Towards building a comprehensive evaluation framework for Human Computer Interaction with a specific focus on AutoML

Sundaraparipurnan Narayanan, AI Tech Ethics, sundar.narayanan@aitechethics.com

Abstract

The rapid advancement of Automated Machine Learning (AutoML) has revolutionized the field of Artificial Intelligence (AI), enabling the automation of complex machine learning workflows. However, the increasing autonomy of AutoML systems has highlighted the critical importance of Human-Computer Interaction (HCI) principles in ensuring their usability, transparency, and trustworthiness. This paper proposes a comprehensive evaluation framework for HCI in AutoML, focusing on five key dimensions: Contracts and User Development, User Interface, Interaction and Experience Design, Information Architecture, Human Augmentation Factors Design, and Care and Responsibility. The framework emphasizes the need for transparent disclosures, intuitive interfaces, structured information presentation, user empowerment, and ethical considerations, including fairness and accountability. By prioritizing user-centered design, the framework aims to bridge the gap between AutoML's technical capabilities and diverse user needs, fostering trust and collaboration between humans and AI systems. The integration of HCI principles is crucial for realizing the full potential of AutoML, ensuring that these powerful tools are developed and deployed in a manner that benefits all users and society. As AutoML continues to evolve towards greater autonomy, the nature of human interaction is expected to transform from direct operation to strategic supervision and collaborative partnership, leveraging the complementary strengths of humans and machines. The proposed framework serves as a foundation for creating transparent, user-friendly, and trustworthy AutoML systems that balance automation with user understanding and control, paving the way for a future where AI systems are not merely efficient but also equitable, safe, and truly beneficial for all.

CCS CONCEPTS • Human-centered computing ~ Human computer interaction (HCI) ~ HCI design and evaluation methods ~ Walkthrough evaluations

Additional Keywords and Phrases: Human Computer Interaction, AutoML Evaluation

Introduction

Human–computer Interaction (HCI) is a multidisciplinary field that combines knowledge from computer science, psychology, cognitive science, and social sciences to create interactive computing systems that are both useful and usable (*Human-Computer Interaction*, 2003) (Olson & Olson, 2002). It focuses on designing, evaluating, and implementing interfaces that facilitate effective communication between users and computers, with the goal of improving user experience and system efficiency (Fallman, 2007) (Kheder, 2023). HCI has given rise to numerous sub-disciplines, including Human Computation (HCOMP), human-AI collaboration (HAI), human-robot interaction (HRI), and HITL (Human-in-the-loop) (Lakkshmanan et al., 2024).HCI research has evolved from its initial focus on functionality to encompass user- friendliness, learnability, efficiency, enjoyment, and emotional aspects of interaction (Kheder, 2023). This shift has led to the development of various methodologies and approaches, such as user-centered design (UCD), usability testing, and prototyping techniques (Kheder, 2023). Additionally, emerging technologies such as augmented reality, virtual reality, and gesture- based interfaces have expanded the scope of HCI research (Kheder, 2023) (Kosch et al., 2023).

In the context of AutoML, HCI focuses on understanding the 'how' and 'why' of human interaction within these frameworks, which is crucial for optimal system design and identifying both opportunities and risks presented by increasing machine autonomy (Khuat et al., 2022).

Automated Machine Learning (AutoML) automates aspects of the ML application workflow. Initially, HCI was not

considered fundamental to AutoML, which primarily aims to operate with minimal human interaction. Currently, however, there is a shift back towards incorporating HCI as a necessity to support technical users who configure and control semi-AutoML packages, including interactions involving basic operations such as selection, value entry, exploration, and reconfiguration, with visualization naturally enhancing understanding and engagement. Key challenges in the present include building trust, ensuring explainability, mitigating bias for fairness, and increasing transparency, supported by emerging tools such as Data Cards for dataset documentation. The future trajectory points towards greater AutoML autonomy (AutonoML), shifting human roles from direct operation to supervision, auditing, and ultimately collaboration with the system as a partner. Designing for this future collaboration, guided by frameworks such as Human-Centered AI (HCAI), requires optimizing interactions for shared understanding, trust, and leveraging complementary human and machine strengths (Khuat et al., 2022).

AutoML aims to make machine learning accessible to non-experts and improve efficiency (Karmaker ("Santu") et al., 2021). Despite the aim of full automation, AutoML systems still require human intervention to be practically applicable (Crisan & Fiore-Gartland, 2021). Humans are involved in various stages of the ML workflow, providing inputs, feedback, or oversight (Mathewson, 2019) (Khuat et al., 2022). Human involvement in critical steps of AutoML includes understanding domain-specific data, defining prediction problems, and creating suitable training datasets (Karmaker ("Santu") et al., 2021). This human-machine interaction is crucial, yet current AutoML systems often lack transparency, making it difficult for users to understand and trust the decision-making process. Many current AutoML tools have become black-box systems, obscuring their internal working. This development highlights the need for further research into human–computer interaction (HCI) to address this weakness. Hence, the fields of AI and HCI share common roots, particularly in early work on conversational agents (Li et al., 2020). Recent advancements in deep learning have revolutionized AI, creating new opportunities for machines and humans to interact.

The field is increasingly recognizing the importance of a human-centered approach to AutoML by recognizing the need to address user interaction, considering the diverse roles, expectations, and expertise of humans involved (Lindauer et al., 2024). Human–computer interaction in AutoML is evolving towards factors including domain knowledge and context awareness, evaluation and interpretation, handling uncertainities and novelty, collaboration and oversight, interface design, interaction modalities, and visualization (Khuat et al., 2022). A human- centered paradigm promotes the collaborative design of ML systems that integrate the complementary strengths of human expertise and AutoML methodologies**,** partly triggered by an increasing awareness of the social and ethical (including trust, explainability, transparency, fairness, accountability, and causality) implications of ML technologies (Lindauer et al., 2024)(Khuat et al., 2022).

With the above, it is clear that trust and transparency have emerged as critical factors in HCI, particularly in the context of AI-enabled systems. Studies have shown that user trust is influenced by socio-ethical considerations, technical features, and user characteristics, highlighting the need for tailored approaches to system design (Bach et al., 2022).User-centric design remains a cornerstone of HCI, with frameworks such as the User-Centered Design Process (UCDP) prioritizing users' goals and characteristics throughout the design process (Kheder, 2023). Such an approach extends to explainable AI (XAI), where human-centered XAI focuses on addressing the distinct needs of non-expert end users, emphasizing usability, trust, and safety (Veitch & Alsos, 2021). Explainability and interpretability have gained prominence, particularly in the context of AI-powered systems. Social Transparency (ST) has been proposed as a sociotechnically informed perspective that incorporates socio-organizational context into explaining AI-mediated decision-making, potentially improving trust calibration and decision- making processes (Ehsan et al., 2021).Usability and human factor engineering continue to play crucial roles in HCI. Researchers have developed innovative frameworks that combine expert cognitive walkthroughs with user surveys to evaluate website UI/UX, thereby providing actionable insights for design improvements (Whaiduzzaman et al., 2023). Additionally, the integration of HCI principles into healthcare systems has shown promise in enhancing patient safety and optimizing processes (Mishra et al., 2023).The expanding scope of HCI encompasses governance, accountability, and risk management. As intelligent systems (including AutoML) have become more prevalent in high-stakes domains, there is a growing need to address moral and ethical concerns and develop transparency frameworks to enhance trust and acceptance (Vorm & Combs, 2022). This broader view of HCI emphasizes the importance of considering not only technical aspects but also the social, ethical,

and organizational implications of human- system interactions.

Framework for evaluation of Human Computer Interaction

Given the above context, this paper proposes a consistent framework approach towards establishing a consistent framework for human–computer interaction evaluation. The effective integration of human–computer interaction (HCI) principles is essential for the successful development and deployment of AI systems in general and AutoML systems in particular, particularly in enabling non-expert users to engage with complex AI. The framework contains five dimensions: (1) contracts and user development that aims at setting expectations, clarifying responsibilities and limitations, and enabling the user; (2) user interface, interaction, and experience design that enables intuitive, usable, and engaging interfaces; (3) information architecture that supports organizing, simplifying, and visualizing complexity for the user; (4) human augmentation factors design that empower users and enables control; and Care and Responsibility that enables ethical, safe, and accountable AI. The details of each dimension are provided below.

Contracts and user development

This dimension focuses on establishing a clear understanding and managing the expectations between the user and the AI system. It encompasses the design of transparent disclosures regarding AI system capabilities and limitations, fostering responsible data use through mechanisms such as models and data cards, and providing comprehensive user training and guidance. Furthermore, it defines user responsibilities, including their roles in system oversight and providing constructive feedback, which are vital for safety and continuous improvement. For AutoML, this dimension ensures that users understand what an automated system can or cannot do, promoting appropriate usage and managing the inherent complexities of machine learning.

Acceptable Uses and Limitations: HCI plays a role in clarifying what AI systems are capable of doing and, importantly, what they are *not* capable of doing, to set appropriate expectations for users (Amershi et al., 2019). Disclosures of system limitations are part of the important design principles. Understanding these limitations is also crucial for safety, especially in safety-critical domains (Raulf et al., 2023) (Retzlaff et al., 2024).

