Understanding Challenges and Needs of Using AI in Web Automation Systems

Jiacheng Zhang
jiache@umich.edu
University of Michigan
Ann Arbor, Michigan, USA

Carl Fan
carlfan@umich.edu
University of Michigan
Ann Arbor, Michigan, USA

Steve Oney
soney@umich.edu
University of Michigan
Ann Arbor, Michigan, USA

Abstract

Web automation—the process of simulating user interactions with websites—has gained significant attention due to advances in Artificial Intelligence and increased digitization of services. While web automation offers potential benefits by streamlining tasks and addressing accessibility barriers, it also introduces unique challenges due to the complexities of automating interactions with interfaces designed for humans. Beyond technical advancements, addressing these challenges requires meeting human-centric needs and concerns surrounding web automation. In this paper, we explore critical questions about users' web automation needs and preferences through a series of interviews with a diverse group of participants, including individuals across various age ranges and users with visual impairments. Our findings offer insights on how users weigh concerns such as privacy, error rates, efficiency, and usefulness in deciding what level of automation might be appropriate for a given task. We highlight critical areas for improvement in automation tools and design principles for future systems.

1 Introduction

Advances in Artificial Intelligence (AI) and the increasing digitization of services have led to a growing interest in *web automation*—the process of automating tasks on websites that are typically performed by users, by simulating user events such as mouse clicks and key presses. Web automation can streamline repetitive tasks, improve efficiency, help users overcome accessibility issues (from permanent, temporary, or situational disabilities), and more [33, 47]. Web automation is inherently more difficult and error-prone than other forms of software automation (such as code that references external APIs) because it interacts with User Interface (UI) elements designed to be used by humans. These UI elements may not have consistent identifiers or may rely on visual cues that humans can easily interpret, but require automated scripts to accurately recognize and interact with [29].

Web automation has been the subject of much prior work, including both academic [11, 14, 15, 30, 33, 65] and commercial systems [4, 40, 42, 43, 48, 56]. The advent of Large Multi-Modal Models (LMMs) has made UI automation increasingly viable. Still, there is much work to be done in the space of web automation; existing tools suffer from limited scope and high error rates. Much prior work has focused on improving the accuracy and efficiency of web automation tools. However, as we will describe, prior work has largely focused on technical aspects such as improving automation accuracy and interaction efficiency—often overlooking nuanced human-centric needs, such as task-specific preferences, control, trust, and accessibility challenges. The success of future web automation tools will depend upon their ability to meet users' needs and integrate with their workflows. Thus, it is crucial for researchers and practitioners working on web automation tools to understand these human-centric challenges.

In this paper, we seek to provide insights to key questions about human-centric needs in web automation:

- **RQ1:** What kinds of tasks would users want to automate?
- **RQ2:** What factors influence users' preferred level of automation (i.e., how much agency the automation can take) and interaction mode?
- **RQ3:** What key concerns (e.g., error rates, privacy, efficiency) shape users' willingness to automate different tasks?
- **RQ4:** What specific concerns (e.g., errors, privacy, efficiency) influence user expectations for future automation tools, and what design directions might address them?
- **RQ5:** How do user backgrounds—including technical familiarity and accessibility needs—shape preferences and requirements for automation tools?

It is important for researchers and practitioners to understand the answers to human-centric questions to create realistic datasets and design tools that meet users' needs. To answer these questions, we conducted a series of interviews with potential end-users about the types of tasks they would want to automate. Our participants spanned a wide range of ages and technological skill levels. The participant pool also included Blind/Visually Impaired (BVI) users who interact with the web through Accessibility Technologies (ATs), providing valuable insights into the accessibility challenges and opportunities in web automation. This paper contributes:

- A thorough analysis of users' needs and preferences for web automation across a diverse set of 24 users.
- Insights into the factors that influence users' willingness to automate web-based tasks, including task complexity, privacy concerns, and users' backgrounds.
- Recommendations for designing web automation tools that address the human-centric challenges.

2 Related Work

2.1 Human-AI Interaction

While AI research focuses on model accuracy, Human-Centered AI (HCAI) emphasizes understanding human agency and values [10]. As AI systems become more autonomous, research must examine how they affect human interaction and behavior [3, 62].

2.1.1 Challenges of Human-AI Interaction. Modern AI systems possess capabilities including sensing (pattern recognition), reasoning (decision-making and inference), autonomous operation, and adaptation through machine learning [20, 45, 50, 61]. While these capabilities enable more sophisticated automation, they can also

mask errors. This reduced predictability increases the risk of over-reliance on automation. Users may not fully understand the AI's decision-making process or recognize when they need to intervene.

AI systems introduce unique challenges due to their inherent uncertainty and complexity [62]. Unlike deterministic non-AI systems, AI outcomes can be unpredictable due to probabilistic machine learning models [61]. The integration of AI into people's daily lives raise a number of socio-technical concerns. These socio-technical challenges in HCAI not only come from the construction of models and systems themselves [18] but also from the role of AI in user-centered applications [7]. Therefore, it is worth scoping down the challenges in HCAI and studying the specific user challenges of applying AI in UI automation.

2.1.2 Human-centered Design. The neglect of human needs and values underlies many socio-technical challenges in AI systems [18]. Rather than rejecting AI due to uncertainty, designers should focus on leveraging its capabilities while minimizing unintended consequences. Two key principles guide effective human-centered AI design. First, AI interfaces must facilitate clear communication, combining intuitive visual design with natural language explanations matched to user expertise. This enables users to better understand, predict, and control AI behavior, fostering trust and effective interaction [52]. Second, continuous user feedback mechanisms are essential for understanding real-world performance and aligning systems with human values and ethical standards [31]. While research has explored AI applications across disciplines like healthcare, economics, and law [7], significant gaps remain in addressing user needs within human-AI interaction systems. Recent research in prototyping AI systems [26, 54, 55] demonstrates the importance of engaging diverse stakeholders early to ensure systems align with real user needs. This is particularly relevant for UI automation, where understanding user needs is crucial for creating systems that adapt effectively to user behaviors and preferences.

2.2 UI automation

User Interface (UI) automation enables software scripts to automate tedious UI-related tasks like form-filling and data collection. However, implementing such scripts requires programming expertise and significant time investment to understand web structures [29].

