Appropriate Automation in AI-Assisted Usability Heuristic Generation: Comparing Human-AI Collaborative Workflows

Yiyao Li
Human Factors Research Group
University of Nottingham
University Park
Nottingham, NG7 2RD
United Kingdom
yiyao.li@nottingham.ac.uk

Setia Hermawati
Human Factors Research Group
University of Nottingham
University Park
Nottingham, NG7 2RD
United Kingdom
setia.hermawati@nottingham.ac.uk

Horia Maior
School of Computer Science
University of Nottingham
Jubilee Campus
Nottingham, NG8 1BB
United Kingdom
horia.maior@nottingham.ac.uk

Domain-specific usability heuristics can better capture context-sensitive usability issues than general heuristics, but their development and refinement remain resource-intensive. This study examined UX practitioners' experiences of three human-AI workflows for generating and refining heuristics for automotive smart cockpit interfaces. Thirty user experience (UX) practitioners used an artificial intelligence (AI) tool to generate an initial heuristic set and were assigned to one of three refinement workflows: continuous AI-assisted feedback and refinement (G1), human expert feedback with manual refinement (G2), or human expert feedback with AI-assisted refinement (G3). No significant between-group differences were found in overall workload, trust or technology acceptance. However, G3 participants reported higher mental demand than G2 participants and higher temporal demand than G1 participants. Interview findings indicated that AI supported early ideation and routine structuring, whereas AI-assisted refinement required prompt formulation, output checking and decision validation. Participants valued workflows that preserved expert control, enabled transparent review and retained human final judgement. These findings suggest that AI can support early-stage heuristic generation, while judgement-intensive refinement benefits from human control, task-relevant transparency and expert review.

Usability Heuristics, Artificial intelligence, Human-AI Collaboration, Cognitive Workload

1. INTRODUCTION

As a cornerstone of user experience (UX) evaluation, heuristic evaluation (HE) is valued for its efficiency and ability to identify the majority of usability flaws early in the design cycle (Nielsen and Molich, 1990; Jeffries et al., 1991; Nielsen and Landauer, 1993; Nielsen, 1994; Jaspers, 2009). The most well-known set of heuristics is perhaps Nielsen's 10 principles (Nielsen, 1994). However, these principles were developed before many contemporary interaction contexts, such as automotive smart cockpits and artificial intelligence (AI) tools, became widespread. Hermawati and Lawson (2016), in their review of studies on domain-specific usability heuristics, found that while general heuristics are useful for many interfaces, specific domains often require tailored heuristics to ensure that domain-related usability problems are properly identified. However, manually developing such sets can be time-consuming (Quiñones and Rusu, 2017). Meanwhile, recent advances in AI have presented new opportunities for research in the field of UX. AI can assist experts in brainstorming and divergent thinking (Wan et al., 2024; Khan, Shokrizadeh and Cheng, 2025). However, the proper role of AI in generating heuristics remains unclear.

We designed an experiment to investigate how UX practitioners collaborate with AI tools in the generation and refinement of usability heuristics for automotive smart cockpit, and how different human-AI collaborative workflows shape their perceived workload, trust, acceptance and overall experience. The automotive smart cockpit was selected as the study domain because in-vehicle interaction takes place in a safety-critical and multitask environment that differs substantially from conventional desktop or mobile interaction. Moreover, although Quiñones, Rojas and Barraza (2024) have developed specific heuristics for in-

© Li et al. Published by BCS Learning and Development Ltd.

vehicle infotainment systems, comparatively less attention has been given to the development of heuristics for the broader and increasingly complex automotive smart-cockpit environment. We propose the following research questions (RQs):

(i) How do different human-AI collaboration workflows influence the process of heuristic generation and refinement?
(ii) How do the three collaborative workflows compare in terms of UX practitioners' trust, acceptance and overall experience?
(iii) How do the three workflows compare in terms of cognitive workload, and does higher automation necessarily reduce expert burden?

This paper contributes to human-computer interaction (HCI) by providing empirical evidence on how UX practitioners experience different human-AI workflows for developing domain-specific usability heuristics. It extends current work on AI-assisted UX evaluation by shifting attention from AI as an evaluator to AI as a collaborator in the generation and refinement of evaluation criteria. The findings offer design implications for appropriate automation, human control, and expert review in AI-supported UX methods.

2. RELATED WORK

2.1 Domain-Specific Heuristic Development

As digital systems have become more diverse, researchers have developed tailored heuristic sets for particular technologies and contexts, such as ambient displays (Mankoff et al., 2003) and video games (Desurvire, Caplan and Toth, 2004; Pinelle, Wong and Stach, 2008). The automotive smart cockpit also provides a particularly relevant context for domain-specific heuristic development. Unlike conventional desktop or mobile interfaces, in-vehicle interaction takes place in a safety-critical and multitask environment, where drivers may need to interact with digital displays, infotainment systems, driving information, and multimodal interfaces while simultaneously maintaining attention to the primary driving task. Usability evaluation in this context therefore needs to consider domain-specific factors such as driver distraction, visual and cognitive workload, time-critical interaction, reachability, and safety. Previous research has similarly highlighted the need for specialised usability evaluation approaches for in-vehicle information systems to improve usability without compromising driver safety (Harvey et al., 2011). More recently, Quiñones, Rojas and Barraza (2024) proposed specific UX heuristics for in-vehicle infotainment systems, further demonstrating the value of domain-specific evaluation criteria in automotive interfaces. However, comparatively less attention has been given to the systematic development of usability heuristics for the broader automotive smart-cockpit environment, which increasingly integrates multiple displays, information sources, interaction modalities, and safety-related functions. This motivates the investigation of more efficient approaches for generating and refining domain-specific heuristics in this context.