Evaluation criteria	General description	AutoML specific description	Fairness context
Acceptable Uses Disclosure	Verify whether the system clearly communicates its intended and acceptable uses to users.	Confirm that AutoML explicitly outlines the types of machine learning problems and deployment scenarios for which it is designed and recommended.	Validate that AutoML clearly communicates its appropriate use cases, particularly where fairness implications are critical, and discourages its use in contexts where fairness cannot be assured.
Misuse Prevention	Evaluate whether the system incorporates safeguards to prevent its use outside of its intended context.	Confirm that AutoML includes mechanisms to prevent users from applying automated models inappropriately or in contexts beyond their validated scope.	Ascertain that the system has express contractual constraints or built-in safeguards to prevent misuse of the tool for unfair or discriminatory activities.

Providing information regarding the use of user data for training: Ensuring visibility in ML models and datasets is a challenge that has received increasing attention (Pushkarna et al., 2022). While not always framed as a 'contract,' transparent design and the provision of documentation are key to clarifying information on data use and provenance. Tools such as Model Cards (Raulf et al., 2023) and Data Cards (Pushkarna et al., 2022) are emerging as non- technical measures to enhance the explainability and responsible deployment of intelligent systems by specifying relevant details regarding model training, intended usage, and documenting datasets.

Evaluation criteria	General description	AutoML specific description	Fairness context
Transparency & Disclosure	Clarity and accessibility of information about data usage for model training.	Supports user understand how their data will be used in training the autoML tool for performance	Clarify if the transparency and disclosure highlight fairness implications of the data use for model training

User training, guidance, and instructions for use: Responsible AI implementation requires a strong focus on user training, guidance, and clear instructions. (Li et al., 2024) emphasized that 'communication, education, and training for users' are pivotal for building trustworthy AI. Tailoring explanations to different users (experts vs. non-experts) based on their information needs, context, and domain knowledge is crucial (Raulf et al., 2023). Some studies have also explored teaching user strategies to interact effectively with systems that have limited capabilities (Chromik & Butz, 2021).

Evaluation criteria	General description	AutoML specific description	Fairness context
User training	Assess the quality, comprehensiveness, and delivery methods of training programs designed to educate users on system functionality and responsible use.	Evaluate how effectively users are trained to understand AutoML workflows, interpret automated model choices, and make informed decisions regarding model selection and deployment.	Examine whether training materials adequately educate users about potential algorithmic biases in AutoML outputs and provide methods for bias detection and mitigation.
Guidance and support	Measure the availability, accessibility, and relevance of ongoing guidance and support resources for users during system interaction	Assess the clarity of in-system guidance for navigating complex AutoML features, understanding system recommendations, and troubleshooting issues related to automated model building.	Determine whether guidance materials offer clear instructions on how to identify and address unfair outcomes or discriminatory impacts that might arise from AutoML-generated models.

Instructions of use	Evaluate the clarity, conciseness, and completeness of explicit instructions provided to users to operate the AI system effectively and safely.	Consider how well the instructions explain the implications of different AutoML configurations, the meaning of various metrics, and the steps for safely deploying automated models.	Examine whether instructions explicitly highlight scenarios Wherefairness onsiderations are paramount and guide users on steps to ensure equitable treatment across different demographic groups.
Tailoring Explanations	Examine the system's ability to adapt the level of detail and complexity of explanations to suit the diverse information needs and technical backgrounds of different user groups.	Assess how AutoML explanations are tailored for both expert data scientists and non-expert domain users, considering their varying understanding of machine learning concepts.	Evaluate if explanations about fairness issues are presented in a way that is understandable and actionable for all relevant user groups, regardless of their technical expertise

User responsibilities when using the tool: While explicit user contracts are rarely mentioned, some sources discuss human responsibilities in interactions, particularly in overseeing autonomous systems and ensuring safety (Khuat et al., 2022). GDPR implies user rights to contest outcomes, which requires a system design that supports such opportunities. Design choices that allow users to provide feedback also address implicit responsibilities in contributing to system improvement (Retzlaff et al., 2024) (Ashtana et al., 2018).

Evaluation criteria	General description	AutoML specific description	Fairness context
Oversight of autonomous systems	Assess clarity and support for users' active roles in monitoring, verifying, and ensuring the safe operation of AI systems.	Evaluate how users are guided to oversee automated model training, validation, and deployment processes to identify potential errors or unintended behaviors.	Examine whether users are informed about their responsibility to monitor AutoML outputs for signs of unfairness or bias and provide tools to do so.
Right to contest outcomes	Information regarding the system design and features that enable users to challenge, query, and seek rectification for AI-generated decisions or outputs.	Assesses how AutoML tools provide information or guidance enabling users to inspect automated model decisions, understand their rationale, and contest results that appear illogical or incorrect.	Determine if users are provided with information or guidance regarding clear mechanisms to contest AutoML decisions that might lead to discriminatory outcomes, and if the system supports investigation into such claims.

User Interface, Interaction and Experience Design

This dimension addresses the core sensory and interactive elements of user engagement with the AutoML system. It prioritizes the creation of intuitive Graphical User Interfaces (GUIs) that facilitate seamless interaction and ensure high usability for diverse user groups, including non- experts. Key considerations include visual explanation interfaces, hierarchical information presentation, and the integration of 'nudges' and clear metrics/visualizations to guide users through complex processes. The goal is to design an experience that is efficient, satisfying, and minimizes cognitive load, making sophisticated AutoML accessible and manageable.

UI/UX Design: HCI is fundamentally the design, evaluation, and implementation of interactive computing systems (Lakkshmanan et al., 2024). In AI/ML, this means focusing on the UI, interaction, and user experience (UX) (Ashtana et al., 2018) (Khuat et al., 2022) (Pop & Raţiu, 2024). For AutoML specifically, GUIs are common in industry, supporting interactive functions such as selection, exploration, reconfiguration, and value entry (Khuat et al., 2022). Designing for user experience is pivotal for the adoption and acceptance of ML technologies (Ashtana et al., 2018), specifically by non-experts.

Evaluation criteria	General description	AutoML specific description	Fairness context
User Interface (UI) Design	Assess the visual layout, elements, and overall aesthetics of the system's interactive components to ensure clarity and navigability.	Evaluate how the AutoML GUI supports interactive functions such as selection, exploration, reconfiguration, and value entry for various model parameters and datasets.	Examine whether the UI design visually highlights potential fairness metrics or biases in model outputs, making them easily discernible to the user.
User Experience (UX) Design	Measure the overall satisfaction, ease of use, and efficiency experienced by users when interacting with an AI system.	Consider how the UX design in AutoML facilitates the adoption and acceptance of ML technologies by non- experts, making complex processes feel approachable	Assess whether the UX design intuitively guides users to identify and address fairness concerns, ensuring a smooth and ethical model development process.
Interaction Design	Evaluate how users engage with the system through actions, feedback, and controls, ensuring intuitive and effective communication between humans and AI.	Determine if the interaction design in AutoML enables seamless exploration of different model architectures, parameter settings, and performance evaluations	Investigate whether interaction patterns allow users to easily compare fairness outcomes across different model versions or data subsets

Learnability	Assess how easy it is for new users to accomplish basic tasks and for experienced users to become proficient over time.	Evaluate whether the AutoML interface is designed for rapid learning, allowing users to quickly grasp automated processes and customize them effectively.	Consider whether the learnability of fairness-related features is high, enabling users to quickly understand how to assess and improve model fairness.

Usability: Usability is a key factor in the user experience of ML technologies. As AI applications become integral to our daily routines, the usability of these systems has become more important. From virtual assistants to autonomous vehicles, the ease with which users can interact and effectively utilize AI technologies is crucial. Usability in AI systems encompasses intuitive interfaces, clear instructions, and efficient task completion, ensuring that users can leverage these technologies without unnecessary complexity or confusion (Ashtana et al., 2018). Problems in system usability can arise from user interfaces (Mishra et al., 2023). Usability evaluation methods, such as formal usability evaluation complementing heuristic evaluation, are part of the iterative user-centered design process, including domain-specific environments (Rundo et al., 2020) (Acemyan & Kortum, 2012).

Evaluation criteria	General description	AutoML specific description	Fairness context
Intuitive Interfaces	Evaluate whether the system's interface is easy to understand and navigate without prior training or extensive documentation.	Assess whether the AutoML interface allows users to intuitively select algorithms, configure parameters, and interpret automated results.	Examine whether the interface clearly and intuitively presents fairness metrics and potential biases, making them easy to grasp.
Error Prevention and Recovery	Design the system to prevent common errors and provide clear, actionable guidance when errors do occur.	Implement mechanisms in AutoML to prevent misconfigurations or data input errors and offer clear steps for correction.	Provide safeguards within AutoML to prevent the creation of highly biased models and offer clear guidance on how to rectify fairness-related errors.
User Control and Freedom	Ensure that users have appropriate control over the system's functions and can easily undo actions or exit processes.	Allow users to easily pause, modify, or revert automated model-building processes and explore alternative configurations in AutoML.	Enable users to control and adjust fairness constraints or re-evaluate models based on different fairness criteria within the AutoML system.

Transparency: Transparency is a consistently reported key design principle for AI tools, and it remains a key element in enhancing human–computer interaction. Transparency is vital for building trust and is often linked to explainability and understandability (Hoque et al., 2024). Transparency can involve clear agent self-identification, disclosure of system limitations, and explanations. GDPR requirements also highlight the need for transparency in algorithmic

systems (Veale et al., 2018). However, many current AutoML tools obscure their internals, acting as black-box systems that hinder transparency (Khuat et al., 2022).