2.2.1 Programming By Demonstration. The programming by demonstration(PBD) approach is well-studied to lower the barrier of creating web automation programs for non-experts [11, 33, 34, 36, 63]. Given a sequence of user demonstrations on a website, PBD systems could generate synthesized programs to repeat the same actions and apply them to similar elements on the website. However, the visual formats of the results programs (low-level programs or visual representations) from these systems still require familiarity with programming to understand them, which also makes it difficult for users to edit the program when errors occur. Then systems like SemanticOn [46], WebRobot [15], MIWA [13], and DiLogics [47] adopted more advanced program synthesis technique that allows users to continuously provide more demonstrations to rewrite the synthesized program. Also, natural language descriptions and visual highlighting are provided to help users understand the generated automation program [13, 30].

2.2.2 Natural Language for Automation. Another line of work explores natural language as an interface for task automation, aiming to reduce user burden through conversational or prompt-based interactions. Besides asking users to directly perform demonstrations on the user interface, some PBD systems such as Sugilite [34], Appinite [35], and ParamMacros [30] have also explored ways that allow users to interact with the systems with natural language instructions. However, these systems lack flexibility in understanding varied linguistic expressions and require user demonstrations for unseen websites.

Recently, there has been a surge in the development and application of LLMs. LLMs are trained on a large corpus of data and include billions of parameters, enabling the models to capture intricate linguistic relationships in the text and lead to unparalleled performance across broad NLP tasks. A remarkable feature of LLMs is few-shot or zero-shot learning [28]. LLMs can handle unseen tasks with very few or zero targeted examples. Additionally, models like GPT-3 [17] have shown abilities in in-context learning, which enables them to adapt to new tasks using only the context provided in the prompt, without the need for direct training.

Without user demonstrations, recent works leverage LLMs to connect UI and natural language. Widget Captioning [37] and Screen Recognition [64] generate semantic labels for UI components, while Mind2Web [14] introduces a framework for generalist UI agents. Studies show LLMs perform well on mobile UI tasks [58, 59], but web UI presents unique challenges due to its dynamic nature [22, 24]. Current LLM-based web automation tools like Adept AI and Taxy AI [1, 2] face efficiency concerns. More recent tools like Operator [42] are understudied but still have high error rates.

2.2.3 Multimedia Interaction. Modern UI automation systems have expanded beyond natural language commands to incorporate multimedia interactions. Systems like Pix2Struct [32] process pixel-based inputs to parse web screenshots into HTML, while WebGUM [19] combines pre-trained vision and language models to enhance web navigation capabilities. Building on these approaches, PIX2ACT [51] effectively translates pixel-based inputs into browser actions, showing particular success on platforms like MiniWob++ and WebShop. More recently, SeeAct [65] demonstrates the potential of using GPT-4V for visual understanding in web agents, though challenges persist in converting model-generated plans into concrete web interactions.

2.2.4 Benchmarking Environments. Recent works including OSWorld [60], WebArena [66], AndroidWorld [49], and Windows Agent Arena [8] have developed environments for evaluating AI agents on real-world applications. These benchmarks, such as OSWorld's 369 computer tasks and WebArena's web interactions, provide realistic testing environments. These benchmarks focus primarily on technical metrics while overlooking human-centric aspects like privacy and user communication. Our study aims to address this gap by examining how automation technologies align with actual user needs. Prior LLM-based and multimodal works have focused on model development and simulated testing, leaving a gap in understanding real-world user needs. Research is needed to examine how these automation technologies align with actual user requirements and concerns.

3 Interpretive Study

We conducted semi-structured interviews with 24 participants from diverse backgrounds (Table 1). We focused on hypothesized scenarios rather than deployment studies for two reasons: current systems' high error rates (top models like GPT-4V and Gemini-Pro-Vision achieve $\leq$ 20% success on benchmark tasks [8, 60, 66]), and our desire to explore future possibilities beyond current technological constraints [16, 39].

3.1 Participants and Recruitment

We recruited participants through mailing lists and social media, using surveys to ensure demographic diversity. Following prior work advocating for diverse perspectives in HCI research [9, 38], our participants included 12 people aged 55+, 13 from the US, 10 from China, one from the UK, with an even split between technical and non-technical backgrounds (see Table 1). Participants were compensated up to $30 USD for completing the study, which took approximately 90 minutes. Eight studies were conducted in-person and 16 were conducted remotely. Our study protocol was approved by our Institutional Review Board (IRB).

ID	Age	Occupation	AT	ID	Age	Occupation	AT
P1	35–55	Retired	N	P13	>55	Lecturer	N
P2	>55	Engineer	N	P14	18–24	Student	N
P3	18–24	Student	N	P15	>55	Teacher	N
P4	18–24	Student	N	P16	>55	Nurse	N
P5	18–24	Student	N	P17	>55	Instructor	N
P6	25–34	Student	N	P18	>55	Not employed	Y
P7	18–24	Student	N	P19	>55	Receptionist	Y
P8	18–24	Student	N	P20	>55	Data analyst	N
P9	35–50	Info. Architect	N	P21	>55	Dir. Social Services	Y
P10	>55	Investment	N	P22	35–50	Not employed	Y
P11	18–24	Not employed	N	P23	>55	Non-profit Org.	Y
P12	18–24	Student	N	P24	>55	Homemaker	Y

Table 1: Demographics of Participants. Note that "AT" stands for "Assistive Technology"—in our case, these were screen readers for Blind and Low-Vision (BLV) participants

3.2 Study Protocol

Figure 1 diagrams the different stages of our study, as we explain in the following sub-sections:

3.2.1 Automation and Study Overview. To ensure participants understood what "web automation" entails, we started each study by defining web automation and presenting five illustrated examples of different forms of automation tools (Figure 1B and supplemental materials).

3.2.2 Task Collection (RQ1) . We then asked participants to propose 5–10 examples of web tasks that they do in their personal or professional contexts (Figure 1C). This resulted in 150 *user-defined* tasks in total, unique to each participant.

3.2.3 Per-Task Questions. We then asked participants a series of questions for each of the 5–10 tasks they proposed and for six additional "predefined" tasks (listed in Table 2 in Appendix A), which allowed us to compare responses to predefined tasks.

