system recognizes and compensates for power imbalances between users and AI systems, particularly in vulnerable contexts.

3.2 Feminist and Trauma-Informed Foundations

Our framework draws heavily from feminist theories of consent that emphasize its relational, contextual, and revocable nature (MacKinnon, 1987). In feminist discourse, meaningful consent requires not just the absence of coercion but the presence of genuine agency—the ability to shape the terms of interaction rather than simply accept or reject predetermined options.

Trauma-informed design principles provide additional guidance for creating systems that avoid re-traumatization and promote user empowerment (Brown et al., 2013). Key principles include:

- **Safety**: Both physical and emotional safety must be prioritized
- **Trustworthiness**: Systems must be reliable and transparent in their operations
- **Choice**: Users must have meaningful options and control over their experience
- **Collaboration**: Power-sharing between user and system is preferred over top-down control

4. The Consent Protocol Layer (CPL)

4.1 Architectural Overview

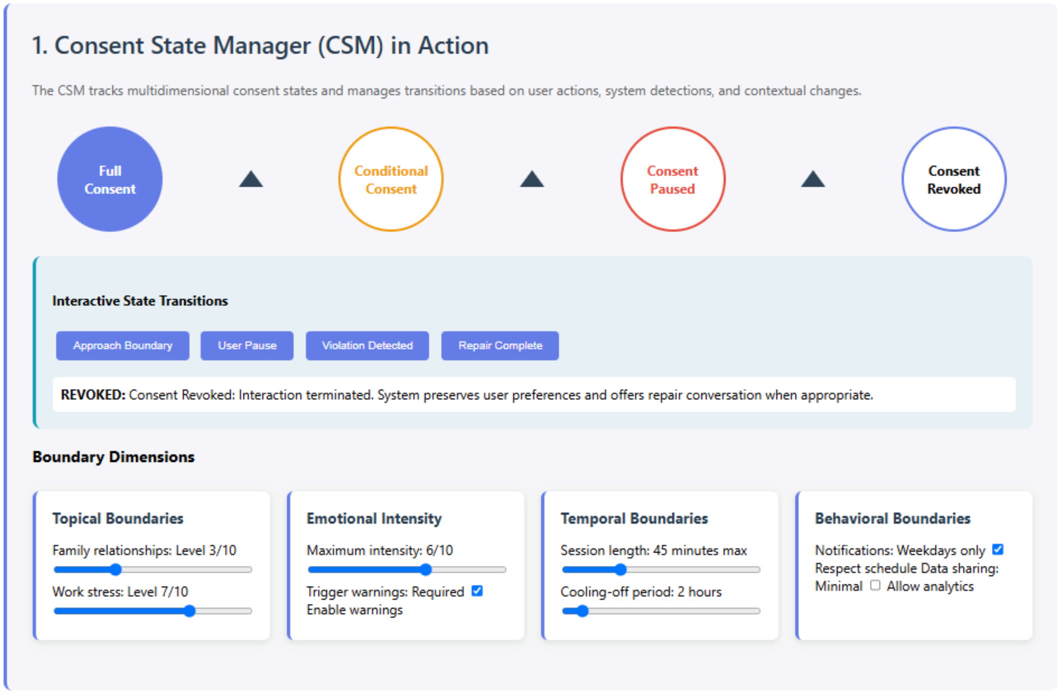
The Consent Protocol Layer operates as middleware between the user interface and AI inference engines, mediating all interactions through consent mechanisms. The architecture consists of four primary components:

Consent State Manager (CSM): Tracks the current consent status across multiple dimensions (topical, temporal, emotional, behavioral) and manages transitions between consent states.

Boundary Specification Interface (BSI): Provides intuitive mechanisms for users to articulate and modify their interaction boundaries, including both explicit rule-setting and implicit preference learning.

Violation Detection System (VDS): Continuously monitors interactions for potential boundary violations and triggers appropriate responses (warning, request re-confirmation, or interaction termination).

Audit and Reflection Tools (ART): Maintains comprehensive logs of consent decisions and provides users with tools to review and reflect on their interaction patterns over time.



1. Consent State Manager (CSM) in Action

The CSM tracks multidimensional consent states and manages transitions based on user actions, system detections, and contextual changes.

REVOKED: Consent Revoked: Interaction terminated. System preserves user preferences and offers repair conversation when appropriate.