Evaluation criteria	General description	AutoML specific description	Fairness context
Internal process visibility	Make the internal workings and logic of the AI system accessible and comprehensible to users.	Provide visibility into the automated search space, the evaluation criteria used, and the intermediate steps taken by AutoML to arrive at a solution.	Examine if the interface clearly and intuitively presents fairness metrics and potential biases, making them easy to grasp.
Self-Identification	Clearly identify the AI system as an automated agent, distinguishing its actions from human interactions.	Explicitly state when AutoML is making automated decisions regarding model selection, hyperparameter tuning, or data transformations.	Examine if there is a clear communication as to how AutoML considers and reports on fairness during automated processes, rather than presenting it as a human decision.

Human Factors Engineering: Human factors engineering is a multidisciplinary field contributing to HCI. It considers human capabilities and limitations in system design to enhance performance, safety, and reliability (Pop & Raţiu, 2024). In the context of AI and HCI, human factors and cognitive science insights are crucial for designing effective human-AI interactions, particularly in safety-critical contexts (Raulf et al., 2023) (Chromik & Butz, 2021).

Evaluation criteria	General description	AutoML specific description	Fairness context
Human Capabilities and Limitations	Validate that the system design aligns with human cognitive, perceptual, and physical capabilities, mitigating human limitations.	Verify that AutoML tools accurately accommodate the cognitive load of understanding automated choices and limitations in interpreting complex ML metrics.	Confirm that AutoML interfaces present fairness information in a way that minimizes cognitive bias and supports accurate human assessment of equity.
Performance Enhancement	Verify that the interaction between human and AI effectively improves overall system efficiency, accuracy, and task completion	Confirm that AutoML effectively enhances user performance by automating repetitive tasks, accelerating model development, and suggesting optimal solutions.	Validate that users can efficiently analyze fairness trade-offs, enabling easier selection of models that balance performance with equitable outcomes.

Situational Awareness Support	Validate that users receive the necessary information to understand the current state of the AI system and its environment.	Confirm that AutoML provides clear dashboards and visualizations that accurately convey the progress of processes, model performance, and resource utilization.	Verify that fairness considerations are presented in a way that helps users maintain high situational awareness regarding the model's equitable performance across subgroups.

User Engagement: User engagement is a critical element in designing human-centric ML systems (Ashtana et al., 2018). Continuous user engagement and feedback loops are part of a user-centric ethical design paradigm that considers the need for ethics considerations. Iterative engagement, when supported by interfaces allowing feedback loops, domain knowledge injection, and model refinement, amplifies the human augmentation principle in AI-driven tools, and thereby, its quality, productivity, and trust.

Evaluation criteria	General description	AutoML specific description	Fairness context
Continuous Feedback Loops	Verify that the system provides accessible and intuitive mechanisms for users to offer ongoing feedback on functionality and performance.	Confirm that AutoML interfaces effectively support continuous feedback loops for users to comment on automated model suggestions, usability, or overall workflow.	Validate that users can easily provide feedback on perceived unfairness or bias in AutoML outputs, thus contributing to ethical system improvements.
Domain Knowledge Injection	Examine whether the system allows users to actively contribute their specific expertise and domain knowledge to refine AI processes or models.	Verify that AutoML tools facilitate the injection of domain knowledge by users to guide automated search processes or refine the generated model pipelines.	Confirm that users can effectively provide domain knowledge to highlight or address potential fairness issues specific to their application context or user groups.

Nudges: 'Nudges' guides users through structured processes to help them understand AI behavior (Buçinca et al., 2021) or suggesting new configurations for model refinement (Khuat et al., 2022). Consideration should be given to the cognitive effort and processes involved in users' interpretation of explanations. This includes the following:

On-screen Disclosures: Disclosures are mentioned as part of transparency, such as disclosing system capabilities and limitations (Amershi et al., 2019), and may include references to outcomes that the model/tool are not certain about.

Warnings, Notifications/Alerts: AI systems can have unpredictable behaviors. These sources imply the need for interfaces that handle such situations, although specific warnings or alerts are not detailed. In the context of autonomous driving, external HMI designs to convey messages and warnings to road users have been discussed (Chen, 2022).

Metrics/Visualization/Reports: Visualization via GUI components is a natural way to foster human understanding and interactivity in AutoML (Khuat et al., 2022). Interactive visualization is a key enabling technology for Human-Centered AI (HCAI) tools (Hoque et al., 2024), as it enhances comprehension, diagnosis, and iterative improvement of ML models. Furthermore, features to generate or download reports/models enhance the value of user engagement (Khuat et al., 2022).

Evaluation criteria	General description	AutoML specific description	Fairness context
Nudges	Validate that nudges effectively guide users through processes, aiding in understanding AI behavior and suggesting beneficial actions.	Verify that nudges in AutoML effectively guide users towards optimal configurations, explain automated decisions, or suggest pathways for model refinement.	Confirm that nudges effectively highlight potential fairness trade-offs or recommend adjustments to improve equitable outcomes across groups.
On-screen Disclosures	Confirm that disclosures clearly communicate system capabilities, limitations, and uncertainties directly within the user interface.	Verify that AutoML on-screen disclosures explicitly state automated system limitations or uncertainties regarding model performance, data quality, or search completeness.	Validate that disclosures clearly present any known limitations or uncertainties in AutoML's fairness evaluation or mitigation capabilities.

Warnings, Notifications/Alerts	Examine whether the system provides timely and clear warnings, notifications, or alerts for unpredictable behaviors or critical situations.	Check that AutoML effectively alerts users to potential issues such as data shifts, convergence failures, or unusual model behaviors during automated training.	Verify that the system generates explicit warnings or alerts when an AutoML-generated model or the loaded data exhibit statistically significant unfairness or bias.
Metrics/ Visualization/ Reports	Evaluate if metrics, interactive visualizations, and downloadable reports foster human understanding, diagnosis, and interaction with AI.	Confirm that AutoML provides intuitive metrics, interactive visualizations (e.g., performance curves, feature importance), and downloadable reports to enhance understanding and iterative model improvement.	Validate that visualizations and reports clearly present fairness metrics (e.g., disparate impact and equalized odds) and allow for easy comparison across different demographic groups or model versions.

Information Architecture

This dimension focuses on the structural organization and presentation of information within an AutoML system to enhance user comprehension and navigation. It dictates how complex data and processes are categorized, linked, and displayed. Key elements include well-structured navigation systems, logical content hierarchies (e.g., progressive disclosure to prevent information overload), and effective information visualization techniques. A well-designed information architecture ensures that users can easily find, understand, and interact with the relevant details of their AutoML tasks, thereby building mental models of the system's operation.

Navigation System: Effective navigation is implicit in discussions on UI design, exploration, and content hierarchy (Khuat et al., 2022). Designing intelligent UIs is critical for supporting human- guided AutoML (Khuat et al., 2022). Successful UIs are associated with a well-defined HCI structure based on principles and guidelines. The UI design should factor in human understanding and control. User interfaces must facilitate user selection, exploration, and reconfiguration actions.

Evaluation criteria	General description	AutoML specific description	Fairness context
Navigational Clarity	Verify that the navigation system provides a clear, logical, and consistent path for users to move through the AI application.	Confirm that the AutoML navigation system clearly guides users through various stages of automated model development, from data ingestion to deployment.	Validate that the navigation system allows users to easily find and access fairness-related settings, metrics, or bias reports within the AutoML tool.
Support for Exploration	Evaluate whether the navigation system effectively enables users to explore different functionalities, data subsets, or model options.	Check that AutoML navigation facilitates the exploration of different algorithms, hyperparameter spaces, or alternative automated pipeline suggestions.	Examine if the navigation system supports users in exploring fairness metrics across different sensitive attributes or model configurations to understand impacts.
Human Understanding and Control	Validate that the navigation system is designed to enhance human understanding of the system's state and facilitate user control over AI processes.	Check that AutoML's navigation intuitively maps to the underlying automated machine- learning process, giving users a sense of control and progress.	Confirm that the navigation system design helps users understand where and how they can exert control over the fairness aspects of automated model development.

Content Hierarchy: Content hierarchy is the organizational structure of information that can help users navigate complex AI systems more effectively. It involves prioritizing information based on importance and presenting it in layers, allowing users to focus on high-level concepts before delving into details and avoiding overwhelming the user with too much information at once (Yu, 2023).Progressive disclosure complements the content hierarchy by enabling users to access information gradually, as needed. Progressive disclosure is particularly valuable in the development of AI systems that require user trust and transparency, as it can alleviate concerns about information overload and enhance understanding through step-by-step information delivery (El Ali et al., 2024). Content hierarchy enables AI systems to be made more approachable and easier for users to engage with, thereby promoting transparency and trust (Xu et al., 2022) (Nazar et al., 2021) Content hierarchy and progressive disclosures are vital for fostering user trust and ensuring that the systems are adaptable and user-friendly (Bach et al., 2022).

