Human-Computer Interaction & Emerging Technologies, Vol. 233, 2026, 125–135
https://doi.org/10.54941/ahfe1007512

An Adversarial Dual-Agent Critical Framework for Intelligent Evaluation and Optimization of Human–Computer Interaction Design

Jiaqi Han[1] and Peiyan Zhong[2]

[1]Qilu University of Technology (Shandong Academy of Sciences), China
[2]Chongqing University, China

ABSTRACT

Critical thinking is a core practice in the field of human-computer interaction and design, aiming to enhance the quality of interaction solutions in multiple dimensions such as experience, technology, and ethics through systematic review. Existing research lacks collaborative deduction and in-depth demonstration of HCI schemes at the behavioral logic, engineering implementation and comprehensive risk levels. This leads to the difficulty in improving love, with feedback remaining superficial and fragmented, making it hard to support high-quality innovation and decision-making in a complex and dynamic interactive context. This paper proposes an adversarial design critical framework based on dual agents. This framework consists of two adversarial agents: (1) Experience Optimization Agent: This agent takes user experience as the core evaluation dimension and quantitatively analyzes the intuitiveness, operational efficiency, and user emotional feedback of the design plan based on interaction design principles, cognitive psychology models, and input user expectations. (2) Constraint Verification Agent: This agent takes technical feasibility and actual conditions as the evaluation dimensions, and based on implementation cost, performance indicators, multi-terminal adaptation requirements and basic design specifications, identifies technical implementation risks, performance defects and compliance issues existing in the design scheme. The results of user experiments show that, compared with traditional schemes, the framework constructed in this paper demonstrates outstanding performance in human-computer interaction tasks, significantly enhancing the comprehensive adoption rate of the generated suggestions, and showing efficient early warning capabilities for key design issues such as logical contradictions and imbalance of resource benefits. The collaborative evaluation mechanism effectively reduces the cognitive load generated when dealing with multi-dimensional feedback, enabling designers to focus more on core design decisions. In complex design contexts such as internationalization and cross-cultural, this framework also demonstrates superior adaptability and strategy generation potential, which is conducive to promoting the evolution of design assistance tools towards an intelligent collaboration paradigm with higher-level cognitive support capabilities.

Keywords: Critical thinking, Human-computer interaction, Design, Dual agents

Received February 20, 2026; Revised March 28, 2026; Accepted April 14, 2026; Available online July 20, 2026

INTRODUCTION

Critical thinking stands as a cornerstone practice within the field of Human-Computer Interaction (HCI) and design. It is essential for systematically deconstructing and examining interaction solutions to enhance their quality across multiple critical dimensions, including user experience, technical robustness, and ethical consideration. This rigorous evaluative process is fundamental for transforming initial concepts into refined, effective, and responsible interactive systems. As digital products and services evolve within increasingly complex and dynamic contexts, the role of structured critical analysis in supporting high-quality innovation and strategic decision-making becomes ever more paramount.

Despite its acknowledged importance, the practical application of critical thinking in contemporary design evaluation often remains suboptimal. Existing approaches and research frequently yield feedback that is fragmented and superficial, lacking a mechanism for collaborative deduction and in-depth demonstration of HCI schemes. Specifically, there is a notable gap in the synergistic and thorough examination of proposals at the intertwined levels of behavioral logic, engineering implementation, and comprehensive risk assessment. This insufficiency leads to significant difficulties in iteratively improving design solutions. Designers are often confronted with disjointed feedback that elevates cognitive load, making it challenging to synthesize actionable insights. Consequently, this hampers the ability to foster meaningful innovation and make informed decisions, particularly in complex, dynamic, and nuanced design contexts such as internationalization and cross-cultural adaptation.

To address these limitations, this paper proposes an Adversarial Design Critical Framework (ADCF) based on a dual-agent architecture. This framework institutionalizes a structured, collaborative critique by employing two specialized adversarial agents:(1) The Experience Optimization Agent (EOA): This agent focuses on user experience as its core evaluation dimension. It performs quantitative analysis of a design's intuitiveness, operational efficiency, and anticipated user emotional feedback, grounding its assessment in established interaction design principles, cognitive psychology models, and specified user expectations. (2) The Constraint Verification Agent (CVA): This agent prioritizes technical feasibility and real-world constraints. It scrutinizes a design sheme for technical implementation risks, performance bottlenecks, multi-platform adaptation challenges, and compliance with fundamental design specifications, based on factors such as development cost, performance metrics, and regulatory requirements.