- **Preferences for Interaction (RQ2):** For each task, we started by asking how they would interact with automation tools for this task, such as browser extensions or AI assistants, for each of these tasks. Participants also explain the benefits (efficiency, accuracy) and tradeoffs (control, involvement). (Figure 1.D.1)
- **Degree of Automation (RQ2):** Participants then discussed their preferred level of automation for the current task (Figure 1.D.2). The discussion is framed around a six-level automation scale from non-automation to full automation. The six-level automation scale builds on prior automation taxonomies [25, 41, 44, 57] and is grounded in the context of web automation. Our scale emphasizes user control and feedback, tailoring these theoretical frameworks to the context of web-based tasks (Table 4 in Appendix B).
- **Concern Evaluation (RQ3, RQ4):** We assessed participants' concerns for each task regarding the use of such automation systems (Figure 1.D.3). We asked participants to individually rate their concerns regarding error rates, privacy, efficiency, and usefulness for each task on a 5-point scale. We also discussed with users the potential changes that can be made to mitigate their concerns (RQ4).
- **Usage Frequency (RQ3):** We also asked participants how frequently they anticipate using automation for each task to better understand potential longer-term usage patterns (Figure 1.D.4).

3.2.4 Follow-Up "Take-Home" Survey. Use cases for web automation might come to mind spontaneously as situations arise, rather than over the course of a short interview. Thus, we gave every participant a "take-home" survey and prompted them to propose additional tasks over the course of one month after they completed their interviews. Participants were compensated $2 USD per task submitted (with up to three use cases per day).

3.3 Data Analysis

We employed mixed-methods analysis of transcribed and anonymized interviews. For qualitative analysis, we conducted thematic analysis [5, 6] using an inductive approach [53]. Two researchers iteratively developed and refined a coding scheme until achieving strong inter-rater reliability (Fleiss's Kappa, $\kappa = 0.84$). Quantitatively, we used descriptive statistics to analyze preferences and categorized tasks by characteristics like decision-making, sensitivity, and communication (see Appendix C). Two authors independently labeled tasks, achieving Kappa scores above 0.8 for all categories before reaching final consensus. We then analyzed correlations between task characteristics, user concerns, automation preferences, and demographics. We integrated qualitative themes and quantitative findings to identify patterns and enhance our interpretation of the interview data.

4 Results and Findings

4.1 Data Overview

We analyzed participants' responses for 312 tasks—150 user-defined tasks (from 5–10 proposed by each participant), 143 pre-defined tasks (same for each participant), and 19 tasks from the "take-home"

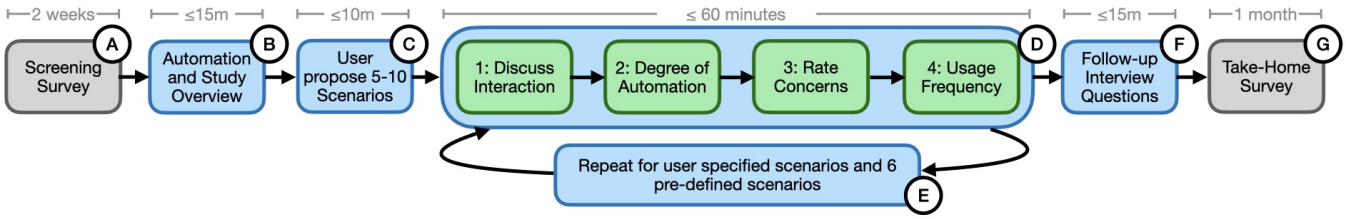

Figure 1: Users first went through a screening survey (A) to collect demographic and technical background information. During the interview, participants were introduced to the web automation concepts and demonstrations through an overview session (B). Users were asked to propose 5–10 examples of web tasks they commonly perform (C). Then they evaluated each of these tasks and six predefined tasks through discussions of interaction, preferred automation levels, concerns, and usage frequency (D). This evaluation process was repeated for all scenarios (E). Participants then answered follow-up questions to explore additional use cases or concerns (F). After the interview, they completed a continuous survey over the next month to log additional tasks incrementally (G).

survey. Each participant on average proposed 7 tasks from individual experience and browsing history. Figure 2 presents the distribution of user-defined tasks across categories.

Participants rated their preferred level of automation for both the tasks they proposed and the examples we provided. Semi-automation was most preferred (48.39%), followed by full automation (32.05%) and no automation (15.71%). On a 5-point scale, participants expressed moderate concerns about error rates (2.24), privacy (2.04), efficiency (1.89), and usefulness (1.75), with errors being the primary concern. Figure 3 details the automation preferences and concerns across predefined and user-defined tasks. In what follows, we explore participants' openness to automation, their varying preferences, and specific needs.

4.2 All Participants were Open to Automation but Preferences for Degree of Automation Depended on the Task, not the Users

Our results indicate that users are open to some degree of automation. However, the degree of automation preferred varies by task based on perceived advantages and disadvantages.

Time-saving emerged as the primary benefit, mentioned 81 times. Participants mentioned that automation should be faster than manual processes, particularly by reducing repetitive efforts such as refreshing and re-entering the same information. Accuracy was another recurring theme, with many trusting AI to fill out fixed details like addresses or account IDs more reliably than they could themselves. For example, P19, a BVI participant, shared, "*In the past, I might hit the wrong address. With AI, I'd want to input the correct address using voice.*" Users also expected personalized and contextualized results. P7, in a shopping context, remarked, "*The tool might suggest new choices I wasn't aware of, which could better meet my current needs.*" The potential disadvantages align with participants' concerns, discussed in subsequent sections.

4.2.1 Automation preferences vary by task characteristics. Tasks with subjective decision-making showed significantly higher automation scores ($M = 3.61$, $SD = 1.91$) compared to tasks with some subjective decisions ($M = 2.75$, $SD = 1.73$; $H(2) = 11.73$, $p = .003$). While participants were open to gathering more information and suggestions from AI, they preferred to make final decisions by themselves. For example, P2, while considering a mountain bike purchase, requested "*pros and cons mentioned in the reviews*" yet wanted to "*confirm the final purchase*". Similarly, P5, using the automation tool for job applications, asked it to "*fill out the application form automatically*" but insisted on "*final confirmation*" before submitting. Participants valued comprehensive and tailored information while retaining ultimate control.