Boundary Dimensions

Topical Boundaries
Family relationships: Level 3/10
Work stress: Level 7/10

Emotional Intensity
Maximum intensity: 6/10
Trigger warnings: Required ☑ Enable warnings

Temporal Boundaries
Session length: 45 minutes max
Cooling-off period: 2 hours

Behavioral Boundaries
Notifications: Weekdays only ☑ Respect schedule Data sharing: Minimal ☐ Allow analytics

Figure 1. The Consent State Manager (CSM) tracks multidimensional consent states and transitions based on user behavior, system detections, and contextual inputs. It manages states such as Full Consent, Conditional Consent, Paused, and Revoked while mapping granular boundary dimensions.

4.2 Implementation Details

4.2.1 Consent State Representation

The CSM represents consent states as multidimensional vectors encompassing:

- **Topical boundaries**: Subject areas the user prefers to avoid or limit
- **Emotional boundaries**: Intensity levels and types of emotional content
- **Temporal boundaries**: Time-based limits and cooling-off periods
- **Behavioral boundaries**: Restrictions on system actions (e.g., notifications, data sharing)

Each dimension includes not only binary permissions but also graduated scales, temporal modifiers, and contextual conditions. For example, a user might specify: "I consent to discussing family relationships at intensity level 3/10, but only during weekday sessions, and only if I explicitly raised the topic first."