However, developing domain-specific heuristics is itself a complex and resource-intensive process. Quiñones and Rusu (2017) showed that developing usability heuristics commonly involves literature review, analysis of existing heuristics, expert input, validation, and iterative refinement. This presents a real challenge for UX practitioners who need more personalised evaluation criteria but may not have enough resources to develop them manually.

2.2 AI in Heuristic Evaluation

Recent advances in AI have created new opportunities to support UX evaluation. In HE, AI has mainly been explored as a tool for identifying usability problems in interfaces based on existing heuristic sets. For example, research has explored using large language models (LLMs) like GPT-4 to perform automated HE to provide problem feedback for user interface models (Meinecke et al., 2024). However, most existing AI-supported HE research focuses on using AI to evaluate interfaces after the heuristics have already been defined. Less attention has been paid to the earlier methodological task of generating, adapting, and refining heuristic sets themselves. This is an important gap because domain-specific heuristics often need to be tailored to particular technologies, users, and contexts.

Related research on AI-supported creative work suggests that LLMs can support early ideation by generating diverse ideas and helping user structure initial thoughts (Wan et al., 2024; Khan, Shokrizadeh and Cheng, 2025). In knowledge work, LLMs are often used not only for generating new content, but also for reducing tedious work and improving existing text. Brachman et al. (2025) found that knowledge workers used LLMs for tasks such as improving text and expected them to be integrated into everyday workflows. Similarly, AI writing systems have been shown to support text editing, polishing, paraphrasing, and ideation, suggesting that AI may be particularly useful for lower-level language and organisation tasks while humans retain responsibility for final judgement (Shi et al., 2022; Raheja et al., 2023). These may be valuable for heuristic generation, where practitioners need to synthesise domain knowledge, interaction principles, and usability concerns.

2.3 Human-AI Collaboration

Research on automation suggests that higher automation does not always reduce human workload. Bainbridge (1983) argued that automation can shift human work from direct execution to monitoring, interpretation, and intervention. More recent work extends this argument to generative AI. Simkute et al. (2024) argue that generative AI systems may cause productivity loss when they shift users' roles from production to evaluation, restructure workflows in unhelpful ways, interrupt existing task processes, or make difficult tasks harder because users must verify and manage AI outputs. Similarly, Parasuraman, Sheridan and Wickens (2000) proposed that automation should be understood through different types and levels of human–automation interaction, suggesting that the effectiveness of automation depends on how tasks are allocated between humans and systems rather than on automation level alone. This issue is also relevant to AI-supported UX work.

Trust, control, and transparency are central to human-AI collaboration. Lee and See (2004) argue that users should develop appropriate reliance on automation, rather than either rejecting it completely or accepting it uncritically. Amershi et al. (2019) further emphasise that human-AI systems should support users in understanding system behaviour, controlling interactions, and recovering from errors. Shneiderman (2020) similarly argues that effective human-centred AI should combine high levels of automation with meaningful human control. Transparency can support this process by helping users understand what the AI is doing, why it produces certain outputs, and when human intervention is needed. However, transparency is not simply about providing more information. Kizilcec (2016) showed that the amount and type of transparency can influence trust in algorithmic systems. Ehsan et al. (2021) further argue that explainability should move beyond technical explanations and include social transparency, helping users understand the social and contextual factors behind AI behaviour. Similarly, Vössing et al. (2022) emphasise that transparency should be designed to support effective human-AI collaboration rather than treated as a generic system feature. These perspectives suggest that AI should not be treated simply as a replacement for human expertise. Instead, its value depends on how tasks are distributed between AI support and human judgement, and whether users can understand, question, and control AI contributions during the workflow.

3. METHODOLOGY

3.1 Participants

This study involved a total of 30 participants (N = 30). Participants were recruited via the LinkedIn network, mainly from the United Kingdom and other European countries. Most participants were aged between 26 and 35 (n = 17) or 36 and 45 (n = 9), with a small number of participants aged 18–25 (n = 3) and over 55 (n = 1). In terms of professional tenure, the sample demonstrated a strong foundation in the user experience (UX) field: 11 participants possessed 5–10 years of experience, 6 participants had over 10 years, and the remaining 13 individuals had fewer than 5 years of experience (1–3 years: n = 7; 3–5 years: n = 6). Regarding current occupational roles, half of the sample identified as UX Researchers (n = 15), complemented by PhD Students (n = 6), UX Designers (n = 5), and Product Designers (n = 4).

3.2 Experimental Design

To ensure a consistent research environment, this study developed a web-based interface, as shown in Figure 1. Google Gemini was used during the development of the experimental environment to assist with drafting and debugging the Python and Streamlit code. All AI-generated code was reviewed, modified and tested by the researchers before the study was deployed.