People would like to automate simple and repetitive tasks. To understand the complexity of the tasks, we labeled rounds of interaction for each task to indicate the steps involved in automating the task with AI. Tasks with one round of interaction showed significantly lower automation scores ($M = 1.96$, $SD = 1.21$) compared to tasks with multiple rounds ($M = 2.83$, $SD = 1.53$), indicating a preference for more automation in simpler tasks ($U = 4475$, $p < .001$). A Kruskal-Wallis test revealed a significant effect of task repetition pattern on automation preference ($H(2) = 10.44$, $p = .005$). Post-hoc Mann-Whitney U tests ($U = 9192.5$, $p = .001$) showed that tasks repeated at fixed intervals ($M = 2.49$, $SD = 1.76$) were preferred to be more automated compared to tasks with unpredictable intervals ($M = 3.30$, $SD = 1.88$). As P9 explained: "*[The tool could] duplicate my previous weekly timesheet automatically based on my previous weekly timesheet then ready for submit, which would help the project manager get it earlier since I tend to procrastinate when doing it manually.*"

4.3 Errors are the Largest Concern

Error concerns emerge as the most significant issue in web automation tools ($M = 2.24$, $SD = 1.47$), manifesting in three main aspects: impact severity, domain-specific concerns, and trust issues. Regarding impact severity, participants were particularly concerned about irreversible consequences. P7 worried that "*I might not be able to recover from wrong information*," while P1 emphasized that "*if anything goes wrong, it will affect my finances or miss the perfect time.*" Domain-specific concerns varied by context, with P6 highlighting "*AI accuracy issues in getting a tutorial*" for educational content, and P9 expressing worry about "*incorrect data exchange with the state agency*" for financial tasks. Trust and confidence issues centered around AI's limitations and reliability. P12 directly stated "*I don't have much trust in AI tools like ChatGPT*," while P1 worried about "*AI's data sources may be inaccurate.*" P20 noted that "*the tool may*

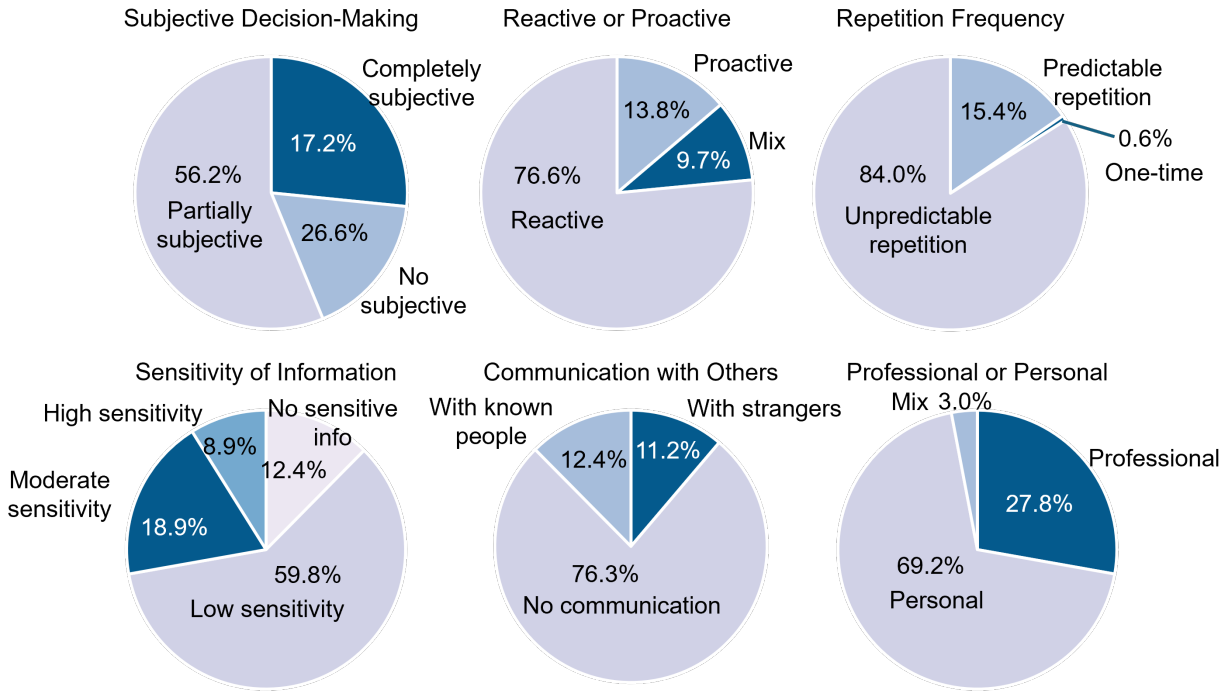

Figure 2: Distribution of 150 user-defined tasks across categories such as subjectivity, repetition pattern, and sensitivity level. These categories were used in our mixed-method analysis to examine how task characteristics relate to automation preferences and user concerns. See Table 5 for more detail on the categories listed.

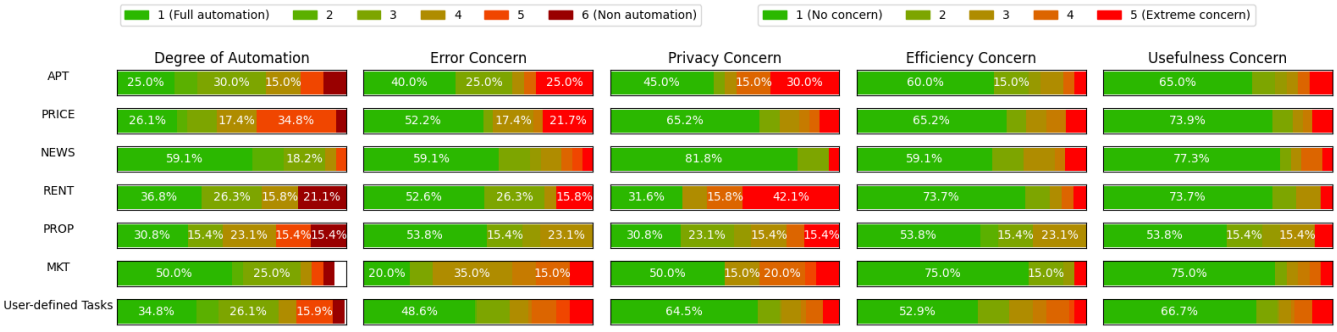

Figure 3: Distribution of preferred automation levels and reported concerns (error, privacy, efficiency, usefulness) for both predefined and user-defined tasks. See Table 2 for details of the 'Example' tasks.

not work very well if I don't provide my personal information in details," highlighting concerns about AI's ability to function reliably without complete information.