By engaging in this adversarial yet collaborative evaluation, the framework aims to generate a more balanced, comprehensive, and actionable critique. It seeks to reduce the cognitive burden associated with processing multi-dimensional feedback, allowing designers to concentrate on core creative and strategic decisions. The primary contributions of this work are threefold. First, we introduce a novel adversarial dual-agent framework that formally balances experiential goals with practical constraints, promoting a more holistic design evaluation. Second, the framework provides quantifiable analysis models for key experiential attributes (e.g., intuitiveness, efficiency)

and critical implementation concerns (e.g., technical risk, compliance), moving critique beyond subjective opinion. Third, through user experiments, we empirically validate the framework's effectiveness. Results demonstrate its superior performance in enhancing the comprehensive adoption rate of generated design suggestions and its efficient early-warning capability for critical issues like logical contradictions and resource-benefit imbalances. The framework also shows promising adaptability and strategic potential in complex design scenarios.

RELATED WORK

Critical Thinking in HCI and Design Evaluation

Critical and reflective perspectives in HCI argue that evaluation should go beyond surface usability defects to interrogate assumptions, values, and socio-technical consequences embedded in interactive artifacts. Reflective design, for instance, frames design work as a practice of making tacit values explicit; Sengers et al. introduced reflective design as a strategy to "identify unconscious values and assumptions built into the way we conceive of design problems," thereby foregrounding hidden commitments in interaction solutions (Sengers et al., 2005). In a related but more provocative tradition, critical design has been positioned as a way to contest dominant narratives of "good" technology by using artifacts to question norms and expose tensions in everyday life; Bardzell and Bardzell articulate what makes such work "critical" in HCI and how it can function as a reflective and ethical lens rather than merely an aesthetic stance (Bardzell and Bardzell, 2013). Research-through-design further strengthens this orientation by emphasizing the role of design artifacts as knowledge-producing instruments and by making evaluation inseparable from iterative reflection and argumentation about trade-offs (Zimmerman et al., 2007). Complementary methods such as cultural probes similarly highlight that in complex human contexts, uncertainty and interpretation are not simply noise but can be productive signals for understanding lived experience and values (Gaver et al., 2004). Collectively, these lines of work suggest that robust evaluation should support multi-perspective reasoning and explainable critique, rather than only enumerating isolated issues.

In contrast, classic inspection-based evaluation methods remain foundational but often yield feedback that is granular yet difficult to synthesize into design-level insights. Heuristic evaluation, popularized in HCI as a fast expert inspection method, asks evaluators to judge interfaces against a set of usability principles (heuristics) and produces lists of violations and recommendations (Nielsen and Molich, 1990). While efficient, its outputs can be inherently fragmented: evaluators identify different subsets of issues, and aggregation becomes a non-trivial interpretive step. Nielsen additionally proposed strengthening heuristics by improving their explanatory power, underscoring that "naming" a violation is insufficient unless evaluators can connect observations to underlying user difficulties and design rationales (Nielsen, 1994). Cognitive walkthroughs approach evaluation through task-based reasoning about learnability and action selection, providing a

theory-motivated lens on whether users can form correct goals and take correct actions step-by-step (Polson et al., 1992). However, both heuristics and walkthroughs typically result in issue-centered feedback, leaving designers to integrate competing comments and resolve trade-offs without an explicit mechanism for structured debate or collaborative deduction.

Empirical work on usability practice further suggests that analysis and synthesis remain pain points in real-world evaluation workflows. For example, survey evidence indicates that analysis in practical usability evaluation is often constrained by time, method limitations, and difficulties translating findings into prioritized redesign action, especially when feedback sources are heterogeneous (Følstad et al., 2012). These challenges become more pronounced in complex design contexts (e.g., cross-cultural interaction, multi-device ecosystems), where designers must integrate usability, experience, ethics, and feasibility constraints in a single decision space. This gap, between critical/reflective aspirations and the fragmented nature of conventional evaluation outputs, motivates frameworks that can systematically organize critique, support multi-dimensional reasoning, and reduce the cognitive load of reconciling conflicting guidance.