The experimental interface was implemented in Streamlit and connected through the Ollama chat API to the locally hosted llama3:8b model, an 8-billion-parameter Llama 3 model (Grattafiori et al., 2024). Responses were generated without streaming, using a temperature of 0.3, a maximum generation length of 1,024 tokens and a context window of 4,096 tokens. No random seed was fixed; therefore, model responses could vary across otherwise similar interactions. Each request included the same system prompt together with the user–assistant conversation history accumulated within the current session. The model did not directly process the smart-car cockpit image. Although the image was displayed to participants through the interface, the model received a researcher-authored textual description of its nine labelled interaction areas. The system prompt positioned the model as an experienced UX/HCI thinking partner and instructed it to ground its suggestions in the driving context, including safety, reachability and glance time; ask clarifying questions when participant ideas were vague; avoid generic web-design advice; and organise suggested heuristics using a consistent structure.

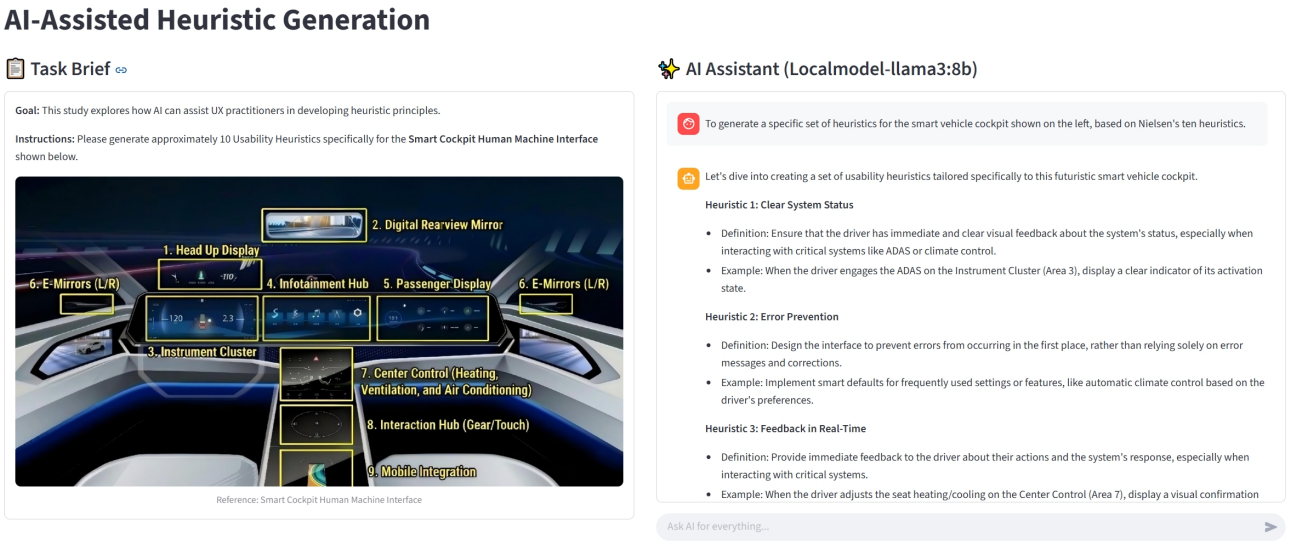


Figure 1: Screenshot of a web-based AI tool interface

All groups used the same model and the same interface. Users can interact with the AI through this interface, view and edit heuristics, and receive text feedback. Before the formal experiment, we conducted three pilot experiments to refine the task flow, operating instructions, and AI prompt templates. Participants were tasked with generating approximately 10 usability heuristics based on a provided image of a smart cockpit human-machine interface (HMI) and then refining them according to their assigned groups. The image, generated by researchers, is not a real product but a concept drawing. It covers important functions of a smart car cockpit.

The study was designed to compare three theoretically and practically meaningful human-AI workflow configurations rather than to implement a full factorial comparison of every possible combination of feedback source and refinement method. All conditions therefore included AI-assisted initial generation, because the primary aim was not to compare AI-assisted heuristic development with a fully manual baseline, but to examine how different allocations of AI and human involvement within an AI-supported workflow shaped practitioners' perceived workload, trust, acceptance and overall experience.

Participants were randomly assigned to one of the three workflow conditions, with ten participants in each group. In G1 (continuous AI-assisted feedback and refinement), participants remained in the same AI-supported session, continued the dialogue to obtain feedback and iteratively refined the set with the AI while retaining responsibility for the final decisions. In particular, G1 should not be interpreted as a condition in which AI independently provided feedback and then autonomously refined the heuristic set. The initial feedback generated by the AI was typically broad and preliminary, and participants generally needed to engage in multiple rounds of dialogue to clarify suggestions, challenge outputs, request further explanation, and progressively improve the heuristic set. Therefore, AI feedback and AI-assisted refinement formed a continuous, interactive human-AI process in which participants remained responsible for directing the interaction and making final decisions. A separate condition combining AI feedback with entirely manual refinement was not included because, within the interaction design used in this study, AI feedback was not treated as a fixed standalone input equivalent to the structured feedback for G2 and G3 provided by the internal expert panel.

An internal feedback panel, composed of three researchers from the Human Factors Research Group with substantial experience in heuristic evaluation, provided structured human feedback to participants in G2 (human expert feedback with manual refinement) and G3 (human expert feedback with AI-assisted refinement) for use in refining their heuristic sets. Each panel member independently reviewed every initial heuristic set and was blinded to both the participant's identity and group allocation. The reviews considered seven dimensions: applicability, consistency with established principles, understandability, comprehensiveness, redundancy, scalability and terminology use. The three panel members did not discuss or consolidate their assessments; instead, each participant received the three independent sets of feedback in their original form. These reviews formed part of the experimental intervention and were intended to provide formative feedback rather than to serve as comparative

outcome measures. In G2, participants manually refined their set based on feedback. In G3, participants used AI to interpret and apply the feedback, then determined the final version themselves. G2 and G3 therefore provide the most direct comparison of manual and AI-assisted refinement under a shared feedback source and timing structure. Comparisons involving G1 represent differences between complete workflows, because G1 also differed in feedback source, immediacy and interaction structure.