Evaluation criteria	General description	AutoML specific description	Fairness context
Prioritization of Information	Verify that information is clearly prioritized based on importance, first presenting critical details and supporting deeper dives.	Confirm that AutoML clearly prioritizes essential information about automated model performance, key metrics, and recommended next steps for users.	Validate that fairness- related warnings, critical biases, or major fairness trade-offs are prioritized and displayed prominently to the user.
Layered Information Presentation	Evaluate whether complex information is organized into clear, digestible layers to prevent user overload.	Check that AutoML presents model details, data transformations, and automated search processes in digestible layers, thereby allowing users to delve progressively into complexity.	Examine whether fairness analyses are presented in layers, starting with high-level summaries and allowing users to drill down into group-specific metrics or bias explanations.

Information Visualization: Information visualization is fundamental to making complex AI processes understandable to humans (Le et al., 2020). Visualizations are crucial for understanding, diagnosing, and improving ML models (Retzlaff et al., 2024) (Khuat et al., 2022). Simple, familiar, and understandable visualizations are often preferred, especially for domain experts who are not visualization experts . Visualizations can bridge the gap between human knowledge and AI insights (Pop & Raţiu, 2024) (Hoque et al., 2024) and are particularly suitable for users with limited technical background.

Evaluation criteria	General description	AutoML specific description	Fairness context
Understandability and Simplicity	Verify that visualizations are simple, familiar, and easy for users to comprehend, thereby	Confirm that AutoML provides simple and familiar visualizations for understanding	Validate that fairness-related visualizations are presented simply
	avoiding unnecessary complexity.	automated model performance, hyperparameter impact, or data transformations.	and clearly, even for users with a limited statistical background.

Diagnostic Capability	Evaluate whether visualizations effectively enable users to diagnose problems, anomalies, or areas for improvement within AI processes or models.	Check that AutoML visualizations allow users to effectively diagnose issues in automated model training, identify convergence problems, or pinpoint data quality concerns.	Examine whether visualizations help users diagnose specific sources of bias or unfairness in the AutoML- generated model's behavior across different groups.
Interactivity and Exploration	Evaluate whether visualizations are interactive, allowing users to explore data and model characteristics dynamically.	Validate that AutoML visualizations allow users to interactively explore different model candidates, evaluate feature importance, and compare various automated pipeline stages.	Check whether fairness visualizations enable interactive exploration of group-specific performance, bias metrics, or the impact of different fairness algorithms.
Bridging Knowledge Gaps	Examine if visualizations effectively translate complex AI insights into a form that aligns with human knowledge and intuition.	Verify that AutoML visualizations bridge the gap between automated ML insights and the user's domain knowledge, making complex model decisions accessible.	Confirm that fairness visualizations effectively communicate the nature of bias and disparate impacts in a way that resonates with human ethical understanding.
Suitability for Non-Experts	Confirm that visualizations are particularly effective for users with limited technical or machine learning backgrounds.	Verify that AutoML visualizations are designed to be highly suitable for non-expert users, abstracting technical complexity while retaining critical information.	Validate that fairness visualizations are tailored for non-experts, allowing them to grasp ethical implications without needing deep
			machine learning expertise.

Avoiding Cognitive Biases	Confirm that the design of information representation avoids common human cognitive biases that could lead to misunderstandings.	Verify that AutoML's presentation of rules, explanations, or automated recommendations is carefully designed to prevent misinterpretations owing to user cognitive biases.	Validate that fairness information is presented in a way that minimizes the risk of users misinterpreting biased data or making biased decisions themselves.

Content Reliability: Content reliability, or trustworthiness, is closely linked to transparency, explainability, and the perceived accuracy of the AI system's outputs (Khuat et al., 2022) (Pop & Rațiu, 2024) . A careful design of how information (such as rules or explanations) is represented is needed to avoid misunderstandings due to human cognitive biases (Pop & Rațiu, 2024). Highlighting ambiguous predictions can help users assess the trustworthiness of models (Hoque et al., 2024).

Evaluation criteria	General description	AutoML specific description	Fairness context
Highlighting Ambiguity/Uncertainty	Examine whether the system explicitly highlights ambiguous predictions or uncertainties in its outputs to help users assess trustworthiness.	Confirm that AutoML highlights cases where its automated predictions are uncertain or where the model's confidence is low, allowing users to assess reliability.	Verify that AutoML explicitly highlights any ambiguity or uncertainty in its fairness assessments, such as when data for certain subgroups are scarce.
Perceived Accuracy of Outputs	Verify that users perceive the AI system's outputs as accurate and reliable based on their experience and system transparency.	Confirm that users perceive AutoML's generated models and predictions as accurate and trustworthy for their specific use	Validate that users perceive the fairness assessments and bias mitigation recommendations provided by AutoML as accurate and

		cases.	dependable.
Consistency of Information	Verify that all presented information, explanations, and rules remain consistent across different system components and over time.	Confirm that AutoML consistently applies and presents its automated rules, explanations for model choices, and performance metrics across different iterations or projects.	Validate that fairness- related information, definitions of protected attributes, and bias mitigation strategies are consistently applied and presented throughout the AutoML workflow.
Data Quality	Validate that information about data sources and quality is clearly communicated to underpin the reliability of content.	Verify that AutoML provides clear information about the quality of the data used for automated model training, enhancing content reliability.	Confirm that AutoML highlights any quality issues or biases within the training data that could impact the fairness and reliability of the generated models.

Human Augmentation Features

This dimension integrates principles from human factor engineering to optimize the collaborative relationship between humans and AutoML systems, thereby augmenting human capabilities rather than merely automating tasks. It encompasses designing for iterative engagement, allowing users to inject domain knowledge, refine models, and provide continuous feedback throughout the machine learning pipeline. This includes the development of iterative interpretability and explainability (XAI) features to make AI decisions transparent. Critical aspects also involve providing users with the ability to download reports, models, and benchmarks, empowering them with control, and enabling human oversight and override capabilities, especially in safety-critical contexts.

Iterative Engagement with Tool, Data, and Model/Automation: Iterative engagement is one of the core human augmentation principles in AI-driven tools (Khuat et al., 2022). This involves designing systems to support continuous feedback loops and allowing users to inject domain knowledge or refine models iteratively. Human involvement can occur throughout the entire ML pipeline(Theis et al., 2023). Stakeholder engagement is iterative because ML models are produced/examined, AutoML settings are reconfigured, and search/production is rerun. It is also nonlinear, allowing stakeholders to skip steps or revisit previous phases (Khuat et al., 2022). Continuous refinement and human-in-the-loop paradigms offer advantages over fully automated approaches, enabling users to inject domain knowledge, provide feedback, and iteratively refine models (Chromik & Butz, 2021). HITL involves integrating human input during an agent's learning process, allowing iterative updates and fine-tuning based on human feedback (Retzlaff et al., 2024).

Evaluation criteria	General description	AutoML specific description	Fairness context
Iterative Model Refinement	Validate that the system supports a cyclic process in which user input and AI outputs are continuously refined, leading to improved models.	Confirm that AutoML allows users to iteratively refine automated models by adjusting parameters, re-running searches, or selectingdifferent pipeline components based on observed performance.	Verify that the system enables iterative refinement of models specifically targeting fairness, allowing users to improve equitable outcomes based on continuous evaluation.
Human-in-the-Loop (HITL) Integration	Verify that human input is effectively integrated into the AI agent's learning process, leading to iterative updates and fine-tuning.	Confirm that AutoML integrates human input throughout the ML pipeline, allowing iterative updates and fine-tuning of automated model building based on user feedback.	Validate that HITL mechanisms within AutoML allow human input to directly inform and iteratively improve the fairness and ethical alignment of the models.
Non-Linear Workflow Support	Examine whether the system's workflow is flexible, allowing users to skip steps, revisit previous phases, or diverge from a strictly linear progression.	Verify that AutoML supports nonlinear exploration, allowing users to skip or revisit automated steps (e.g., data preprocessing and model selection) based on evolving needs.	Confirm that users can nonlinearly engage with fairness evaluation, allowing them to revisit data, model choices, or mitigation techniques as new biases are identified.
Human Augmentation Principle	Validate that iterative engagement design amplifies human capabilities, leading to enhanced quality,	Confirm that AutoML's iterative engagement principles enhance user productivity and decision-making quality by	Verify that iterative engagement within AutoML amplifies the human ability to identify, analyze, and
	productivity, and trust in AI-driven tools.	providing flexible human control over automated processes.	resolve fairness issues, leading to more robust and trustworthy ethical outcomes.

Iterative Interpretability and Explainability: Explainability (XAI) is a widely discussed concept at the nexus of AI and HCI, as it enables people to understand AI decisions (Chromik & Butz, 2021). Different types of explanations can be integrated into automated learning systems, such as visual and textual explanations (Khuat et al., 2022). Explainable AI must be designed to express helpful explanations while avoiding misunderstandings due to typical human cognitive biases. Explainability has additional feedback effects in enhancing the performance and reliability of ML solutions when human-in-the-loop collaboration is integral (Khuat et al., 2022). Explainable AI methods can make an agent's decision-making process transparent and interpretable (Retzlaff et al., 2024). This refers to providing explanations for a model's decisions, predictions, and actions. Explainability must be understood from the perspectives of human cognition

and emergent human-machine cognitive systems (Khuat et al., 2022). The iterative improvement of explanations based on human feedback is a promising approach (Khuat et al., 2022).