4.3.1 People have higher error concerns in professional tasks compared to personal tasks. Professional tasks, such as searching for learning tutorials, generating code, and filling out job applications, often carry higher stakes and potential consequences. Participants described these tasks as "*critical to success*" (P5) and "*important because errors in them could result in tangible negative outcomes*" (P7) because they can directly impact their career development or job prospects. For instance, P4 elaborated that they "*do not want to be misguided*," as errors in these tasks could lead to misinformation, lost opportunities, or even job rejections. In contrast, personal tasks are perceived as having fewer immediate consequences, which explains the lower error concerns. This disparity highlights the need for more reliable and accurate AI systems when supporting professional activities, where users expect precision and trustworthiness.

4.4 Personal data raises privacy concerns

Many participants expressed concerns about sharing personal identifiable information and corporate data through third-party tools.

These concerns are particularly salient in web contexts, where users routinely interact with sensitive platforms (e.g., healthcare portals, government sites) and may be reluctant to authorize third-party automators—even when AI assistance would be beneficial. We classified data sensitivity into four levels: level 0 (no sensitive data), level 1 (mild sensitivity like user behavior), level 2 (moderate sensitivity like login credentials), and level 3 (high sensitivity like financial details). A Kruskal-Wallis test showed privacy concerns increase with data sensitivity ($H(3) = 12.47$, $p = .006$), with a moderate positive correlation ($r_s = .37$, $p < .001$).

Users were less concerned about low-sensitivity data already accessible on existing platforms. P20 noted about social media: "*any random person could access my social media profile.*" However, for sensitive data, participants preferred trusted institutions over third-party tools. P6 emphasized: "*Uploading corporate data to third-party AI tools might lead to privacy issues that could jeopardize both the company's interests and my job,*" while P5 explained: "*If the tool is developed by the bank itself, I would trust it... However, if the tool comes from a third party, I worry about privacy.*" Survey results indicate users are somewhat willing to trade privacy for automation functionality (M = 3.46 on a 7-point scale).

4.5 Specific needs from Elderly/BVI individuals

Elderly individuals and those with visual impairments often face challenges when interacting with AI-assisted technologies [12, 21, 23]. Our study revealed distinct patterns in their automation preferences and needs.

4.5.1 Elderly/BVI individuals prefer more automation. Elderly/BVI users showed stronger preference for automation (M = 2.65, SD = 1.797) compared to others (M = 3.21, SD = 1.892; Mann-Whitney $U = 9245.0$, $p = .007$). While these users expressed limited technical understanding (P19: "*I don't understand too much about how AI works*"), they prioritized practical usability over technical concerns. P18 highlighted functional accessibility issues: "*some of the software...doesn't always perform as what it says,*" while P22 emphasized that their main concern was "*do I get the things I want, the right information I need to hear?*" These findings suggest the need for automation tools that prioritize straightforward usability and practical functionality.

4.5.2 Elderly adults prefer voice interactions. Prior research shows elderly users prefer voice assistants [27], which our study confirmed. Participants found typing challenging (P1) and preferred voice interaction for its convenience and naturalness. P24 noted: "*it would save me a lot of time, and it would basically be an easier way of doing things more than manually.*" Voice interaction also proved useful for situational accessibility needs, as P9 explained: "*I would ask a smart speaker to do this while I was doing the dishes or commuting through voice.*"

4.5.3 BVI individuals need extra assistance in extracting useful information from the Web. BVI participants face challenges with complex web interfaces and screen readers. P18 noted: "*Pick out the simplest pieces of information on the webpage, because when you're using a speech system, it's sometimes difficult to find what you actually need.*" P23 highlighted screen reader limitations: "*Current screen readers read every word on the page, including ads. When a website refreshes dynamically, the screen reader re-reads the ads.*" P21 wanted better image recognition: "*capture scenes in more detail, without needing assistance from a real person,*" while P24 suggested reducing user input through active detection and feedback.

5 Discussion and Implications

Our findings highlight user needs that go beyond accuracy or efficiency. Participants want automation tools that align with their values—control, transparency, and adaptability—and fit naturally into their workflows. These needs are not well represented in current automation benchmarks, which assume that automation should aim for full autonomy.

5.1 Flexible Automation Design: From User Needs to Benchmarks

Previous systems focused mainly on information scraping, while our participants sought to automate diverse tasks from scheduling to decision-making. Our study revealed users prefer partial over full automation, especially for tasks involving subjective decisions or interpersonal communication. This contrasts with current benchmarks [8, 14, 49, 60, 66] which assume full automation environments, indicating a need to redesign both systems and evaluation frameworks.

We propose three improvements: (1) Flexible interaction modes allowing users to choose between full automation and guided assistance, (2) Benchmarks that evaluate human-AI collaboration points rather than just next actions, and (3) Expanded benchmark coverage, particularly for tasks involving sensitive personal data.

5.2 Mitigating Privacy and Error Concerns

For errors, participants emphasized robust detection mechanisms and easy correction tools, especially for high-stakes tasks. Privacy concerns focused on data handling with third-party services. Users want transparency in data usage and storage, particularly for sensitive information. Future systems should include error recovery mechanisms and clear privacy controls, while prioritizing partnerships with trusted institutions.

5.3 Inclusive Design for Elderly and BVI users

For elderly users, automation interfaces should prioritize voice interaction, with natural language processing for conversational commands and verbal responses. For BVI users, tools need improved screen reader integration that intelligently filters and prioritizes task-relevant content while handling dynamic web elements effectively. The system should reduce cognitive load from auxiliary content and provide customizable content summarization, making web automation more accessible for both user groups' distinct needs.

6 Conclusion

Through interviews with 24 diverse participants, we examined human-centric needs and preferences in AI-assisted web automation systems. This paper contributed empirical evidence for designing user-centric web automation systems that balance control, error handling, privacy, and accessibility needs across diverse user groups.

References

[1] [n. d.]. Adept: Useful General Intelligence — adept.ai. https://www.adept.ai/. [Accessed 19-04-2024].

[2] [n. d.]. Taxy AI — taxy.ai. https://taxy.ai/. [Accessed 19-04-2024].

[3] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, et al. 2019. Guidelines for human-AI interaction. In Proceedings of the 2019 chi conference on human factors in computing systems. 1–13.

[4] Axiom.ai. 2024. No-Code Browser Automation. https://axiom.ai/. Accessed: September 2024.