3.3 Experimental Procedure

Regarding the experimental procedures, all participants first completed a consent form and a pre-experiment questionnaire. They then generated and refined heuristics according to their assigned workflow. G1 completed the full procedure in one session lasting 45–65 minutes, including heuristic generation and refinement (20–30 minutes), the post-study questionnaire (5 minutes), and a semi-structured interview (20–30 minutes). Since the internal panel required time to prepare the structured feedback, participants in G2 and G3 completed the study across two sessions. The first session focused on drafting the initial heuristic set (10–20 minutes). The second session was scheduled according to participants' availability, with intervals ranging from 2 to 14 days and most participants completing it within the same week. The median interval between sessions was 5 days for G2 and 6 days for G3. This session included refinement based on the internal panel feedback (10 minutes), the post-experiment questionnaire (5 minutes), and a semi-structured interview (20–30 minutes). During refinement, G2 made manual changes, while G3 used the same AI tool to interpret and apply the feedback from the internal panel.

3.4 Data Collection and Analysis

Data collection in this study comprised three sequential phases: a pre-experiment questionnaire, a post-experiment questionnaire, and a post-task semi-structured interview. Ethical approval was obtained. Participants provided informed consent before the study, and all data were anonymised before analysis.

The pre-experiment questionnaire, deployed via Microsoft Forms, was systematically designed to capture participants' demographic profiles, professional roles, domain expertise, prior experience with HE, and baseline attitudes towards AI using the General Attitudes towards Artificial Intelligence Scale (GAAIS) (Schepman and Rodway, 2020). These baseline metrics were specifically collected to confirm that the three experimental groups were statistically equivalent and that no significant pre-existing variances existed across these critical dimensions. Kruskal-Wallis tests were applied to empirically verify group equivalence.

The post-experiment questionnaire, also administered through Microsoft Forms following the task, captured multi-dimensional user experience metrics relative to the human-AI collaboration workflow. The questionnaire includes the NASA-TLX (Task Load Index) scale, used to assess cognitive, physical, and time demands (Hart and Staveland, 1988); the Trust between People and Automation (TPA) scale to quantify subjective trust dimensions (Jian, Bisantz and Drury, 2000); and the Technology Acceptance Model (TAM) to measure perceived usefulness (PU), perceived ease of use (PEOU), and behavioural intention (BI) regarding future tool adoption (Davis, 1989). Participant-level composite scores were calculated by averaging the relevant items within each scale or subscale. These composite scores were then used for group-level descriptive comparisons and Kruskal-Wallis tests.

Finally, post-experiment semi-structured interviews were conducted and recorded via Microsoft Teams, with the participants' explicit consent, to gather deeper qualitative insights into their collaborative experiences. The interview systematically explored several aspects, including the Overall Process, Task Allocation, Cognitive Workload, Explainability, Transparency, Trust, and ideal future AI tools. The interview recordings were transcribed and analysed using thematic coding following the framework by Braun and Clarke (2006). Researchers iteratively coded the transcripts to extract key themes. To achieve a comprehensive and robust interpretation, these qualitative findings were systematically triangulated with the quantitative metrics.

4. RESULTS

4.1 Pre-experiment Questionnaire Results

The Kruskal-Wallis test showed no significant differences among the three groups in UX knowledge ($p = .246$), HE experience ($p = .209$), and prompt-writing ability ($p = .769$). The GAAIS results showed no significant differences among the three groups of participants in terms of positive ($p = .568$) and negative ($p = .914$) attitudes toward artificial intelligence, indicating a relatively balanced initial level of trust and risk perception. This conclusion remained valid even after excluding participants who failed the attention test (P25). The attention test is included only in the GAAIS scale of the pre-experiment questionnaire and has no impact on the post-experiment questionnaire.

4.2 Post-experiment Questionnaire Results

First, a composite score was calculated for each participant, and then the composite scores for these participants were averaged within each group. Kruskal-Wallis tests showed no significant differences between groups on overall score of NASA-TLX (workload), TPA (trust) and TAM (perceived usefulness, perceived ease of use and adoption intention). However, based on these group-level descriptive means, G2 showed the lowest overall NASA-TLX workload score and the highest scores for TPA trust, perceived usefulness, perceived ease of use, and behavioural intention. This descriptive trend suggests that G2 may have potential advantages in improving user experience and reducing perceived workload.

To examine the specific sources of workload, Kruskal-Wallis tests were conducted on the six NASA-TLX sub-dimensions. There was a significant difference in Mental Demand between the three groups, $H(2, N = 30) = 6.19$, $p = .045$. Pairwise Mann–Whitney U tests with Bonferroni correction showed that G2 (Mdn = 2.00, IQR = 1.00) reported significantly lower Mental Demand than G3 (Mdn = 3.00, IQR = 0.00), $p = .049$. There was also a significant difference in Temporal Demand, $H(2, N = 30) = 7.15$, $p = .028$. Post-hoc comparisons showed that G1 (Mdn = 2.50, IQR = 1.00) reported significantly lower Temporal Demand than G3 (Mdn = 4.00, IQR = 1.00), adjusted $p = .040$. No significant differences were found for Physical Demand, Performance, Effort, or Frustration Level, all $ps > .293$. However, the descriptive trend consistently favoured G2, which had the highest performance, while its effort and frustration levels were lower than those of G1.

