Artificial Intelligence and Social Computing, Vol. 163, 2025, 215–224
https://doi.org/10.54941/ahfe1006053

AI Tool Compliance Reporting: A Heuristic Analysis of Survey Data Using Natural Language Processing

Aimee Kendall Roundtree

Texas State University, San Marcos, TX 78666, USA

ABSTRACT

This study examined how well New York City's public AI tools reported good design practices for users. It analyzes 76 reports about algorithmic tools using a mix of computer methods (natural language processing), human review, and Nielsen's ten common heuristics for good usability, such as showing system status, giving users control, and providing help. The tools often followed some of these rules—especially those that support transparency, user control, and clear design. But others, like helping users prevent mistakes or reducing memory load, were rarely used. Agencies may be focusing more on making tools technically sound and less on making them easy and fair to use. We also looked at the language in the reports and found differences based on heuristic. Some used more formal or technical words, while others were simpler and more user-friendly. This study's findings confirm earlier ones that public trust in AI depends on transparency and fairness. More work is needed to include all users, especially regarding high-risk tools like those used in healthcare or law enforcement. Future studies should involve users and designers directly and look at tools across more sectors to improve design and fairness in public AI.

Keywords: Artificial intelligence, Compliance records, Compliance reporting, Social computing

INTRODUCTION

Understanding user experience (UX) in artificial intelligence (AI) applications must influence decisions in public services, healthcare, and law enforcement. Evaluating usability also helps promote transparency, trust, and equity in algorithmic decision-making. This study applies Jakob Nielsen's ten usability heuristics to analyze algorithmic tools disclosed by New York City agencies in the Algorithmic Tools Compliance Record. Seventy-six reports were evaluated using natural language processing (NLP) and human coding to identify heuristic-related language in tool descriptions, including their purpose, datasets, and vendor information. The study found that the heuristics represented focus on risk mitigation and reliability. However, other heuristics appear less frequently, which raises concerns about system transparency, intuitiveness, and cognitive load, suggesting gaps in user-centered design and documentation. The analysis shows agencies may prioritize technical functionality and risk management over user experience elements. Furthermore, limited report details may have caused underrepresentation of several heuristics. We must incorporate broader UX principles

into public sector system design. Overall, while public tools show strengths in reliability, greater attention to transparency, simplicity, and support will enhance user experience.

BACKGROUND

Public attitudes are ambivalent about AI. There is optimism about how AI can improve healthcare, but also fear about job loss, data misuse, and decision-making (Roundtree, 2024). Trust hinges on transparency and user involvement, so AI programs must factor in public sentiment, ethical awareness, and social acceptance.

Several studies deepen this focus on transparency and fairness by emphasizing the interdependence of technical and ethical design. A review of 42 peer-reviewed studies on AI explainability, interpretability, fairness, and privacy found that these dimensions collectively underpin public trust—AI cannot be fair if it is incomprehensible or lacks consent (Roundtree, 2023). For AI to be trustworthy, end-user must be able to interpret AI and algorithms must be transparent.

These concerns are particularly pronounced in the deployment of facial recognition technologies (FRT). For example, FRT exhibits disproportionately high error rates for racial minorities, raising serious ethical and legal implications when used in law enforcement (Roundtree, 2021, June). Sixteen industry and professional codes of ethics for FRT find that most focus on professionalism or compliance, but few offer guidance on operationalizing transparency or incorporating public feedback (Roundtree, 2022). Ethical AI must move beyond abstract principles to participatory, enforceable frameworks.

Participatory design and socio-technical perspectives offer crucial frameworks for developing ethical and inclusive AI systems. Prioritizing user concerns—particularly from marginalized communities—highlights the importance of participatory design, especially through industry-academic partnerships that develop facial recognition systems for particular sectors (Roundtree, 2021, October). Applying actor-network theory (ANT) to AI ethics, responsibility in AI systems must be distributed among human and nonhuman agents (Roundtree, 2020). This challenges conventional ethical models and supports the thesis that AI governance must reflect the complexity of socio-technical systems.

The demand for human-centered design is echoed across UX literature. A review of 359 studies finds that most AI design tools prioritize automation over empathy, leading to tools that misalign with designers' workflows or users' needs (Lu et al., 2024). They advocate for designer-centric datasets and evaluation metrics to shift focus toward ethical, effective AI support. Automation in UX can also introduce job stress and creative disempowerment unless implemented with clear explainability and human oversight (Stige et al., 2024). Ethical AI design must center on human experience, distribute responsibility, and resist automation that undermines empathy, transparency, and user agency.