Automated and Agent-Based Design Assistance Tools

To address the cost and scalability limitations of manual evaluation, HCI has long explored automation for usability assessment and design support. A landmark survey by Ivory and Hearst systematically reviewed automated usability evaluation approaches and categorized techniques ranging from guideline-based checking to model-based and empirical-data-driven methods, highlighting both promise and persistent challenges such as validity, generalization, and the gap between detected issues and actionable design improvements (Ivory and Hearst, 2001). Building on such foundations, broader work in UI software tooling emphasizes that interactive system construction and evaluation can be augmented by intelligent support and reusable infrastructure, yet practical impact hinges on how well tools align with real design workflows and how interpretable their outputs are to designers (Myers et al., 2000).

A key limitation in many automated tools is that they often behave like "single-lens" evaluators: they are strong at detecting a particular type of defect (e.g., guideline violations, layout constraints, performance thresholds) but weak at balancing competing objectives (e.g., delight vs. feasibility, novelty vs. compliance) and at explaining trade-offs in a way that designers can trust. Mixed-initiative interaction offers a useful conceptual bridge here: instead of full automation, systems collaborate with humans by taking initiative when helpful and deferring when uncertainty or value judgments dominate; Horvitz articulated principles for mixed-initiative user interfaces that operationalize such collaboration and decision-making under uncertainty (Horvitz, 1999). While mixed-initiative systems can reduce designer effort, they still frequently lack a formalized mechanism for structured critique that pits competing perspectives against each other to stress-test design decisions.

In design practice, critique is not merely about detection but about argumentation—why something is problematic, what alternative satisfies

constraints, and how to prioritize. A recent systematic review of the Design Critique method consolidates how critique is used across software and interaction design, and it underscores that critique quality depends on framing, grounding evidence, and making the rationale actionable (Alabood et al., 2023). This observation aligns with a broader gap in current intelligent design assistants: even when tools can generate recommendations, they rarely provide an internal structure that supports adversarial reasoning (e.g., one agent advocating experience improvements while another challenges feasibility/compliance), nor do they explicitly negotiate conflicts in a traceable way.

Against this backdrop, critical/speculative design perspectives further reinforce why "optimization" alone is insufficient—some design goals are inherently value-laden and benefit from structured contestation. Speculative design argues for exploring alternative futures and interrogating assumptions through artifacts; Dunne and Raby's work is frequently cited in HCI as a foundation for using design to question what ought to be built, not just what can be built (Dunne and Raby, 2013). Therefore, an agent-based assistant that only "improves usability" without confronting value conflicts or feasibility constraints may systematically under-serve real-world design decision-making. These limitations collectively motivate the need for frameworks that combine automated evaluation capacity, mixed-initiative collaboration, and critique-centered reasoning that can explicitly represent and reconcile conflict, precisely the gap addressed by adversarial dual-agent approaches.

METHOD: ADVERSARIAL DUAL-AGENT FRAMEWORK

Framework Overview

This section presents the overall architecture of our proposed Adversarial Dual-Agent Critical Framework (ADCF). The framework comprises two parallel-working agents: the Experience Optimization Agent (EOA) and the Constraint Verification Agent (CVA). Each agent evaluates a design proposal from a different primary focus—the EOA emphasizes the optimization and enhancement of user experience, while the CVA focuses on inspecting constraints such as technical feasibility and compliance. Each agent independently generates its evaluation results and recommendation report. When their conclusions conflict, the framework initiates an adversarial negotiation mechanism to resolve discrepancies through iterative interactions between the agents. During this adversarial negotiation process, the EOA and CVA challenge each other's suggestions and adjust their own recommendations accordingly, ultimately producing a consolidated evaluation report that achieves a balanced trade-off between elevating user experience and satisfying practical constraints. As illustrated in the framework flowchart, the entire process involves user requirement input, parallel processing by the dual agents, the adversarial negotiation phase, and the final generation of design suggestions.