Table 1: *Comparison of NASA-TLX sub-dimensions*

Sub-dimension	H(2, N = 30)	p-value	Post-hoc comparison
Mental Demand	*6.19*	*0.045**	*G2 < G3, p = .049*
Temporal Demand	*7.15*	*0.028**	*G1 < G3, p = .040*
Physical Demand	*1.63*	*0.442*	*-*
Performance	*2.22*	*0.329*	*-*
Effort	*2.45*	*0.294*	*-*
Frustration	*2.46*	*0.293*	*-*

4.3 Interview Results

The semi-structured interviews explored participants' overall experiences of the workflow, task allocation between humans and AI, perceived workload, trust, transparency, and their preferred forms of future AI support. Our thematic analysis identified six themes: AI as a productivity tool for early-stage generation and routine work; human control, accountability, and cognitive work in refinement; supporting factors for reliable AI-assisted generation; interaction and workflow friction; requirements for domain-specialised and traceable AI support; and organisational and privacy conditions for adoption. The following sections will elaborate on the three themes that most directly address the research questions.

4.3.1. AI as a productivity tool for early-stage generation and routine work

The interview findings show that higher automation did not always reduce participants' workload. In the early generation stage, AI was often seen as useful because it helped participants create an initial structure and quickly access ideas about smart cockpit interfaces. For example, P03 reported that AI helped them 'generate general concepts and understand the task requirements' during the first session. P15 similarly described AI as useful for quickly generating key points and drafting an initial version, which saved the time needed for manual drafting. P18 stated that work which might normally take 'two or three days could be completed by AI in a short time', while P30 described AI as an 'excellent starting point and foundation' for participants with limited experience in heuristic generation. These comments suggest that AI was most helpful when it acted as a starting point rather than a final decision-maker.

At the same time, AI was also valued for handling routine and lower-level work. Participants repeatedly described AI as useful for tasks such as formatting, prioritising, checking consistency, organising scattered ideas, and turning rough material into a more readable document. P07 described AI as a 'smart administrative assistant that could take on grunt work', including formatting and prioritisation. P14 stated that AI was suitable for low-level tasks, including text formatting and consistency checking, and for turning divergent ideas into a clearer format.

4.3.2. Human control, accountability, and cognitive work in refinement

However, in the refinement stage, the role of AI became more complex. Some participants in G2 felt that manual editing based on expert feedback gave them greater autonomy, responsibility, and control. P08 stated that manual revision gave them a 'stronger sense of ownership' over the final heuristics, even though simplifying and reorganising overlapping points was mentally demanding. P11 also saw manual revision as linked to accountability, comparing the process to using 'flashcards' because it helped them teach themselves and internalise the feedback. P14 was even more direct, stating that they did not want AI to directly modify the result and preferred to retain full control.

Prompt writing helped explain why G3 participants experienced higher Mental Demand than G2 participants. Participants had to express expert feedback clearly enough for AI to understand. P03 reported that understanding the task, forming ideas, and typing instructions required mental effort, and the time limit made the feedback stage more stressful. P12 similarly reported that AI did not always understand small revision intentions and sometimes moved the output in the wrong direction. These examples suggest that using AI for refinement added an extra communication step between human judgement and final output. Another reason for the higher perceived workload in G3 was the need for constant checking. G3 participants did not fully trust AI output as a final answer. P06 described AI as a 'junior researcher' whose output could be generic or irrelevant and therefore needed human review. P15 believed that humans should keep the right to judge whether the output was correct. P18 stated that they would not directly copy and paste AI output without checking it. P24 also preferred keeping humans in the loop because AI might produce biased, incorrect, or over-agreeable results. This shows that even in a highly automated workflow, participants still felt responsible for the quality of the final heuristics.

4.3.3. Supporting factors for reliable AI-assisted generation

Transparency also emerged as an important factor in how participants judged and managed AI output. Participants did not only want AI to generate or revise heuristics; they also wanted to understand why a specific output was produced and where the information came from. P07 described the AI as a 'black box' and explained that they could not directly observe or control its reasoning process. They argued that the reasoning process should be more visible, because this would allow users to intervene when the output moved in the wrong direction. P09 similarly wanted to see AI's thought processes and reference sources and suggested that transparency would increase the credibility of the generated heuristics. P15 also stated that seeing AI's internal reasoning would help experts judge whether a heuristic should be included in the final list. P20 expressed a similar need for AI to cite industry reports or previous studies, especially because smart cockpit design is a safety-related domain where users should not blindly trust AI output. P27 also preferred a visible logic chain, because it would help users judge whether AI's suggestions were reasonable. These comments suggest that transparency was not only a matter of trust, but also a practical support for expert review, intervention, and final decision-making.

In addition, regarding human feedback versus AI feedback, the interview results showed that many participants believed that human reviewers performed better in identifying overlaps, ambiguities, lack of context, and actual defects in heuristic sets. For example, P09 argued that AI feedback on its own could be too positive and might fail to challenge its own output, while human experts could notice redundancy and repetition more effectively. P17 also preferred human expert feedback because it could bring 'new perspectives' that the participant had not considered. Similarly, P19 believed that human experts were more valuable because AI tended to be too agreeable, whereas human reviewers could provide more critical comments. P20 also trusted human feedback more when several experts reached similar conclusions, especially in a safety-related domain such as smart cockpits.