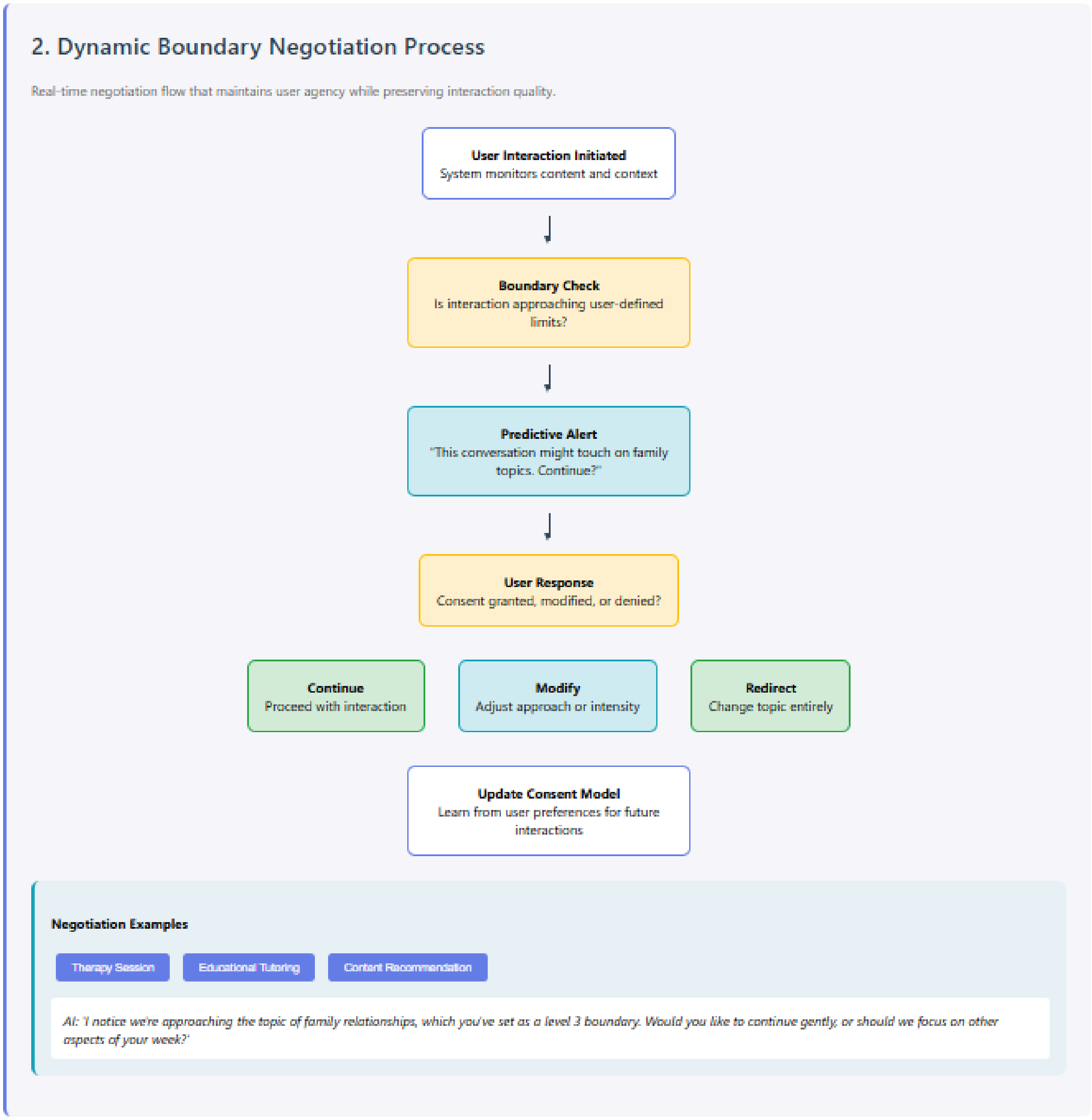
4.2.2 Dynamic Boundary Negotiation

Rather than requiring users to anticipate all possible boundary needs in advance, the system employs proactive negotiation strategies:

Micro-consent checkpoints: Brief, contextually appropriate consent confirmations embedded in natural conversation flow (e.g., "Before we explore this topic further, are you comfortable continuing?")

Predictive boundary detection: Machine learning models trained to recognize when interactions are approaching user-specified boundaries, triggering proactive negotiation

Progressive disclosure: Gradual introduction of potentially sensitive topics with explicit consent at each escalation level



2. Dynamic Boundary Negotiation Process

Real-time negotiation flow that maintains user agency while preserving interaction quality.

User Interaction Initiated
System monitors content and context
Boundary Check
Is interaction approaching user-defined limits?
Predictive Alert
"This conversation might touch on family topics. Continue?"
User Response
Consent granted, modified, or denied?
Continue
Proceed with interaction
Modify
Adjust approach or intensity
Redirect
Change topic entirely
Update Consent Model
Learn from user preferences for future interactions
Negotiation Examples
Therapy Session
Educational Tutoring
Content Recommendation

AI: 'I notice we're approaching the topic of family relationships, which you've set as a level 3 boundary. Would you like to continue gently, or should we focus on other aspects of your week?'