Experience Optimization Agent (EOA)

The EOA specializes in assessing and improving the user experience of a design proposal. This agent analyzes and quantifies the design across several user-experience-related dimensions, deriving an overall experience score through a comprehensive scoring function, and subsequently provides corresponding optimization suggestions. The evaluation dimensions and the quantitative model of the EOA are detailed below.

Evaluation Dimensions

The EOA considers three key dimensions for user experience evaluation. The first is Intuitiveness, which measures how easily users can understand the interface and accomplish tasks. It can be quantified using metrics such as task completion time, error rate, and learning curve slope; more favorable metrics result in a higher intuitiveness score. The second dimension is Operational Efficiency, reflecting how efficiently users can perform tasks. This is evaluated based on factors like the number of steps, click count, and total time required to complete a goal; simpler steps and shorter completion times yield a higher operational efficiency score. The final dimension is User Emotional Feedback, which assesses users' subjective emotions and satisfaction during interaction. Standard emotion models (e.g., the Pleasure-Arousal-Dominance model) or direct user satisfaction ratings can be employed to quantify emotional experience; a higher degree of positive emotion corresponds to a higher score in this dimension.

Quantitative Model

Within the EOA's quantitative model, we calculate an overall user experience rating by synthesizing scores from all dimensions. Specifically, an experience scoring function (see Formula 1) is defined. This function combines the scores for Intuitiveness (I), Operational Efficiency (E), and Emotional Feedback (F) via a weighted sum using corresponding weight coefficients, *w*. These weight coefficients can be adjusted according to specific application contexts to reflect the relative importance of each dimension in the overall experience evaluation. By tuning the values of *w*, the framework can dynamically emphasize certain user experience elements under different design requirements while ensuring all dimensions are considered.

Furthermore, we specify the calculation method for each sub-dimension score. For instance, the Intuitiveness score I is computed based on a normalized combination of task completion time and error rate: shorter completion times and lower error rates, after normalization, yield a higher I value. Similarly, the Emotional Feedback score F is calculated based on the alignment between the design outcome and user expectations—the better the design meets or exceeds user expectations for experience, the higher the F score. The Operational Efficiency score E is directly determined from relevant efficiency metrics (e.g., calculated based on the number of task steps and average completion time, where fewer steps or shorter average times result in a higher E), quantifying the convenience of the design for task execution. After obtaining the three scores I, E, and F, the EOA can quantitatively assess

the overall user experience level of the design proposal and provide targeted optimization suggestions accordingly.

Constraint Verification Agent (CVA)

Next, we introduce the Constraint Verification Agent (CVA). The CVA is responsible for evaluating the design proposal from the perspective of real-world constraints, focusing on identifying risks in technical implementation, cross-platform adaptation issues, and defects in specification compliance. Through analysis of these aspects, the CVA can pinpoint factors that may hinder the implementation of the design proposal and propose corresponding refinement requirements. The main evaluation dimensions considered by the CVA and its quantitative verification model are explained below.

Evaluation Dimensions

The CVA considers three primary constraint evaluation dimensions. The first is Technical Feasibility, which examines the viability of technically implementing the design, including factors such as development cost and whether performance metrics (e.g., response time, memory usage) meet requirements. Lower implementation costs and easier-to-achieve performance targets result in a higher technical feasibility score. The second is Multi-platform Adaptability, measuring the compatibility and consistency of the design across different devices or platforms. This involves checking the adaptation of interface layouts and interaction consistency on various terminals, ensuring the design maintains a good user experience across diverse environments like web and mobile. The third dimension is Compliance, assessing the design's adherence to relevant standards and guidelines. Specifically, it involves verifying whether the design conforms to industry standards and platform-specific guidelines, such as accessibility standards (WCAG) and official Human Interface Guidelines for specific platforms (e.g., iOS/Android). Violations of these guidelines will lower the compliance score.

Verification Model

Within the CVA's verification model, we synthesize the scores from the above dimensions into an overall constraint satisfaction metric. For this purpose, a constraint satisfaction function is defined (see Formula 2) to integrate the scores for Technical Feasibility, Adaptability, and Compliance, quantitatively measuring the overall degree to which the design proposal satisfies various constraints. The sub-item scores are derived from the quantitative assessment of their respective dimensions. For example, the Technical Feasibility score (Tech) is calculated in a weighted manner based on implementation cost and performance: lower required implementation costs and fewer performance defects yield a higher Tech score.