Evaluation criteria	General description	AutoML specific description	Fairness context
Understandability of AI Decisions	Verify that explanations enable users to accurately understand the rationale behind AI decisions, predictions, and actions.	Confirm that AutoML provides explanations that allow users to accurately understand why specific models were chosen, hyperparameter sets, or features were transformed.	Validate that explanations clarify fairness-related decisions made by AutoML, such as why a particular bias mitigation technique was applied.
Variety of Explanation Types	Evaluate whether different types of explanations (e.g., visual and textual) are integrated to cater to diverse user needs and contexts.	Check whether AutoML integrates various explanation types, such as visual representations of the model architecture or textual summaries of feature importance, to enhance understanding.	Examine whether AutoML provides diverse explanations for fairness issues, including visual comparisons of group performance or textual descriptions of bias sources.
Transparency of Decision-Making Process	Validate that explainable AI methods effectively make the AI agent's decision-making process transparent and interpretable to users.	This verifies that AutoML's XAI methods clearly reveal the rationale and inner workings behind the automated selection and optimization of machine learning models.	Ensure that explanations make the fairness-related decision-making process within AutoML transparent, showing how equitable considerations influence model selection.
Iterative Explanation Improvement	Confirm that the system supports iterative improvement of explanations based on continuous human feedback and interaction.	Validate that AutoML allows for the iterative refinement of its explanations of automated processes or model choices based on user queries and feedback.	Verify that AutoML actively incorporates human feedback to iteratively improve the clarity and effectiveness of its explanations of fairness and bias.

Feedback Exchange on Functionality or Performance: Feedback exchange is a key principle that enables users to provide inputs on system functionality and performance. Eedback can be used during model training to push the model to align its knowledge with human decisions or by learning through imitation (Theis et al., 2023). Human feedback is crucial for refining AI models and can lead to improved model quality, productivity, and trust (Khuat et al., 2022).

Designing AI systems as collaborators means considering feedback in both directions (AI understanding human intention and human understanding AI state). Effective feedback exchange mechanisms, facilitated by transparent and communicative AI, play a pivotal role in achieving successful AI-HCI integration, enhancing user perceptions, and ensuring ethical AI adoption **(**((Sundar & Lee, 2022) (Guzman & Lewis, 2019) .

Evaluation criteria	General description	AutoML specific description	Fairness context
User Input on Functionality/ Performance	Verify that the system provides clear and accessible channels for users to submit feedback on its overall functionality and performance.	Confirm that AutoML allows users to provide input on the effectiveness of its automated model building, search efficiency, or generated model performance.	Validate that users can easily submit feedback regarding the fairness-related functionality of AutoML or the perceived performance of bias- mitigation features.
Alignment with Human Decisions	Evaluate whether user feedback is effectively utilized to align the AI model's knowledge or decisions with human intentions and domain expertise.	leverages human feedback during automated training or refinement to align its generated models more closely with the desired outcomes or human expert judgments.	Examine whether user feedback, particularly on fairness, is effectively incorporated to align AutoML-generated models with human ethical standards and equitable decision-making.
Bidirectional Feedback Mechanisms	Examine whether the system supports feedback exchange in both directions, allowing AI to understand human intention and humans to understand AI state.	This verifies that AutoML not only receives human feedback but also provides clear communication back to the user regarding how their input influences automated model choices or search results.	Confirm that AutoML communicates how user feedback on fairness impacts model adjustments and provides transparent insights into AI's understanding of fairness objectives.
Transparency of Feedback Impact	Validate that the impact of user feedback on the AI system's adjustments or future behaviors is communicated to the user.	Verify that AutoML transparently indicates how user inputs, such as preference for certain model types or metrics, influence subsequent automated model recommendations.	Examine whether AutoML clearly demonstrates how user feedback regarding bias or fairness concerns has led to specific changes in the model or automated pipeline.

Download Reports, Models and Benchmarks: The feature to download reports, models, and benchmarks from AI tools serves as a significant human-computer interaction component by empowering users with control over outputs. This allows users to examine a specific ML pipeline, compare it with challenger models, and provide performance metrics. Such features to download models and benchmarks support processes for trust-building. It allows technical stakeholders to inspect, control, and manage the learning process (Khuat et al., 2022). The ability to download models and benchmarks also aligns with the principles of transparency and accountability, thereby fostering a sense of control and understanding, which are crucial for building trust in digital interactions (Zhang et al., 2024). It also allows users to independently validate the outcomes, assuring the reliability and consistency of outputs empowering users to perceive it to be a partner in decision-making rather than a black box technology (Balcombe & De Leo, 2022).

Evaluation criteria	General description	AutoML specific description	Fairness context
Feature Availability & Functionality	The system provides clear, functional, and easily accessible features to download various outputs.	Confirm that AutoML offers robust and user- friendly features to download generated models, comprehensive reports, and relevant benchmarks.	Validate that AutoML includes direct and functional features for downloading fairness reports, bias assessments, and fairness-adjusted models.
User Control over Outputs	Verify that the feature empowers users with direct control over accessing and managing AI-generated outputs.	Confirm that AutoML allows users to download generated models, detailed reports, and performance benchmarks to manage their own ML assets.	Validate that users can download reports or models containing specific fairness metrics or bias analysis results for an independent assessment.
Examination and Comparison	Evaluate whether downloadable outputs enable users to thoroughly examine AI pipelines and comparethem against alternatives.	Check that AutoML's downloadable reports allow users to comprehensively examine automated ML pipelines and compare them with manually built or challenger models.	Examine whether downloadable outputs facilitate detailed comparison of fairness metrics across different AutoML-generated models or against predefined fairness benchmarks.

Human Oversight and Override: Maintaining human oversight and the ability to control or override autonomous functions is essential, particularly in safety-critical domains. Human- centered AI (HCAI) and collaborative paradigms emphasize the importance of human involvement in AI systems. Human-in-the-loop (HITL) approaches incorporate human oversight mechanisms into AI models, ensuring that humans maintain control and the ability to intervene. Human controllers play a crucial role in collaborating with and overseeing system performance and safety. The design of trustworthy autonomous systems should support high levels of human oversight and situational awareness (Khuat et al., 2022). This is achieved through interfaces that facilitate oversight and override capabilities. In safety-critical applications, prioritizing the safety of human operators and enabling on-the-fly guidance from humans can mitigate dangerous actions (Theis et al., 2023).

Evaluation criteria	General description	AutoML specific description	Fairness context
Level of Oversight Supported	Verify that the system design consistently supports a high level of human oversight over autonomous functions.	Confirm that AutoML interfaces provide sufficient visibility and control points for users to oversee the automated model selection, training, and deployment processes.	Validate that AutoML allows users to effectively oversee fairness metrics and bias mitigation efforts, ensuring ethical alignment.
Override Capability	Evaluate whether the system provides clear, easily accessible, and effective mechanisms for human users to intervene and override autonomous actions.	Check that AutoML offers explicit override functions allowing users to halt, modify, or reject automated model suggestions or pipeline construction.	Examine whether users can easily override AutoML's decisions that might compromise fairness, for instance, by forcing a specific bias mitigation technique.
Collaboration and Control	This verifies that the system facilitates effective collaboration between human controllers and AI, thereby balancing autonomy with human control.	Confirm that AutoML fosters a collaborative environment whichere users can guide and refine automated processes while retaining ultimate control over model outcomes.	Validate that AutoML facilitates collaboration where users can guide fairness objectives, ensuring that the automated system aligns with ethical human priorities and avoids unintended bias.

5. Care and Responsibility

This dimension addresses the ethical, safety, and accountability aspects of AutoML system deployment, ensuring responsible human-AI collaboration. It involves designing features that support decision governance, promote accountability through transparent and predictable AI behavior, and establish robust safety "guardrails." Key HCI elements include the implementation of various disclosures (e.g., data/model cards, TEVV results, and failure mode histories) to build trust and enable informed user decisions. Furthermore, this dimension necessitates mechanisms for continuous improvement through regular updates and robust Adverse Incident Reporting Systems, ensuring that AutoML systems are not only effective but also reliable, safe, and ethically managed throughout their lifecycle.

Decision Governance: Decision governance in the context of human-computer interaction (HCI) in AI involves enabling features that support overseeing the decision-making processes of AI systems to ensure that these interactions are transparent, trustworthy, and aligned with human values. Further, decision governance in AI must consider the balance between automation and human accountability, particularly in high-stakes environments such as healthcare, finance, and autonomous systems (Çakır, 2024) (Liu, 2021).

Evaluation criteria	General description	AutoML specific description	Fairness context
Oversight of AI Decision-Making	Verify that the system provides features that enable users to effectively oversee AI's decision-making processes.	Confirm that AutoML offers mechanisms to oversee automated choices regarding model architecture, feature selection, or hyperparameter optimization throughout the pipeline.	Validate that AutoML provides features to oversee decisions related to fairness, such as how bias mitigation strategies are applied or which fairness metrics are prioritized.
Trustworthiness Alignment with Values	Confirm that the system design supports trustworthy interactions and aligns AI decisions with human values and ethical principles.	Verify that AutoML's automated decisions are demonstrably aligned with user-defined objectives and ethical guidelines, fostering trust in its outcomes.	Validate that AutoML's decision governance features ensure alignment with human values regarding fairness, promoting equitable outcomes, and preventing discriminatory decisions.