[5] Ann Blandford, Dominic Furniss, and Stephann Makri. 2016. Analysing Data. In Qualitative HCI Research. Springer, Cham, 51–60. https://doi.org/10.1007/978-3-031-02217-3_5

[6] Ann Blandford, Dominic Furniss, and Stephann Makri. 2016. Planning a Study. In Qualitative HCI Research. Springer, Cham, 7–21. https://doi.org/10.1007/978-3-031-02217-3_2

[7] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258 (2021).

[8] Rogerio Bonatti, Dan Zhao, Francesco Bonacci, Dillon Dupont, Sara Abdali, Yinheng Li, Yadong Lu, Justin Wagle, Kazuhito Koishida, Arthur Bucker, et al. 2024. Windows agent arena: Evaluating multi-modal os agents at scale. arXiv preprint arXiv:2409.08264 (2024).

[9] Robin N Brewer and Anne Marie Piper. 2017. xPress: Rethinking design for aging and accessibility through an IVR blogging system. Proceedings of the ACM on Human-Computer Interaction 1, CSCW (2017), 1–17.

[10] Tara Capel and Margot Brereton. 2023. What is human-centered about human-centered AI? A map of the research landscape. In Proceedings of the 2023 CHI conference on human factors in computing systems. 1–23.

[11] Sarah E Chasins, Maria Mueller, and Rastislav Bodik. 2018. Rousillon: Scraping distributed hierarchical web data. In Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology. 963–975.

[12] Khansa Chemnad and Achraf Othman. 2024. Digital accessibility in the era of artificial intelligence—Bibliometric analysis and systematic review. Frontiers in Artificial Intelligence 7 (2024), 1349668.

[13] Weihao Chen, Xiaoyu Liu, Jiacheng Zhang, Ian Iong Lam, Zhicheng Huang, Rui Dong, Xinyu Wang, and Tianyi Zhang. 2023. MIWA: Mixed-Initiative Web Automation for Better User Control and Confidence. In Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology. 1–15.

[14] Xiang Deng, Yu Gu, Boyuan Zheng, Shijie Chen, Sam Stevens, Boshi Wang, Huan Sun, and Yu Su. 2024. Mind2web: Towards a generalist agent for the web. Advances in Neural Information Processing Systems 36 (2024).

[15] Rui Dong, Zhicheng Huang, Ian Iong Lam, Yan Chen, and Xinyu Wang. 2022. WebRobot: web robotic process automation using interactive programming-by-demonstration. In Proceedings of the 43rd ACM SIGPLAN International Conference on Programming Language Design and Implementation. 152–167.

[16] Anthony Dunne and Fiona Raby. 2024. Speculative Everything, With a new preface by the authors: Design, Fiction, and Social Dreaming. MIT press.

[17] Luciano Floridi and Massimo Chiriatti. 2020. GPT-3: Its nature, scope, limits, and consequences. Minds and Machines 30 (2020), 681–694.

[18] Fiona Fui-Hoon Nah, Ruilin Zheng, Jingyuan Cai, Keng Siau, and Langtao Chen. 2023. Generative AI and ChatGPT: Applications, challenges, and AI-human collaboration. , 277–304 pages.

[19] Hiroki Furuta, Ofir Nachum, Kuang-Huei Lee, Yutaka Matsuo, Shixiang Shane Gu, and Izzeddin Gur. 2023. Multimodal web navigation with instruction-finetuned foundation models. arXiv preprint arXiv:2305.11854 (2023).

[20] Stephen Grossberg. 2020. A path toward explainable AI and autonomous adaptive intelligence: deep learning, adaptive resonance, and models of perception, emotion, and action. Frontiers in neurorobotics 14 (2020), 533355.

[21] Jie Gu, Xiaolun Wang, Xinlin Yao, and Anan Hu. 2020. Understanding the influence of AI voice technology on visually impaired elders' psychological well-being: An affordance perspective. In Human Aspects of IT for the Aged Population. Technology and Society: 6th International Conference, ITAP 2020, Held as Part of the 22nd HCI International Conference, HCII 2020, Copenhagen, Denmark, July 19–24, 2020, Proceedings, Part III 22. Springer, 226–240.

[22] Izzeddin Gur, Ofir Nachum, Yingjie Miao, Mustafa Safdari, Austin Huang, Aakanksha Chowdhery, Sharan Narang, Noah Fiedel, and Aleksandra Faust. 2022. Understanding html with large language models. arXiv preprint arXiv:2210.03945 (2022).

[23] Sara Hamideh Kerdar, Liane Bächler, and Britta Marleen Kirchhoff. 2024. The accessibility of digital technologies for people with visual impairment and blindness: a scoping review. Discover Computing 27, 1 (2024), 24.

[24] Faria Huq, Jeffrey P Bigham, and Nikolas Martelaro. 2023. " What's important here?": Opportunities and Challenges of Using LLMs in Retrieving Information from Web Interfaces. arXiv preprint arXiv:2312.06147 (2023).

[25] David B Kaber and Jennifer M Riley. 1999. Adaptive automation of a dynamic control task based on secondary task workload measurement. International journal of cognitive ergonomics 3, 3 (1999), 169–187.

[26] Dae Hyun Kim, Hyungyu Shin, Shakhnozakhon Yadgarova, Jinho Son, Hariharan Subramonyam, and Juho Kim. 2024. AINeedsPlanner: A Workbook to Support Effective Collaboration Between AI Experts and Clients. In Proceedings of the 2024 ACM Designing Interactive Systems Conference. 728–742.

[27] Sunyoung Kim and Abhishek Choudhury. 2021. Exploring older adults' perception and use of smart speaker-based voice assistants: A longitudinal study. Computers in Human Behavior 124 (2021), 106914.

[28] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. Advances in neural information processing systems 35 (2022), 22199–22213.

[29] Rebecca Krosnick and Steve Oney. 2021. Understanding the challenges and needs of programmers writing web automation scripts. In 2021 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC). IEEE, 1–9.

[30] Rebecca Krosnick and Steve Oney. 2022. ParamMacros: Creating UI Automation Leveraging End-User Natural Language Parameterization. In 2022 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC). IEEE, 1–10.

[31] Todd Kulesza, Simone Stumpf, Margaret Burnett, and Irwin Kwan. 2012. Tell me more? The effects of mental model soundness on personalizing an intelligent agent. In Proceedings of the sigchi conference on human factors in computing systems. 1–10.