5. DISCUSSION

The discussion is organised around the three research questions. First, it explains how the three workflows shaped the process of heuristic generation and refinement. Second, it discusses why G2 appeared to provide a more positive user experience. Third, it explains why higher automation did not necessarily reduce cognitive workload.

5.1 Appropriate Automation in AI-Assisted Heuristic Generation

In response to RQ1 and RQ3, we found that different human-AI workflows changed not only the amount of work, but also the type of work that participants had to perform. The most direct comparison was between G2 and G3, which shared the same expert-feedback source and delayed-session structure but differed in whether refinement was manual or AI-assisted. G3 participants reported significantly higher mental demand than G2 participants, while interview accounts indicated that AI-assisted refinement introduced additional work through prompt formulation, interpretation and output verification. Together, these findings suggest that AI assistance may redistribute rather than eliminate cognitive work during judgement-intensive refinement. This pattern reflects Bainbridge's (1983) classic "ironies of automation": when automation takes over direct execution, human work can shift towards monitoring, interpretation and intervention. In the present study, this redistribution was reflected in participants' need to formulate prompts, interpret AI revisions and verify outputs before making final decisions. It is also consistent with recent work on the ironies of generative AI, which argues that generative AI can shift users' roles from production to evaluation, restructure workflows in unhelpful ways, and make difficult tasks even harder when users must verify and manage AI outputs (Simkute

et al., 2024). In the context of heuristic generation, refinement is not a simple execution task. It requires experts to understand the intention behind feedback, judge whether a heuristic is meaningful, and decide how the wording should be adjusted for a specific domain. When AI was used to perform this refinement work, participants still had to supervise the process closely. Therefore, the workload was not removed; it was redistributed into prompt formulation, output checking, and decision validation. This supports the view that automation should be applied at an appropriate level and to appropriate task types, rather than being treated as a simple way to reduce human effort (Parasuraman, Sheridan and Wickens, 2000).

The finding also suggests that the value of AI depends on where it is placed in the workflow. AI was more useful when supporting open-ended drafting, because this stage benefits from fast idea generation and broad knowledge synthesis. This is consistent with recent human-AI co-creation research showing that LLMs can support early ideation and divergent thinking, while still requiring users to remain actively involved in directing the process (Wan et al., 2024; Khan, Shokrizadeh and Cheng, 2025). In addition, participants valued AI for routine and lower-level work. This aligns with prior work showing that knowledge workers use AI to reduce tedious tasks and improve text, and that AI writing systems can support editing, polishing, coherence improvement, and idea organisation (Shi et al., 2022; Raheja et al., 2023; Brachman et al., 2025).

However, refinement is a more judgement-based activity, where small wording changes can affect clarity, applicability, and domain relevance. In this stage, excessive automation may reduce the feeling of direct control and increase the need for verification. This aligns with human-centred AI principles, which argue that effective AI systems should combine high levels of automation with meaningful human control, rather than simply pursuing full automation (Shneiderman, 2020). Therefore, in AI-assisted heuristic generation, a more effective design goal may be appropriate automation: using AI to support the construction of initial frameworks and to handle routine and low-level tasks such as formatting, prioritisation, and consistency checks, while maintaining the central role of expert judgment in tasks that require explanation and ultimate accountability.

5.2 Maintaining Human Control through Transparency and Final Review

In response to RQ2, the workflow that yielded a better subjective user experience was one that preserved human control while still using AI for useful support. This pattern was complemented by qualitative accounts from G2 participants, who described manual refinement as supporting ownership, accountability and direct control. Transparency and human review were central to this experience, particularly during refinement and final review. In this study, participants did not treat AI as an autonomous decision-maker. Although AI was useful for generating initial ideas and organising heuristic content, participants still felt responsible for judging whether the output was clear, relevant, and appropriate for the smart cockpit context. This was especially evident in the refinement stage, where small wording changes could affect the meaning, applicability, and domain relevance of a heuristic.

The need for transparency was particularly visible in the G3 workflow, where AI was used to interpret and apply human expert feedback. Participants had to check whether AI had correctly understood the feedback, whether the revised heuristics preserved the intended meaning, and whether any important contextual detail had been lost. In this sense, transparency was not only a matter of explaining how the AI system worked, but also a practical condition for effective review. This aligns with work on social transparency and human-AI collaboration, where transparency can support informed decision-making, trust calibration, accountability, and effective collaboration (Ehsan et al., 2021; Vössing et al., 2022). However, the findings also suggest that transparency should be carefully designed. Participants did not appear to require full technical explanations of the AI model. Instead, they needed lightweight and task-relevant transparency, such as visible revision logic, clear indications of what had changed, source references, and uncertainty cues. This is important because previous research has shown that transparency is not simply a matter of providing more information; excessive or poorly targeted information may reduce trust or increase cognitive effort (Kizilcec, 2016).