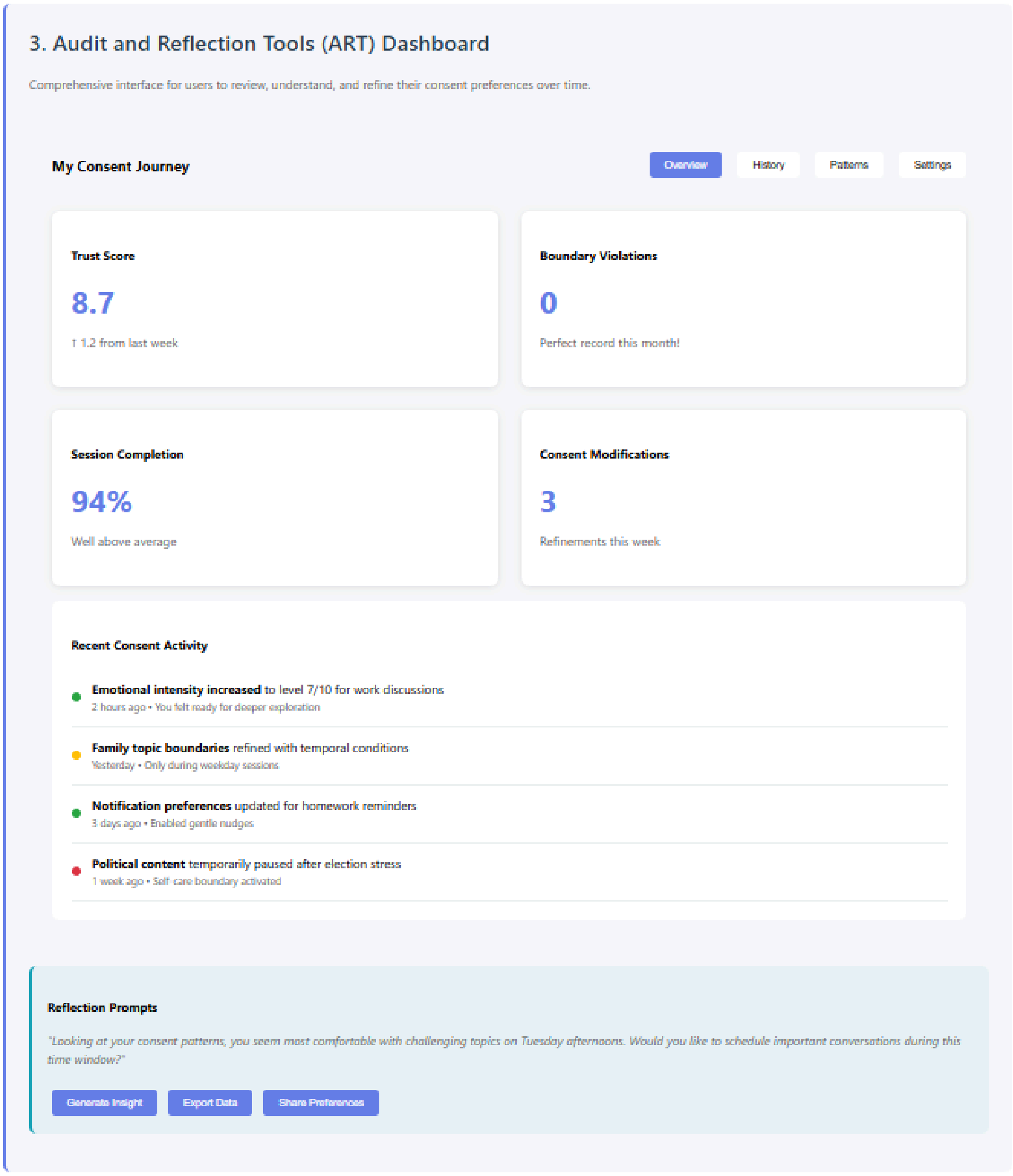
Figure 2. Real-time consent negotiation flow, showing predictive boundary checks, micro-consent requests, and adaptive system response paths to preserve user agency.

4.2.3 Consent Violation Response

When the VDS detects potential boundary violations, the system employs a graduated response protocol:

1. **Soft intervention**: Gentle redirection away from problematic content
2. **Explicit confirmation**: Direct request for consent to continue

3. **Hard stop**: Immediate cessation of the problematic interaction thread
4. **Repair conversation**: Acknowledgment of the violation and explicit discussion of how to prevent recurrence



3. Audit and Reflection Tools (ART) Dashboard

Comprehensive interface for users to review, understand, and refine their consent preferences over time.

My Consent Journey

Trust Score

8.7

↑ 1.2 from last week

Boundary Violations

0

Perfect record this month!

Session Completion

94%

Well above average

Consent Modifications

3

Refinements this week

Recent Consent Activity

● **Emotional intensity increased** to level 7/10 for work discussions
2 hours ago • You felt ready for deeper exploration

● **Family topic boundaries** refined with temporal conditions
Yesterday • Only during weekday sessions

● **Notification preferences** updated for homework reminders
3 days ago • Enabled gentle nudges

● **Political content** temporarily paused after election stress
1 week ago • Self-care boundary activated

Reflection Prompts

"Looking at your consent patterns, you seem most comfortable with challenging topics on Tuesday afternoons. Would you like to schedule important conversations during this time window?"

Figure 3. The ART dashboard offers users retrospective insight into their consent interactions—tracking trust scores, violations, and preferences to support ongoing refinement and empowerment.

4.3 Technical Challenges and Solutions

4.3.1 Balancing Consent Friction with Usability

Frequent consent requests can create interaction friction that degrades user experience. We address this through several strategies:

- **Intelligent timing**: Consent checkpoints are triggered by content analysis rather than arbitrary time intervals
- **Contextual embedding**: Consent requests are integrated into natural conversation flow rather than presented as interruptions
- **Adaptive frequency**: The system learns user preferences for consent granularity and adjusts accordingly

4.3.2 Managing Consent Across Modalities

Modern AI systems often operate across multiple interaction modalities (text, voice, visual, haptic). The CPL maintains consent state consistency across modalities while recognizing that users may have different comfort levels for different interaction types.