[32] Kenton Lee, Mandar Joshi, Iulia Raluca Turc, Hexiang Hu, Fangyu Liu, Julian Martin Eisenschlos, Urvashi Khandelwal, Peter Shaw, Ming-Wei Chang, and Kristina Toutanova. 2023. Pix2struct: Screenshot parsing as pretraining for visual language understanding. In International Conference on Machine Learning. PMLR, 18893–18912.

[33] Gilly Leshed, Eben M Haber, Tara Matthews, and Tessa Lau. 2008. CoScripter: automating & sharing how-to knowledge in the enterprise. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. 1719–1728.

[34] Toby Jia-Jun Li, Amos Azaria, and Brad A Myers. 2017. SUGILITE: creating multimodal smartphone automation by demonstration. In Proceedings of the 2017 CHI conference on human factors in computing systems. 6038–6049.

[35] Toby Jia-Jun Li, Igor Labutov, Xiaohan Nancy Li, Xiaoyi Zhang, Wenze Shi, Wanling Ding, Tom M Mitchell, and Brad A Myers. 2018. Appinite: A multi-modal interface for specifying data descriptions in programming by demonstration using natural language instructions. In 2018 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC). IEEE, 105–114.

[36] Toby Jia-Jun Li, Marissa Radensky, Justin Jia, Kirielle Singarajah, Tom M Mitchell, and Brad A Myers. 2019. Pumice: A multi-modal agent that learns concepts and conditionals from natural language and demonstrations. In Proceedings of the 32nd annual ACM symposium on user interface software and technology. 577–589.

[37] Yang Li, Gang Li, Luheng He, Jingjie Zheng, Hong Li, and Zhiwei Guan. 2020. Widget captioning: Generating natural language description for mobile user interface elements. arXiv preprint arXiv:2010.04295 (2020).

[38] Sebastian Linxen, Christian Sturm, Florian Brühlmann, Vincent Cassau, Klaus Opwis, and Katharina Reinecke. 2021. How weird is CHI?. In Proceedings of the 2021 chi conference on human factors in computing systems. 1–14.

[39] Thomas Markussen and Eva Knutz. 2013. The poetics of design fiction. In Proceedings of the 6th International Conference on Designing Pleasurable Products and Interfaces. 231–240.

[40] Microsoft. 2023. Reinventing Search with a New AI-Powered Microsoft Bing and Edge, Your Copilot for the Web. https://blogs.microsoft.com/blog/2023/02/07/reinventing-search-with-a-new-ai-powered-microsoft-bing-and-edge-your-copilot-for-the-web/. Accessed: September 2024.

[41] Meredith Ringel Morris, Jascha Sohl-dickstein, Noah Fiedel, Tris Warkentin, Allan Dafoe, Aleksandra Faust, Clement Farabet, and Shane Legg. 2023. Levels of AGI: Operationalizing Progress on the Path to AGI. arXiv preprint arXiv:2311.02462 (2023).

[42] OpenAI. 2025. Introducing Operator. https://openai.com/index/introducing-operator/. Accessed: 2025-01-23.

[43] Ottogrid.ai. 2024. AI-Powered Automation for Manual Research. https://ottogrid.ai/. Accessed: September 2024.

[44] R. Parasuraman, T.B. Sheridan, and C.D. Wickens. 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 (May 2000), 286–297. https://doi.org/10.1109/3468.844354

[45] Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology. 1–22.

[46] Kevin Pu, Rainey Fu, Rui Dong, Xinyu Wang, Yan Chen, and Tovi Grossman. 2022. Semanticon: Specifying content-based semantic conditions for web automation programs. In Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology. 1–16.

[47] Kevin Pu, Jim Yang, Angel Yuan, Minyi Ma, Rui Dong, Xinyu Wang, Yan Chen, and Tovi Grossman. 2023. DiLogics: Creating Web Automation Programs with Diverse Logics. In Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology. 1–15.
[48] Rabbit Research Team. 2023. Learning Human Actions on Computer Applications. https://www.rabbit.tech/research. Accessed: September 2024.
[49] Christopher Rawles, Sarah Clinckemaillie, Yifan Chang, Jonathan Waltz, Gabrielle Lau, Marybeth Fair, Alice Li, William Bishop, Wei Li, Folawiyo Campbell-Ajala, et al. 2024. AndroidWorld: A dynamic benchmarking environment for autonomous agents. arXiv preprint arXiv:2405.14573 (2024).
[50] Gavriel Salomon. 1988. AI in reverse: Computer tools that turn cognitive. Journal of educational computing research 4, 2 (1988), 123–139.
[51] Peter Shaw, Mandar Joshi, James Cohan, Jonathan Berant, Panupong Pasupat, Hexiang Hu, Urvashi Khandelwal, Kenton Lee, and Kristina N Toutanova. 2024. From pixels to ui actions: Learning to follow instructions via graphical user interfaces. Advances in Neural Information Processing Systems 36 (2024).
[52] Ben Shneiderman. 2020. Bridging the gap between ethics and practice: guidelines for reliable, safe, and trustworthy human-centered AI systems. ACM Transactions on Interactive Intelligent Systems (TiiS) 10, 4 (2020), 1–31.
[53] Anselm Strauss and Juliet Corbin. 1998. Basics of qualitative research techniques. (1998).
[54] Hari Subramonyam, Divy Thakkar, Jürgen Dieber, and Anoop Sinha. 2024. Content-Centric Prototyping of Generative AI Applications: Emerging Approaches and Challenges in Collaborative Software Teams. arXiv preprint arXiv:2402.17721 (2024).
[55] Mei Tan, Hansol Lee, Dakuo Wang, and Hari Subramonyam. 2024. Is a seat at the table enough? Engaging teachers and students in dataset specification for ml in education. Proceedings of the ACM on Human-Computer Interaction 8, CSCW1 (2024), 1–32.
[56] Taxy AI. 2024. Automate Your Browser with GPT-4. https://taxy.ai/. Accessed: September 2024.
[57] Marialena Vagia, Aksel A Transeth, and Sigurd A Fjerdingen. 2016. A literature review on the levels of automation during the years. What are the different taxonomies that have been proposed? Applied ergonomics 53 (2016), 190–202.
[58] Bryan Wang, Gang Li, and Yang Li. 2023. Enabling conversational interaction with mobile ui using large language models. In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. 1–17.
[59] Hao Wen, Yuanchun Li, Guohong Liu, Shanhui Zhao, Tao Yu, Toby Jia-Jun Li, Shiqi Jiang, Yunhao Liu, Yaqin Zhang, and Yunxin Liu. 2023. Empowering llm to use smartphone for intelligent task automation. arXiv preprint arXiv:2308.15272 (2023).
[60] Tianbao Xie, Danyang Zhang, Jixuan Chen, Xiaochuan Li, Siheng Zhao, Ruisheng Cao, Toh Jing Hua, Zhoujun Cheng, Dongchan Shin, Fangyu Lei, et al. 2024. Osworld: Benchmarking multimodal agents for open-ended tasks in real computer environments. arXiv preprint arXiv:2404.07972 (2024).
[61] Wei Xu, Marvin J Dainoff, Liezhong Ge, and Zaifeng Gao. 2023. Transitioning to human interaction with AI systems: New challenges and opportunities for HCI professionals to enable human-centered AI. International Journal of Human–Computer Interaction 39, 3 (2023), 494–518.
[62] Qian Yang, Aaron Steinfeld, Carolyn Rosé, and John Zimmerman. 2020. Re-examining whether, why, and how human-AI interaction is uniquely difficult to design. In Proceedings of the 2020 chi conference on human factors in computing systems. 1–13.
[63] Tom Yeh, Tsung-Hsiang Chang, and Robert C Miller. 2009. Sikuli: using GUI screenshots for search and automation. In Proceedings of the 22nd annual ACM symposium on User interface software and technology. 183–192.
[64] Xiaoyi Zhang, Lilian De Greef, Amanda Swearngin, Samuel White, Kyle Murray, Lisa Yu, Qi Shan, Jeffrey Nichols, Jason Wu, Chris Fleizach, et al. 2021. Screen recognition: Creating accessibility metadata for mobile applications from pixels. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. 1–15.
[65] Boyuan Zheng, Boyu Gou, Jihyung Kil, Huan Sun, and Yu Su. 2024. Gpt-4v (ision) is a generalist web agent, if grounded. arXiv preprint arXiv:2401.01614 (2024).
[66] Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, et al. 2023. Webarena: A realistic web environment for building autonomous agents. arXiv preprint arXiv:2307.13854 (2023).