The Compliance score (Comp) depends on the number and severity of design guideline violations—more violations result in a lower Comp score (if there are no violations, it can be considered a full score for compliance).

The impact of Multi-platform Adaptability is reflected by penalizing cross-platform inconsistency issues in the overall score: if the design has severe adaptation problems on certain terminals, the overall constraint satisfaction rating will be correspondingly reduced. By integrating various factors in the manner described, the CVA can quantitatively evaluate how well the design proposal meets different types of real-world constraints and, based on this, identify high-risk issues that require focused attention and improvement.

Adversarial Negotiation Mechanism

When the evaluation results of the EOA and CVA conflict (e.g., the EOA suggests introducing complex interactions to enhance user experience, while the CVA deems the design technically infeasible or non-compliant), the framework activates an adversarial negotiation mechanism to resolve the disagreement. This mechanism seeks a compromise through iterative interaction between the two agents. For this purpose, we introduce a trade-off function (see Formula 3) used during the negotiation process to balance considerations of user experience and constraint satisfaction. This trade-off function incorporates a context adjustment coefficient, α, to dynamically adjust the weighting of the two factors based on the design phase: in the early stages of conceptual design, a higher α value can be set to emphasize user experience (favoring the EOA's opinion); in later stages of detailed design and implementation, α is lowered to increase consideration for implementation constraints (favoring the CVA's opinion). Through the dynamic adjustment of this parameter, the framework can flexibly trade off between user experience optimization and implementation feasibility according to the project's progress stage.

During the adversarial negotiation process, guided by the trade-off function and related rules, the two agents continuously exchange views and adjust their respective suggestions until a mutually acceptable improvement plan is reached. Upon completion of the negotiation, the framework outputs a consolidated list of design suggestions that incorporates perspectives from both sides. Each suggestion is accompanied by a priority indication and annotated with the specific manner in which the conflict was resolved. This explanatory, annotated output enables designers to clearly understand the rationale behind each recommendation and ensures that critical issues involving conflicts between user experience and real-world constraints have been adequately weighed and resolved in the final proposal.

EVALUATION

Experimental Setup

A mixed-methods evaluation combining real-world user experiments and automated scenario testing was designed to comprehensively assess ADCF. Participants: 32 interaction designers and UX researchers with varying expertise (junior, intermediate, senior) were recruited and randomly assigned to an experimental group (using ADCF-assisted tools) and a control group (using traditional heuristic checklists and simulated expert review documents). Design Tasks: Three real-world HCI design challenges of differing complexity were selected:

Task A (Medium Complexity): Redesign the device control panel for a smart home mobile app to optimize operational efficiency and intuitiveness.

Task B (High Complexity): Design a cross-platform (Web, Mobile, Smartwatch) product browsing and purchasing flow for an e-commerce platform, balancing experience consistency with platform-specific constraints.

Task C (Internationalization): Design a personalized news feed interface for a reading app supporting multiple languages and considering cultural factors (e.g., reading direction, color preferences).

Procedure: Participants completed tasks individually. The experimental group used a prototype tool integrating ADCF, which provided real-time evaluation reports and suggestions synthesized from the EOA and CVA. The control group referenced a comprehensive but discrete heuristic checklist and a simulated expert review document. Final design solutions were collected for subsequent quantitative and qualitative analysis.

Simulation Experiment

To quantitatively test the framework's capability in detecting issues and generating strategies across a large, diverse set of scenarios, a test suite containing 120 predefined design defects was constructed. These defects covered logical contradictions, technical infeasibility, compliance violations, and cross-cultural adaptation issues.

Baseline Methods

ADCF was compared against two typical traditional design evaluation paradigms:

Heuristic Checklist (HC): A composite checklist based on Nielsen's heuristics and platform-specific design guidelines (e.g., Material Design, iOS HIG).

Asynchronous Expert Review (AER): Simulated independent review documents from two senior designers (one focused on experience, one on technology), to be integrated by the designer.