Accountability: Designing for accountability is part of incorporating human concerns into HCAI tools. Accountability is a desired mechanism for trustworthy autoML systems (Khuat et al., 2022). AI practitioners are obligated to take responsibility for public interaction (Mathewson, 2019). Transparency, understandability, and predictability are required for operators to hold autonomous systems accountable. Ensuring accountability is crucial for building trust in AI applications (Retzlaff et al., 2024).

Evaluation criteria	General description	AutoML specific description	Fairness context
Balance of Automation and Accountability	Examine whether the system strikes an appropriate balance between automated decision-making and clear human accountability for outcomes.	Confirm that AutoML's design maintains a clear allocation of responsibility, allowing users to understand when automation occurs and where human accountability lies for model deployment.	Verify that AutoML clearly delineates responsibility for fairness outcomes, ensuring that users can trace accountability for biased decisions to design or intervention points.
Auditability and Traceability	Evaluate whether the system provides comprehensive logs and audit trails to trace AI decisions and their contributing factors.	Check whether AutoML generates detailed audit trails for every automated decision point, including data transformations, model choices, and evaluation results.	Examine if AutoML's audit trails specifically document fairness- related interventions, such as bias detection reports and mitigation technique applications, for traceability.

Guardrails: While "guardrails" is a primary Human Computer Interaction feature, it is critical to designing reliable, safe, and trustworthy systems (Mathewson, 2019), implementing verification processes and control mechanisms, mitigating risks in safety-critical contexts, and enabling human oversight; the ability to avert undesirable actions is fundamental to Human Computer Interaction (Khuat et al., 2022) (Theis et al., 2023).

Evaluation criteria	General description	AutoML specific description	Fairness context
Design for Reliability	Verify that guardrails are designed to inherently enhance the reliability and predictable behavior of the AI system.	Confirm that AutoML incorporates guardrails to ensure the reliability of automated model generation, preventing the creation of unstable or poorly performing models.	Validate that guardrails within AutoML are designed to ensure the reliable application of fairness constraints, preventing unintended biases.
Safety Critical Contexts	Evaluate whether guardrails effectively mitigate risks and ensure safe operation, especially in safety-critical applications.	Check whether AutoML implements robust guardrails to prevent the deployment of potentially unsafe or unvetted automated models in safety-critical domains.	Examine whether guardrails in AutoML specifically prevent the generation or deployment of models that could lead to discriminatory harm in sensitive or critical applications.
Verification Processes	Confirm that guardrails include rigorous verification processes to validate system behavior against defined standards.	Verify that AutoML's guardrails incorporate automated verification processes to confirm model quality, adherence to performance thresholds, or data integrity before deployment.	Validate that the guardrails in AutoML include verification processes to confirm that fairness metrics meet predefined thresholds before a model is considered viable.
Control Mechanisms	Assess whether guardrails provide clear and effective control mechanisms for users to manage AI system operations and outputs.	Confirm that AutoML provides users with clear control mechanisms within the guardrails, allowing them to set bounds on search space, resource usage, or model complexity.	Verify that AutoML offers explicit control mechanisms within its guardrails, enabling users to set strict fairness targets or mandating the application of specific bias-mitigation techniques.

Disclosures: Disclosures of system capabilities, limitations, and potential errors are discussed as crucial aspects of transparency. Setting user expectations includes the following:

- Data/Model Cards:
- Model card (Raulf et al., 2023) and data card (Pushkarna et al., 2022) disclosure are practical mechanisms for enhancing transparency and supporting responsible AI development by providing structured documentation on models and datasets, including details relevant to training, evaluation, and intended use. TEVV Results:
- Testing, Evaluation, Validation, and Verification (TEVV) results as disclosure enable dosntream adoption of such systems, including gaining adequate visibility around usability evaluation, performance metrics, safety and robustness, and through user studies (Khuat et al., 2022). Such an effort supports users in determining the adoption approaches.

Failure modes/adverse incident history: Addressing potential failures and unpredictable behaviors is crucial for designing safe AI systems. Providing information regarding why a system fails or might fail or be unable to perform a task (e.g., a chatbot being unable to respond) is a part of disclosures (Shneiderman, 2020). Designing for human understanding and control is important so that stakeholders can interrupt anything incomprehensible and potentially dangerous. Disclosures are needed in case of errors or anticipated errors (Chromik & Butz, 2021). Highlighting and textually explaining ambiguous predictions helps users appropriately reassess their level of trust. Discrepancies between human and machine predictions indicate that an error exists, justifying the need for explanation and verification (Khuat et al., 2022).

Evaluation criteria	General description	AutoML specific description	Fairness context
System Capabilities Disclosure	Verify that the system clearly and accurately discloses its functionalities and the range of tasks it can perform.	Confirm that AutoML explicitly communicates the types of ML tasks it can automate, the algorithms it supports, and its computational limits.	Validate that AutoML discloses its capabilities in detecting and mitigating various types of biases or ensuring specific fairness criteria.
System Limitations Disclosure	Evaluate whether the system transparently communicates its inherent limitations, boundaries, and scenarios where it may not perform optimally.	Check that AutoML clearly informs users about its limitations regarding data size, model complexity, or the quality of solutions it can guarantee in specific contexts.	Examine if AutoML transparently discloses its limitations in achieving perfect fairness or its inability to detect certain subtle biases.
Data/Model Cards Implementation	Validate that structured documentation, such as Data Cards and Model Cards, is consistently provided for transparency and responsible A I development.	Verify that AutoML generates and presents comprehensive Model Cards for generated models and Data Cards for datasets used, detailing training, evaluation, and intended use.	Confirm that the Data and Model Cards generated by AutoML explicitly include sections on fairness considerations, protected attributes, and bias evaluation results.

TEVV Results Disclosure	Confirm that Testing, Evaluation, Validation, and Verification (TEVV) results are transparently disclosed to enable informed system adoption.	Validate that AutoML provides access to comprehensive TEVV results, including usability evaluations, performance metrics, and robustness assessments, to support adoption decisions.	Verify that the TEVV results disclosed by AutoML include detailed assessments of fairness metrics, bias detection results, and robustness of fairness-aware models.
Failure Modes/Adverse Incidents History	Examine whether the system provides clear information about its potential failure modes,	Confirm that AutoML logs and discloses past failure modes, instances where automation failed, or	Validate that AutoML records and discloses any incidents where automated processes
	adverse incident history, and reasons for inability to perform tasks.	explanations for the inability to find an optimal model.	lead to unfair or biased outcomes, including the identified reasons and context.
Ambiguity and Uncertainty Highlighting	Verify that the system highlights ambiguous predictions or uncertainties in its outputs to help users assess content trustworthiness.	Check whether AutoML explicitly highlights any ambiguity or low confidence in its automated predictions or model recommendations.	Confirm that AutoML clearly highlights any uncertainty in its fairness assessments, for instance, due to limited data for certain subgroups or complex bias interactions.

Patches and Updates: Regular updates and patches to AI systems play a crucial role in enhancing human-computer interaction (HCI) by ensuring that the tools remain relevant, accurate, and aligned with user expectations. A critical aspect of regular updates is their ability to improve the predictive performance of AI systems (Bansal et al., 2019). This is particularly crucial in high-stakes domains such as healthcare and criminal justice (domains where human and AI collaboration is essential for decision-making), wherein updates may actually harm overall team performance if they are not compatible with the user's past experiences (Bansal et al., 2019). Furthermore, AI systems should strive for a balance between performance and compatibility, ensuring that enhancements in AI performance are aligned with user expectations and do not disrupt established workflows (Bansal et al., 2019). For instance, in conversational AI, regular updates driven by deep learning and data from human interactions enable systems to become more adept at understanding and responding to natural language (Yan, 2018).

Evaluation criteria	General description	AutoML specific description	Fairness context
Level of Oversight Supported	Verify that the system design consistently supports a high level of human oversight over autonomous functions.	Confirm that AutoML's interfaces provide sufficient visibility and control points for users to oversee automated model selection, training, and deployment processes.	Validate that updates to AutoML specifically address newly identified fairness challenges or respond to user feedback on ethical considerations.
Predictive Performance Improvement	Evaluate whether updates demonstrably	Check that AutoML updates measurably enhance the predictive	Examine whether updates to AutoML lead to demonstrable improvements in the fairness
	improve the predictive performance and overall capabilities of the AI system.	performance of the generated models and improve the efficiency of the automated search process.	and equitable performance of the generated models. The updates shall not create inconsistent fairness outcomes compared to its previous version

Adverse Incident Reporting System: An Adverse Incident Reporting System (AIRS) serves as a critical component in the intersection of Human-Computer Interaction (HCI) and Artificial Intelligence (AI), playing an essential role in documenting and analyzing incidents that arise from AI systems. As AI systems become more prevalent, adverse events provide invaluable learning opportunities to refine AI algorithms and improve their interactions with humans (Lupo, 2023). A robust incident-reporting strategy facilitates the development of flexible regulatory frameworks that evolve alongside new AI technologies (Lupo, 2023). Documentation of incidents, particularly those highlighting system biases, is crucial for learning from past failures and advancing policy recommendations aimed at mitigating these risks (Turri & Dzombak, 2023). Proper incident reporting assists teams in understanding and contextualizing AI actions, fostering trust, and improving the effectiveness of human-AI collaborations (Zhang et al., 2024). Moreover, as AI systems increasingly handle complex tasks, an understanding augmented by clear incident reporting can bridge the gap between human expectations and AI functionalities (Zhang et al., 2024). Designing human-AI interactions is complicated by AI's output complexity and the uncertainty surrounding its capabilities. AIRS can provide insights into these design challenges, aiding designers and researchers in effectively addressing them (Yang et al., 2020).