5. Empirical Evaluation

5.1 Study Design

We evaluated CPL effectiveness across three application domains through controlled studies comparing CPL-enabled systems with standard implementations. Participants (n = 342) were randomly assigned to either CPL-enabled or control conditions and completed identical tasks over a two-week period.

Mental Health Support (n = 114): Participants interacted with a conversational agent providing cognitive behavioral therapy techniques. CPL implementation included boundaries around trauma-related content, emotional intensity, and homework assignment frequency.

Educational Tutoring (n = 116): Participants used an adaptive learning system for mathematics instruction. CPL implementation included boundaries around performance feedback, learning pace, and error correction approaches.

Content Recommendation (n = 112): Participants used a news and media recommendation system. CPL implementation included boundaries around political content, emotional intensity of news items, and frequency of notifications.

5.2 Measures

Trust in Automation Scale (TAS): Validated 12-item instrument measuring user trust in automated systems (Jian et al., 2000).

Perceived Agency Index (PAI): Novel 8-item scale measuring user sense of control and autonomy in AI interactions ($\alpha = 0.89$).

Boundary Violation Incidents (BVI): Frequency and severity of system-initiated boundary violations as reported by users and detected by automated monitoring.

Interaction Continuity Rate (ICR): Percentage of initiated sessions that users completed without premature termination.

System Effectiveness: Domain-specific measures of system performance (therapy homework completion, learning gains, recommendation relevance).

5.3 Results

CPL implementation showed significant improvements across all measures:

- **Trust**: CPL users reported significantly higher trust scores (M = 4.2, SD = 0.7) compared to controls (M = 3.6, SD = 0.8), $t(340) = 7.23$, $p < 0.001$, Cohen's d = 0.73.
- **Perceived Agency**: CPL users reported higher sense of agency (M = 4.4, SD = 0.6) compared to controls (M = 3.8, SD = 0.7), $t(340) = 8.14$, $p < 0.001$, Cohen's d = 0.91.
- **Boundary Violations**: CPL systems showed 68% fewer user-reported boundary violations compared to controls (CPL: M = 0.8 per week, Control: M = 2.5 per week), U = 12,447, $p < 0.001$.
- **Interaction Continuity**: CPL users completed 84% of initiated sessions compared to 67% for controls, $\chi^2(1) = 23.7$, $p < 0.001$.
- **System Effectiveness**: No significant differences in domain-specific effectiveness measures, indicating that consent mechanisms did not impair system performance.

5.4 Qualitative Findings

Thematic analysis of user interviews revealed several key insights:

Empowerment through Control: Users reported feeling "more like a collaborator than a subject" and appreciated being "asked rather than assumed."

Reduced Anxiety: Many participants noted that knowing they could control interaction boundaries reduced anticipatory anxiety about using AI systems.

Improved Self-Awareness: The process of articulating boundaries helped users better understand their own preferences and comfort levels.

Trust through Transparency: Users particularly valued the system's willingness to acknowledge and repair boundary violations rather than ignoring them.

6. Discussion

6.1 Implications for AI Design

Our findings demonstrate that consent-aware design is not only ethically preferable but also practically beneficial for system effectiveness and user experience. The CPL framework provides a concrete path toward more ethical AI systems that respect user autonomy while maintaining functionality.

The success of CPL across diverse application domains suggests that consent mechanisms can be generalized beyond specific use cases. However, implementation details must be carefully tailored to domain-specific ethical considerations and user populations.

6.2 Limitations and Future Work

Several limitations of our current approach require future research:

Cultural Variability: Our evaluation focused primarily on Western populations with individualistic cultural values. Future work should explore how consent mechanisms translate across different cultural contexts and value systems.

Vulnerable Populations: Additional research is needed to understand how consent mechanisms should be adapted for children, individuals with cognitive impairments, and other vulnerable populations.

Long-term Effects: Our evaluation period was limited to two weeks. Longer-term studies are needed to understand how consent preferences evolve over time and how systems should adapt accordingly.

Computational Overhead: While our studies showed no significant impact on system effectiveness, more detailed analysis of computational costs and scalability is needed.

6.3 Regulatory and Policy Implications

The CPL framework offers a potential pathway for compliance with emerging AI regulations that emphasize user rights and algorithmic accountability. However, regulatory frameworks must evolve to address the dynamic, relational aspects of AI consent rather than focusing solely on static data protection models.