Evaluation Metrics & Key Results

A multi-dimensional assessment was conducted using the following metrics:

Suggestion Adoption & Design Improvement: The ADCF group demonstrated a significantly higher Suggestion Adoption Rate (SAR: 78% vs. 52% in HC, 61% in AER). The Scheme Improvement Rate (SIR) assessed by blind experts was also markedly higher for ADCF-generated revisions. Critical Issue Detection & Early Warning: In simulated testing, ADCF achieved a high F1-Score (0.89) in identifying predefined defects, particularly excelling in detecting logical inconsistencies. In user tasks, it provided earlier warnings for critical technical feasibility issues compared to baselines. Design Efficiency & Cognitive Load: While initial task completion times were comparable, the ADCF group reported significantly lower NASA-TLX scores (avg. 42 vs. 58 in control) and lower feedback integration difficulty, indicating reduced cognitive load. Cross-cultural Scenario Adaptability: For Task C, ADCF generated a greater number of context-specific adaptation

strategies, which were rated higher in relevance and feasibility (avg. 4.2/5) by experts compared to generic checklist prompts.

Conclusion

This paper addresses the critical issue of fragmented feedback and the lack of systematic, collaborative deduction in traditional HCI design evaluation by proposing an innovative Adversarial Dual-Agent Critical Framework (ADCF). By formalizing the critical thinking process into two adversarial yet collaborative agents, the Experience Optimization Agent and the Constraint Verification Agent, the research establishes a dynamic assessment system capable of concurrent deep experiential insight and rigorous technical-constraint verification. Results from user experiments and simulated testing consistently show that the framework not only significantly enhances the comprehensive quality and adoption rate of design suggestions but also provides effective early warning for underlying design risks such as logical contradictions and resource-benefit imbalances. More importantly, its integrated, explanatory feedback effectively reduces designers' cognitive load when processing multi-dimensional information, leading to more focused and efficient design decisions. Its promising performance in complex contexts like cross-cultural design further underscores the framework's potential in advancing design assistance tools from automated checking towards an intelligent collaboration paradigm with higher-order cognitive support capabilities. Future work will focus on expanding the agents' knowledge bases and exploring their application in collaborative, multi-stakeholder design environments.

REFERENCES

A. Dunne and F. Raby, Speculative Everything: Design, Fiction, and Social Dreaming. MIT Press, 2013.

A. Følstad, E. L.-C. Law, and K. Hornbæk, "Analysis in practical usability evaluation: A survey study," Proc. CHI 2012, 2012.

B. A. Myers, S. E. Hudson, and R. Pausch, "Past, Present, and Future of User Interface Software Tools," ACM Transactions on Computer-Human Interaction (TOCHI), 2000.

E. Horvitz, "Principles of Mixed-Initiative User Interfaces," Proc. CHI 1999, 1999.

J. Bardzell and S. Bardzell, "What is 'critical' about critical design?" Proc. CHI 2013, 2013.

J. Nielsen, "Enhancing the explanatory power of usability heuristics," Proc. CHI 1994, 1994.

J. Nielsen and R. Molich, "Heuristic evaluation of user interfaces," Proc. CHI 1990, 1990.

J. Zimmerman, J. Forlizzi, and S. Evenson, "Research Through Design as a Method for Interaction Design Research in HCI," Proc. CHI 2007, 2007.

L. Alabood, Z. Aminolroaya, D. Yim, O. Addam, and F. Maurer, "A systematic literature review of the Design Critique method," Information and Software Technology, 2023.

M. Y. Ivory and M. A. Hearst, "The state of the art in automating usability evaluation of user interfaces," ACM Computing Surveys, 2001.

P. G. Polson, C. Lewis, J. Rieman, and C. Wharton, "Cognitive walkthroughs: a method for theory-based evaluation of user interfaces," International Journal of Man-Machine Studies, 1992. DOI: https://doi.org/10.1016/0020-7373(92)90039-N

P. Sengers, K. Boehner, S. David, and J. "Jofish" Kaye, "Reflective Design," Proc. Critical Computing '05, 2005.

W. Gaver, A. Boucher, S. Pennington, and B. Walker, "Cultural probes and the value of uncertainty," interactions, vol. 11, no. 5, pp. 53–56, 2004.