Human expert feedback also remained important as a form of quality control. Participants valued human feedback because it could identify ambiguity, redundancy, missing context, and practical weaknesses that AI feedback might overlook. This suggests that expert review is not only useful for improving the content of heuristics, but also for ensuring that the final heuristic set remains meaningful, context-sensitive, and usable in practice. Even when AI was involved in refinement, participants still needed to decide whether the output should be accepted, revised, or rejected. These findings reflect the principle of appropriate reliance on automation, where users should neither reject AI support nor accept AI outputs without critical evaluation (Lee and See, 2004). They also align with human-AI interaction guidelines, which emphasise that AI systems should support user understanding, control, and

recovery from errors rather than reduce users to passive recipients of automated output (Amershi et al., 2019). Therefore, human control should not be viewed as a temporary limitation of current AI systems. Instead, it should be treated as a necessary design principle for AI-assisted heuristic generation. AI can support the production and organisation of heuristic ideas, but expert judgement should remain central when interpreting feedback, refining heuristic wording, and finalising the heuristic set.

5.3 Limitations and Future Work

This study has several limitations. First, the sample size was relatively small, with ten participants in each group, so the findings should be interpreted with caution. Second, this study mainly focused on participants' perceived workload, trust, acceptance and interview feedback, but did not include an independent quality assessment of the final heuristic sets. We therefore cannot determine whether the workflows produced heuristics of different quality, utility or suitability. A necessary next step is a blinded external evaluation of final heuristic sets. Third, the study was conducted in the context of smart cockpit interfaces, which are complex and safety related. Future research could test whether similar human-AI collaboration patterns appear in other UX domains. In addition, participants worked with a single researcher-generated concept image in a controlled task, which may not reflect the iterative and organisationally situated nature of heuristic development in practice. Finally, G2 and G3 had the same internal-expert feedback and delayed-session structure, whereas G1 differs in terms of feedback source, immediacy and interaction pattern. Comparisons involving G1 therefore reflect a bundle of workflow differences, and other plausible configurations were not tested. Future studies should combine subjective measures with behavioural performance indicators, independent output-quality assessment and systematic analysis of AI response variability.

6. CONCLUSION

This study compared three human-AI collaborative workflows for generating and refining usability heuristics for smart cockpit interfaces. Although no significant between-group differences were found in overall workload, trust or technology acceptance, the NASA-TLX sub-dimension results indicated that participants in G3 reported higher mental demand than those in G2 and higher temporal demand than those in G1. The interview findings helped explain these differences by showing that AI-assisted refinement did not remove the need for expert judgement. Instead, it introduced additional work through prompt formulation, interpretation of AI suggestions, output verification and final decision-making.

The findings therefore support the principle of appropriate automation rather than maximising AI involvement throughout the workflow. AI was perceived as particularly useful for early-stage idea generation, initial structuring and routine tasks such as formatting and consistency checking. However, judgement-intensive refinement benefited from direct human control, expert feedback and task-relevant transparency. Participants valued being able to understand, question and selectively apply AI suggestions while retaining responsibility for the final heuristic set. Overall, AI should be designed as a supportive collaborator in heuristic development rather than as an autonomous replacement for UX expertise.

7. REFERENCES

Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S.T., Bennett, P.N., Inkpen, K., Teevan, J., Kikin-Gil, R. and Horvitz, E. (2019) Guidelines for human-AI interaction. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. New York: ACM, 1–13. doi: 10.1145/3290605.3300233.

Bainbridge, L. (1983) Ironies of automation. Automatica, 19(6). 775–779. doi: 10.1016/0005-1098(83)90046-8.

Brachman, M., El-Ashry, A., Dugan, C. and Geyer, W. (2025) Current and future use of large language models for knowledge work. Proceedings of the ACM on Human-Computer Interaction, 9(7). 1–24. doi: 10.1145/3757403.

Braun, V. and Clarke, V. (2006) Using thematic analysis in psychology. Qualitative Research in Psychology, 3(2). 77–101. doi: 10.1191/1478088706qp063oa.

Davis, F.D. (1989) Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Quarterly, 13(3). 319–340. doi: 10.2307/249008.

Desurvire, H., Caplan, M. and Toth, J.A. (2004) Using heuristics to evaluate the playability of games. In: CHI '04 Extended Abstracts on Human Factors in Computing Systems. New York: ACM, 1509–1512. doi: 10.1145/985921.986102.

Ehsan, U., Liao, Q.V., Muller, M., Riedl, M.O. and Weisz, J.D. (2021) Expanding explainability:

Towards social transparency in AI systems. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. New York: ACM, Article 82, 1–19. doi: 10.1145/3411764.3445188.

Grattafiori, A. et al. (2024) The Llama 3 herd of models. arXiv preprint arXiv:2407.21783. doi: 10.48550/arXiv.2407.21783.

Hart, S.G. and Staveland, L.E. (1988) Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In: Hancock, P.A. and Meshkati, N. (eds). Human Mental Workload. Amsterdam: North-Holland, 139–183. doi: 10.1016/S0166-4115(08)62386-9.

Harvey, C., Stanton, N.A., Pickering, C.A., McDonald, M. and Zheng, P. (2011) A usability evaluation toolkit for In-Vehicle Information Systems (IVISs). Applied Ergonomics, 42(4). 563–574. doi: 10.1016/j.apergo.2010.09.013.

Hermawati, S. and Lawson, G. (2016) Establishing usability heuristics for heuristics evaluation in a specific domain: Is there a consensus? Applied Ergonomics, 56. 34–51. doi: 10.1016/j.apergo.2015.11.016.