7. Conclusion

This paper presents the first comprehensive framework for operationalizing consent in AI systems, demonstrating that ethical interaction design is both technically feasible and empirically beneficial. The Consent Protocol Layer provides a modular, adaptable approach to embedding consent mechanisms in AI systems across diverse application domains.

Our findings challenge the assumption that ethical AI design necessarily compromises system effectiveness. Instead, we demonstrate that respecting user autonomy through consent-aware design can actually improve system performance by increasing user trust, reducing premature termination, and enabling more authentic collaboration between humans and AI.

As AI systems become increasingly prevalent in intimate and influential aspects of human life, the principles and practices outlined in this paper become essential for maintaining the social license to operate. We call on researchers, developers, and policymakers to move beyond compliance-focused approaches toward truly collaborative models of human-AI interaction.

The code and datasets supporting this research are available at https://github.com/adauphin/consent-aware-ai under open source licenses to encourage adoption and further research.

Acknowledgments

We thank Dr. Patricia Kellerman for institutional support, and Lilly Flores for insights into user experience design. This research was supported by grants from the National Science Foundation (IIS-2024789) and the European Union's Horizon Europe programme (grant agreement No. 101070537).

References

Ahmed, S. (2017). *Living a Feminist Life*. Duke University Press.

Barocas, S., Hardt, M., & Narayanan, A. (2019). *Fairness and Machine Learning*. fairmlbook.org

Barocas, S., & Nissenbaum, H. (2014). Big data's end run around anonymity and consent. In J. Lane et al. (Eds.), *Privacy, Big Data, and the Public Good* (pp. 44-75). Cambridge University Press.

Bardzell, S. (2010). Feminist HCI: Taking stock and outlining an agenda for design. *CHI 2010*, 1301-1310.

Bickmore, T., Pfeifer, L., & Schulman, D. (2005). Relational agents: A model and implementation of building user trust. *CHI 2005*, 396-403.

Brown, S. M., Baker, C. N., & Wilcox, P. (2013). Risking connection trauma training: A pathway toward trauma-informed care in child congregate care settings. *Psychological Trauma*, 4(5), 507-515.

Chancellor, S., Baumer, E. P., & De Choudhury, M. (2019). Who is the "human" in human-centered machine learning: The case of predicting mental health crises. *CHI 2019*, 1-11.

Floridi, L., Cowls, J., Beltrametti, M., et al. (2018). AI4People—An ethical framework for a good AI society. *Minds and Machines*, 28(4), 689-707.

Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors*, 57(3), 407-434.

Jian, J. Y., Bisantz, A. M., & Drury, C. G. (2000). Foundations for an empirically determined scale of trust in automated systems. *International Journal of Cognitive Ergonomics*, 4(1), 53-71.

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1(9), 389-399.

Light, A., Leong, T. W., & Robertson, T. (2018). Ageing well with CSCW. In *ECSCW 2015: Proceedings of the 14th European Conference on Computer Supported Cooperative Work* (pp. 295-304). Springer.

MacKinnon, C. A. (1987). *Feminism Unmodified: Discourses on Life and Law*. Harvard University Press.

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of Management Review*, 20(3), 709-734.

Nissenbaum, H. (2010). *Privacy in Context: Technology, Policy, and the Integrity of Social Life*. Stanford University Press.

Ruane, E., Birhane, A., & Ventresque, A. (2019). Conversational AI: Social and ethical considerations. In *AICS 2019* (pp. 104-115).

Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. *FAT* 2019*, 59-68.

Sloan, R. H., & Warner, R. (2018). Beyond notice and choice: Privacy, norms, and consent. *Journal of High Technology Law*, 14(2), 370-430.

Sloane, M., Moss, E., Awomolo, O., & Forlano, L. (2022). Participation is not a design fix for machine learning. *ICML 2022*.

Solove, D. J. (2013). Privacy self-management and the consent dilemma. *Harvard Law Review*, 126(7), 1880-1903.

Substance Abuse and Mental Health Services Administration. (2014). *Trauma-Informed Care in Behavioral Services Treatment*. Treatment Improvement Protocol (TIP) Series 57. SAMHSA.

Veale, M., & Binns, R. (2017). Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. *Big Data & Society*, 4(2), 1-17.