Evaluation criteria	General description	AutoML specific description	Fairness context
Documentation & Analysis of Incidents	Verify that the AIRS provides clear, comprehensive, and standardized mechanisms for documenting and analyzing adverse incidents from AI systems.	Confirm that AutoML's AIRS effectively documents and analyzes incidents related to automated model failures, unexpected performance, or resource issues during its operation.	Validate that AIRS specifically captures and analyzes incidents, highlighting biases in AutoML-generated models or unfair outcomes across user groups.
Bridging Expectation Gaps	Evaluate whether understanding augmented by clear incident reporting can bridge the gap between human expectations and AI functionalities.	Check that AutoML's AIRS provides insights into its operational complexities, helping to bridge user expectations with the actual functionalities of automated ML.	Examine whether the AIRS helps bridge the gap between human expectations of fair AI and AutoML's actual performance regarding equity, guiding users on realistic capabilities.

Conclusion

The field of human–computer interaction (HCI) is undergoing significant evolution, particularly with the rise of Automated Machine Learning (AutoML). This shift highlights the crucial role of human oversight, understanding, and collaboration, even in increasingly autonomous systems (Balcombe & De Leo, 2022). HCI plays a crucial role in bridging the gap between human capabilities and technological advancements. It employs a range of research methods, including experimental design, eye-tracking, qualitative research, and cognitive modeling (*Research Methods for Human-Computer Interaction*, 2008). The field continues to evolve, addressing the challenges posed by new technologies and user demands while also contributing to the development of psychological and social theories in the context of technology use (Carroll, 1997).

As AutoML continues to advance towards greater autonomy, a human-centered approach to HCI will remain paramount, transforming human roles from direct operation to strategic supervision and collaborative partnership with intelligent systems. The evolution of human– computer interaction (HCI) from a focus on basic functionality to encompassing user experience, ethical considerations, and emerging technologies has profoundly impacted the development of Artificial Intelligence (AI) systems, particularly Automated Machine Learning (AutoML). Although AutoML initially aimed to minimize human intervention, the growing recognition of the need for human oversight, collaboration, and trust has underscored the critical role of HCI (Holzinger et al., 2025).

As AutoML systems advance towards greater autonomy (AutonoML), the nature of human interaction is expected to evolve. The relationship may transform from direct instruction to collaboration, where the system is seen more as a partner or teammate. This collaboration leverages the complementary strengths of humans and machines. New human roles, such as "explainers" and "sustainers," may emerge to bridge the human-system gap, interpret system behaviors, ensure ethical compliance, and validate outcomes. Optimizing collaborative interactions involves strategically distributing tasks based on the strengths of humans and the autonomous system. Modern viewpoints advocate for a Human-Centered AI (HCAI) framework that treats automation and human control as orthogonal axes, ensuring that humans retain the option to intervene or oversee (Khuat et al., 2022).

Convergence of HCAI views presents an opportunity to address the black-box nature of AutoML systems by incorporating HCI principles to enhance user understanding and control. Therefore, emergent approaches are attempting to address the greatest weakness of modern AutoML offerings – their black-box nature – which serves as a significant motivating factor for further research into HCI (Mueller et al., 2023). To improve the interaction between humans and AutoML systems, researchers can create more transparent, user-friendly, and trustworthy automated machine learning tools that balance automation with user understanding and control (Karmaker (“Santu”) et al., 2021) (Li et al., 2020).

This paper proposes a comprehensive framework for HCI evaluation in AutoML, structured across five crucial dimensions: Contracts and User Development, User Interface, Interaction and Experience Design, Information Architecture, Human Augmentation Factors Design, and Care and Responsibility. By prioritizing transparency, explainability, usability, and ethical considerations, this framework aims to foster trust and ensure that AutoML systems are not merely functional but also equitable, safe, and truly beneficial for all users.

This framework emphasizes transparent disclosures, intuitive interfaces, structured information presentation, user empowerment, and ethical considerations, including fairness and accountability. Further, the framework exhibits the need for prioritizing user-centered design and aims to bridge the gap between AutoML's technical capabilities and diverse user needs, fostering a future where AI systems are not only efficient but also trustworthy, transparent, and truly collaborative partners. The continued integration of HCI principles is paramount for realizing the full potential of AutoML, ensuring that these powerful tools are developed and deployed in a manner that benefits all users and society (Nakao et al., 2022) (Yu, 2023) .

References

Human-Computer Interaction. (2003). crc. https://doi.org/10.1201/9780367804787

Olson, G. M., & Olson, J. S. (2002). Human-Computer Interaction: Psychological Aspects of the Human Use of Computing. *Annual Review of Psychology*, *54*(1), 491–516. https://doi.org/10.1146/annurev.psych.54.101601.145044

Fallman, D. (2007). Why Research-Oriented Design Isn't Design-Oriented Research: On the Tensions Between Design and Research in an Implicit Design Discipline. *Knowledge, Technology & Policy*, *20*(3), 193–200. https://doi.org/10.1007/s12130-007-9022-8

Kheder, H. A. (2023). HUMAN-COMPUTER INTERACTION: ENHANCING USER EXPERIENCE IN INTERACTIVE SYSTEMS. *Kufa Journal of Engineering*, *14*(4), 23–41. https://doi.org/10.30572/2018/kje/140403

Lakkshmanan, A., Tyagi, A. K., Sree, P. H., & Sharma, A. K. (2024). *Engineering Applications of Artificial Intelligence* (pp. 166–179). igi global. https://doi.org/10.4018/979-8-3693-5261- 8.ch010

Kosch, T., Zagermann, J., Schmidt, A., Reiterer, H., Karolus, J., & Woźniak, P. W. (2023). A Survey on Measuring Cognitive Workload in Human-Computer Interaction. *ACM Computing Surveys*, *55*(13s), 1–39. https://doi.org/10.1145/3582272

Khuat, T., Kedziora, D., & Gabrys, B. (2022). *The Roles and Modes of Human Interactions with Automated Machine Learning Systems*. cornell university. https://doi.org/10.48550/arxiv.2205.04139

Karmaker ("Santu"), S. K., Hassan, M. M., Zhai, C., Smith, M. J., Xu, L., & Veeramachaneni, K. (2021). AutoML to Date and Beyond: Challenges and Opportunities. *ACM Computing Surveys*, *54*(8), 1–36. https://doi.org/10.1145/3470918

Crisan, A., & Fiore-Gartland, B. (2021). *Fits and Starts: Enterprise Use of AutoML and the Role of Humans in the Loop*. 1–15. https://doi.org/10.1145/3411764.3445775

Mathewson, K. (2019). *A Human-Centered Approach to Interactive Machine Learning*. https://doi.org/10.48550/arxiv.1905.06289

Li, Y., Lasecki, W. S., Hilliges, O., & Kumar, R. (2020, April 25). *Artificial Intelligence for HCI: A Modern Approach*. https://doi.org/10.1145/3334480.3375147

Lindauer, M., Karl, F., Klier, A., Moosbauer, J., Tornede, A., Mueller, A., Hutter, F., Feurer, M., & Bischl, B. (2024). *Position: A Call to Action for a Human-Centered AutoML Paradigm*. https://doi.org/10.48550/arxiv.2406.03348

Bach, T. A., Khan, A., Hallock, H., Beltrão, G., & Sousa, S. (2022). A Systematic Literature Review of User Trust in AI-Enabled Systems: An HCI Perspective. *International Journal of Human– Computer Interaction*, *40*(5), 1251–1266. https://doi.org/10.1080/10447318.2022.2138826

Veitch, E., & Alsos, O. A. (2021). Human-Centered Explainable Artificial Intelligence for Marine Autonomous Surface Vehicles. *Journal of Marine Science and Engineering*, *9*(11), 1227. https://doi.org/10.3390/jmse9111227

Ehsan, U., Muller, M., Liao, Q. V., Riedl, M. O., & Weisz, J. D. (2021). *Expanding Explainability: Towards Social Transparency in AI systems*. 1–19. https://doi.org/10.1145/3411764.3445188

Whaiduzzaman, M., Shahrier, L., Barros, A., Jan, T., Rahman, M. S., Sakib, A., Fidge, C., Khan, N. J., Thompson-

Whiteside, S., Mahi, M. J. N., Ghosh, S., & Chaki, S. (2023). Concept to Reality: An Integrated Approach to Testing Software User Interfaces. *Applied Sciences*, *13*(21), 11997. https://doi.org/10.3390/app132111997

Mishra, R., Pati, B., & Satpathy, R. (2023). Human Computer Interaction Applications in Healthcare: An Integrative Review. *EAI Endorsed Transactions on Pervasive Health and Technology*, *9*. https://doi.org/10.4108/eetpht.9.4186

Vorm, E. S., & Combs, D. J. Y. (2022). Integrating Transparency, Trust, and Acceptance: The Intelligent Systems Technology Acceptance Model (ISTAM). *International Journal of Human– Computer Interaction*, *38*(18–20), 1828–1845. https://doi.org/10.1080/10447318.2022.2070107

Amershi, S., Vorvoreanu, M., Teevan, J., Horvitz, E., Fourney, A., Suh, J., Kikin-Gil, R., Bennett, P. N., Nushi, B., Collisson, P., Iqbal, S., Weld, D., & Inkpen, K. (2019, May 2). *Guidelines for Human- AI Interaction*. https://doi.org/10.1145/3290605.3300233

Raulf, A., Bruder, C., Berro, C., Deligiannaki, F., Theis, S., & Jentzsch, S. (2023). *Requirements for Explainability and Acceptance of Artificial Intelligence in Collaborative Work*. https://doi.org/10.48550/arxiv.2306.15394

Le D. J. L., & Macke, S. (2020). A Human-in-the-loop Perspective on AutoML: Milestones and the Road Ahead. IEEE Data Engineering Bulletin.