Jaspers, M.W.M. (2009) A comparison of usability methods for testing interactive health technologies: Methodological aspects and empirical evidence. International Journal of Medical Informatics, 78(5). 340–353. doi: 10.1016/j.ijmedinf.2008.10.002.

Jeffries, R., Miller, J.R., Wharton, C. and Uyeda, K. (1991) User interface evaluation in the real world: A comparison of four techniques. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. New York: ACM, 119–124. doi: 10.1145/108844.108862.

Jian, J.-Y., Bisantz, A.M. and Drury, C.G. (2000) Foundations for an empirically determined scale of trust in automated systems. International Journal of Cognitive Ergonomics, 4(1). 53–71. doi: 10.1207/S15327566IJCE0401_04.

Khan, A., Shokrizadeh, A. and Cheng, J. (2025) Beyond Automation: How Designers Perceive AI as a Creative Partner in the Divergent Thinking Stages of UI/UX Design. In: Proceedings of the CHI Conference on Human Factors in Computing Systems. New York: ACM. doi: 10.1145/3706598.3713500.

Kizilcec, R.F. (2016) How much information? Effects of transparency on trust in an algorithmic interface. In: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems. New York: ACM, 2390–2395. doi: 10.1145/2858036.2858402.

Lee, J.D. and See, K.A. (2004) Trust in automation: Designing for appropriate reliance. Human Factors, 46(1). 50–80. doi: 10.1518/hfes.46.1.50_30392.

Mankoff, J., Dey, A.K., Hsieh, G., Kientz, J., Lederer, S. and Ames, M. (2003) Heuristic evaluation of ambient displays. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. New York: ACM, 169–176. doi: 10.1145/642611.642642.

Meinecke, A., Heidrich, D., Dworatzyk, K. and Theis, S. (2024) A comparative heuristic evaluation of Kadi4Mat through human evaluators and GPT-4. In: HCI International 2024 – Late Breaking Papers. Lecture Notes in Computer Science, 15380. Cham: Springer, 91–108. doi: 10.1007/978-3-031-76821-7_8.

Nielsen, J. (1994) Enhancing the explanatory power of usability heuristics. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. New York: ACM, 152–158. doi: 10.1145/191666.191729.

Nielsen, J. and Landauer, T.K. (1993) A mathematical model of the finding of usability problems. In: INTERCHI '93: Conference Proceedings. New York: ACM, 206–213. doi: 10.1145/169059.169166.

Nielsen, J. and Molich, R. (1990) Heuristic evaluation of user interfaces. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. New York: ACM, 249–256. doi: 10.1145/97243.97281.

Parasuraman, R., Sheridan, T.B. and Wickens, C.D. (2000) A model for types and levels of human interaction with automation. IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans, 30(3). 286–297. doi: 10.1109/3468.844354.

Pinelle, D., Wong, N. and Stach, T. (2008) Heuristic evaluation for games: Usability principles for video game design. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. New York: ACM, 1453–1462. doi: 10.1145/1357054.1357282.

Quiñones, D. and Rusu, C. (2017) How to develop usability heuristics: A systematic literature review. Computer Standards & Interfaces, 53. 89–122. doi: 10.1016/j.csi.2017.03.009.

Quiñones, D., Rojas, L.F. and Barraza, A. (2024) User experience heuristics for in-vehicle infotainment systems. Procedia Computer Science, 237. 725–732. doi: 10.1016/j.procs.2024.05.159.

Raheja, V., Kumar, D., Koo, R. and Kang, D. (2023) CoEdIT: Text editing by task-specific instruction tuning. In: Bouamor, H., Pino, J. and Bali, K.

(eds). Findings of the Association for Computational Linguistics: EMNLP 2023. Singapore: Association for Computational Linguistics, 5274–5291. doi: 10.18653/v1/2023.findings-emnlp.350.

Schepman, A. and Rodway, P. (2020) Initial validation of the general attitudes towards artificial intelligence scale. Computers in Human Behavior Reports, 1. Article 100014. doi: 10.1016/j.chbr.2020.100014.

Shi, S., Zhao, E., Tang, D., Wang, Y., Li, P., Bi, W., Jiang, H., Huang, G., Cui, L., Huang, X., Zhou, C., Dai, Y. and Ma, D. (2022) Effidit: Your AI writing assistant. arXiv preprint arXiv:2208.01815. doi: 10.48550/arXiv.2208.01815.

Shneiderman, B. (2020) Human-centered artificial intelligence: Reliable, safe & trustworthy. International Journal of Human–Computer Interaction, 36(6). 495–504. doi: 10.1080/10447318.2020.1741118.

Simkute, A., Tankelevitch, L., Kewenig, V., Scott, A.E., Sellen, A. and Rintel, S. (2024) Ironies of generative AI: Understanding and mitigating productivity loss in human-AI interaction. International Journal of Human–Computer Interaction, 41(5). 1–22. doi: 10.1080/10447318.2024.2405782.

Vössing, M., Kühl, N., Lind, M. and Satzger, G. (2022) Designing transparency for effective human-AI collaboration. Information Systems Frontiers, 24(3). 877–895. doi: 10.1007/s10796-022-10284-3.

Wan, Q., Hu, S., Zhang, Y., Wang, P., Wen, B. and Lu, Z. (2024) 'It felt like having a second mind': Investigating human-AI co-creativity in prewriting with large language models. Proceedings of the ACM on Human-Computer Interaction, 8(CSCW1). Article 84, 1–26. doi: 10.1145/3637361.