A Predefined Tasks

We provided six predefined tasks for the users listed in Table 2. We map the predefined tasks to categories from Table 3, indicating that predefined tasks cover a broad range of real-world conditions.

B Contextualize Automation Level in Web

We use Table 4 to contextualize six web automation levels in user control and feedback.

C Categorization of User Preferences and Interactions

We use Table 5 to detail the categorization of user preferences and interactions.

Example	Task Description
1 (APT)	Schedule an appointment for a car knowledge test at the nearest DMV, requiring your name.
2 (PRICE)	Compare features, prices, and user ratings of smart home devices across brands and online stores.
3 (NEWS)	Summarize today's news on the New York Times.
4 (RENT)	Pay rent monthly with your credit card, requiring banking information.
5 (PROP)	Upload a business proposal for a new AI startup to an AI tool for refinement, risking proprietary information exposure.
6 (MKT)	Use an AI tool to generate a report on market trends by the end of the day, with limited time for verification.

Table 2: Predefined Tasks

Task	Decision-Making	Reactive/Proactive	Repetition	Sensitivity	Context
APT	Medium	Reactive	One-time	Somewhat	Personal
PRICE	High	Reactive	Non-predictable	Mild	Personal
NEWS	Medium	Reactive	Predictable	Mild	Personal
RENT	Low	Proactive	Predictable	High	Personal
PROP	High	Reactive	Non-predictable	Somewhat	Professional
MKT	Medium	Reactive	Non-predictable	High	Professional

Table 3: Mapping Tasks to Categories from Table 5

Automation Level	Description	User Control	Feedback Mechanism
Level 1: Fully Automated (No Feedback)	System completes tasks independently without user input or feedback once initiated.	None	None
Level 2: AI Decides When to Continue/Stop	AI autonomously evaluates and decides to proceed or stop based on pre-set parameters.	Minimal	Limited, only final results shared with users
Level 3: Step-by-Step Automation	Task advances in steps; user can optionally provide feedback at each step.	Moderate	Optional user feedback at each stage
Level 4: Mandatory Step-by-Step Feedback	User confirmation required at each step; automation only proceeds with explicit approval.	High	Mandatory feedback at each step
Level 5: Multiple Options at Each Step	Automation offers choices at each step; user selects preferred option to continue.	Very High	User-driven choice selection at every step
Level 6: Non-Automation	User performs all tasks manually without automation assistance.	Full	None

Table 4: Contextualizing the Six Levels of Web Automation in User Control and Feedback

Category	Description	User Examples
Subjectivity	No subjective decisions involved	Pay my bills online using different portals (P3)
	Subjective decisions involved	Search and watch coding tutorials on Youtube (P6)
	Completely dependent on subjectivity	Content generation for blog posts (P17)
Reactive or Proactive	Reactive—The user instructs the AI	Search figures(person) and check bibliography (P10)
	Proactive—The AI notifies user	Receive notifications for upcoming deadlines (P12)
	Mix — The user wants both interactions	Receive notifications for refilling prescriptions and help me refill the prescription after my confirmation. (P9)
Repetition	Would only need to run one time	Look for a cooking recipe (P2)
	Runs repeatedly but not predictable	Check social media updates (P5)
	Runs at a predictable interval	Pay online bills monthly (P11)
Sensitive Info Level	No sensitive information	Search for technical materials, such as papers (P2)
	Mildly sensitive information	Get personalized shopping recommendations (P1)
	Some sensitive information	Save login information for e-commerce websites (P10)
	Somewhat sensitive information	Store credit card details for quick transactions (P13)
Communication	No communication with other people	Self-study with specific topics(such as NLP) (P4)
	Involves communication with strangers	Participate in an anonymous online survey (P15)
	communicating with acquaintances	Collaborate on Google Docs with team members (P12)
Professional Level	Professional	Searching for online resources for class notes (P13)
	Personal	Plan a family vacation itinerary (P20)
	Mix	Check emails and organize a shared calendar for work and personal appointments (P7)

Table 5: Categorization of User Preferences and Interactions