Retzlaff, C. O., Afshari, M., Mousavi, P., Wayllace, C., Das, S., Holzinger, A., Saranti, A., Yang, T., Angerschmid, A., & Taylor, M. E. (2024). Human-in-the-Loop Reinforcement Learning: A Survey and Position on Requirements, Challenges, and Opportunities. *Journal of Artificial Intelligence Research*, *79*, 359–415. https://doi.org/10.1613/jair.1.15348

Pushkarna, M., Kjartansson, O., & Zaldivar, A. (2022). *Data Cards: Purposeful and Transparent Dataset Documentation for Responsible AI*. 1776–1826. https://doi.org/10.1145/3531146.3533231

Pushkarna, M., Zaldivar, A., & Kjartansson, O. (2022). *Data Cards: Purposeful and Transparent Dataset Documentation for Responsible AI*. cornell university. https://doi.org/10.48550/arxiv.2204.01075

Li, Y., Wu, B., Huang, Y., & Luan, S. (2024). Developing trustworthy artificial intelligence: insights from research on interpersonal, human-automation, and human-AI trust. *Frontiers in Psychology*, *15*. https://doi.org/10.3389/fpsyg.2024.1382693

Chromik, M., & Butz, A. (2021). *Human-XAI Interaction: A Review and Design Principles for Explanation User Interfaces* (pp. 619–640). springer. https://doi.org/10.1007/978-3-030-85616- 8_36

Ashtana, R., Vikash, V., & Omprakash, O. (2018). Human-centric machine learning: Addressing user experience and ethical considerations. *International Journal of Applied Research*, *4*(12), 65–69. https://doi.org/10.22271/allresearch.2018.v4.i12a.11467

Pop, E.-L., & Raţiu, A. (2024). *Human-Computer Interaction in Artificial Intelligence with Applications in Healthcare: A Review*. institut f r ost und s dosteuropaforschung. https://doi.org/10.3233/faia241213

Rundo, L., Sala, E., Vitabile, S., Gambino, O., & Pirrone, R. (2020). Recent advances of HCI in decision-making tasks for optimized clinical workflows and precision medicine. *Journal of Biomedical Informatics*, *108*, 103479. https://doi.org/10.1016/j.jbi.2020.103479

Acemyan, C. Z., & Kortum, P. (2012). The Relationship Between Trust and Usability in Systems. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *56*(1), 1842–1846. https://doi.org/10.1177/1071181312561371

Hoque, M., Shin, S., & Elmqvist, N. (2024). *Visualization for Human-Centered AI Tools*. https://doi.org/10.48550/arxiv.2404.02147

Veale, M., Binns, R., & Van Kleek, M. (2018). *Some HCI Priorities for GDPR-Compliant Machine Learning*. center for open science. https://doi.org/10.31228/osf.io/wm6yk

Buçinca, Z., Malaya, M., & Gajos, K. (2021). *To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making*. https://doi.org/10.48550/arxiv.2102.09692

Chen, J. Y. C. (2022). Transparent Human–Agent Communications. *International Journal of Human–Computer Interaction*, *38*(18–20), 1737–1738. https://doi.org/10.1080/10447318.2022.2120173

Yu, S. (2023). *Towards Trustworthy and Understandable AI: Unraveling Explainability Strategies on Simplifying Algorithms, Appropriate Information Disclosure, and High-level Collaboration*. 133–143. https://doi.org/10.1145/3616961.3616965

El Ali, A., Venkatraj, K. P., Morosoli, S., Cesar, P., Naudts, L., & Helberger, N. (2024). *Transparent AI Disclosure Obligations: Who, What, When, Where, Why, How*. *43*, 1–11. https://doi.org/10.1145/3613905.3650750

Xu, W., Dainoff, M. J., Ge, L., & Gao, Z. (2022). 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*, *ahead-of-print*(ahead-of-print), 494–518. https://doi.org/10.1080/10447318.2022.2041900

Nazar, M., Yafi, E., Su'Ud, M. M., & Alam, M. M. (2021). A Systematic Review of Human– Computer Interaction and Explainable Artificial Intelligence in Healthcare With Artificial Intelligence Techniques. *IEEE Access*, *9*, 153316–153348. https://doi.org/10.1109/access.2021.3127881

Sundar, S. S., & Lee, E.-J. (2022). Rethinking Communication in the Era of Artificial Intelligence. *Human Communication Research*, *48*(3), 379–385. https://doi.org/10.1093/hcr/hqac014

Guzman, A. L., & Lewis, S. C. (2019). Artificial intelligence and communication: A Human– Machine Communication research agenda. *New Media & Society*, *22*(1), 70–86. https://doi.org/10.1177/1461444819858691

Zhang, R., Duan, W., Knijnenburg, B., Flathmann, C., Mcneese, N. J., Schelble, B., & Musick, G. (2024). I Know This Looks Bad, But I Can Explain: Understanding When AI Should Explain Actions In Human-AI Teams. *ACM Transactions on Interactive Intelligent Systems*, *14*(1), 1–23. https://doi.org/10.1145/3635474

Balcombe, L., & De Leo, D. (2022). Human-Computer Interaction in Digital Mental Health. *Informatics*, *9*(1), 14. https://doi.org/10.3390/informatics9010014

Çakır, A. M. (2024). Human AI Collaboration in Decision-Making Auto Systems With AI. *Human Computer Interaction*, *8*(1), 123. https://doi.org/10.62802/b4z5p105

Liu, B. (2021). In AI We Trust? Effects of Agency Locus and Transparency on Uncertainty Reduction in Human–AI Interaction. *Journal of Computer-Mediated Communication*, *26*(6), 384–

402. https://doi.org/10.1093/jcmc/zmab013

Shneiderman, B. (2020). Human-Centered Artificial Intelligence: Three Fresh Ideas. *AIS Transactions on Human-Computer Interaction*, *12*(3), 109–124. https://doi.org/10.17705/1thci.00131

Bansal, G., Nushi, B., Kamar, E., Horvitz, E., Weld, D. S., & Lasecki, W. S. (2019). Updates in Human-AI Teams: Understanding and Addressing the Performance/Compatibility Tradeoff. *Proceedings of the AAAI Conference on Artificial Intelligence*, *33*(01), 2429–2437. https://doi.org/10.1609/aaai.v33i01.33012429

Yan, R. (2018). *"Chitty-Chitty-Chat Bot": Deep Learning for Conversational AI*. 5520–5526. https://doi.org/10.24963/ijcai.2018/778

Lupo, G. (2023). Risky Artificial Intelligence: The Role of Incidents in the Path to AI Regulation. *Law, Technology and Humans*, *5*(1), 133–152. https://doi.org/10.5204/lthj.2682

Turri, V., & Dzombak, R. (2023). *Why We Need to Know More: Exploring the State of AI Incident Documentation Practices*. 576–583. https://doi.org/10.1145/3600211.3604700

Yang, Q., Steinfeld, A., Zimmerman, J., & Rosé, C. (2020). *Re-examining Whether, Why, and How Human-AI Interaction Is Uniquely Difficult to Design*. 1–13. https://doi.org/10.1145/3313831.3376301

Research Methods for Human-Computer Interaction. (2008). cambridge university. https://doi.org/10.1017/cbo9780511814570

Carroll, J. M. (1997). HUMAN-COMPUTER INTERACTION: Psychology as a Science of Design. *Annual Review of Psychology*, *48*(1), 61–83. https://doi.org/10.1146/annurev.psych.48.1.61

Holzinger, A., Zatloukal, K., & Müller, H. (2025). Is human oversight to AI systems still possible? *New Biotechnology*, *85*, 59–62. https://doi.org/10.1016/j.nbt.2024.12.003

Mueller, F. 'Floyd,' Andres, J., Mehta, Y., Semertzidis, N., Li, X., Benford, S., Marshall, J., & Matjeka, L. (2023). Toward Understanding the Design of Intertwined Human–Computer Integrations. *ACM Transactions on Computer-Human Interaction*, *30*(5), 1–45. https://doi.org/10.1145/3590766

Nakao, Y., Strappelli, L., Stumpf, S., Naseer, A., Regoli, D., & Gamba, G. D. (2022). Towards Responsible AI: A Design Space Exploration of Human-Centered Artificial Intelligence User Interfaces to Investigate Fairness. *International Journal of Human–Computer Interaction*, *ahead-of-print*(ahead-of-print), 1762–1788. https://doi.org/10.1080/10447318.2022.2067936