The emphasis on human-centeredness also applies to conversational AI systems. Boundary-aware design respects privacy, disclosure norms, and user autonomy (Zheng et al., 2022). Drawing from Communication Privacy Management Theory, AI should navigate disclosure, identity, and temporal boundaries to uphold user dignity in mediated interaction (Palen and Dourish, 2003). Systems that facilitate interaction while adapting to complex social dynamics reinforce ethical engagement through context-sensitivity and respect.

Practical challenges around explainability and user control continue to hinder AI's integration in UX. While AI can dynamically adapt interfaces, misalignment with human workflows risks fatigue and disengagement (Johnston et al., 2019). Despite their sophistication, AI tools remain underused due to incompatibility with the non-linear, exploratory nature of design (Abbas et al., 2022). A knowledge gap between AI developers and UX designers demonstrates the need for collaborative tools that support shared understanding rather than opacity (Yang et al., 2020). Bridging these gaps requires co-creation of tools with designers and users to improve human-machine collaboration.

The literature also highlights the need for inclusive and participatory design to counteract algorithmic bias. Participatory methods ensure that AI serves marginalized users rather than reproducing inequities (McKenna-Aspell et al., 2022). AI can make participation easy, but relying too much on automated profiling runs the risk of reinforcing stereotypes (Wallach et al., 2020). These studies affirm that inclusivity is not only ethical but essential for accurate, effective, and trusted AI.

At a systems level, researchers warn against unchecked automation without ethical frameworks. AI risks displacing human intuition in design unless systems preserve creativity and contextual reasoning (Verganti et al., 2020). Emerging risks include diminished autonomy, explainability loss, and diffused accountability (Koch, 2017). Koch calls for design models embedding transparency and user control as foundational—not optional—features.

AI must be human-centered and transparent to earn public trust. Scholars warn against premature automation, call for inclusive design, and advocate participatory engagement to mitigate bias. Yet questions remain: Do designers meaningfully integrate public voices? What mechanisms ensure transparency without overload? And how can governance evolve to hold human and nonhuman agents accountable? Future research must explore these questions through interdisciplinary, empirical inquiry to ensure AI fulfils its promise ethically and fairly.

METHODS

The dataset included 76 reports with detailed descriptions of algorithmic tools, data characteristics, vendor information, and usage dates. Text from these columns was consolidated into a single field for analysis. Each heuristic was represented by keywords and definitions related to its core principles. For content analysis, natural language processing (NLP) methods using Python

and LIWC analyzed occurrences of heuristic-related concepts in the text and categorized and quantified adherence to each heuristic. Human coding verified computer coding.

We applied NLP techniques to categorize NYC agency algorithmic tool reports using Jakob Nielsen's usability heuristics as classification labels. The primary objective was to align unstructured text describing algorithmic tools with one of ten usability heuristics, based on semantic similarity. We used heuristic definitions and category names as our labeled dataset, treating each definition as representative text for its corresponding principle. From this, we constructed a training corpus mapping each heuristic to a body of descriptive language. To prepare textual data for classification, we applied TF-IDF (Term Frequency-Inverse Document Frequency) vectorization, converting text into numerical format that highlights word importance within and across documents. We trained a logistic regression classifier—a robust, interpretable model—on this TF-IDF-transformed text to distinguish between heuristic categories. Though trained on a small number of definitions, the model learned characteristic language patterns for each heuristic. We then applied this model to the algorithmic tool reports. Each description was tokenized, vectorized, and classified into one of the predefined heuristic categories. The classifier output the closest heuristic for each report. This NLP approach combines shallow learning methods (TF-IDF and logistic regression) with supervised classification to annotate short, structured reports with UX-relevant categories. The method is transparent, scalable, and modifiable for usability-focused text classification in public sector and compliance settings.

LIWC (Linguistic Inquiry and Word Count) is software that assesses psychological, cognitive, and social dimensions through language. It operates by scanning digital text files and comparing each word against a pre-established dictionary of over 12,000 words, stems, phrases, and emoticons. These words are categorized into subdictionaries such as emotion, cognition, social processes, and more. The software calculates the percentage of words in a text that fall into each category, producing a comprehensive profile of linguistic features. LIWC's advanced capabilities include the Meaning Extraction Method (MEM), which performs factor analyses to uncover dominant themes; the Narrative Arc module for story structure; and Language Style Matching to evaluate linguistic synchrony. The Contextualizer module adds qualitative depth by extracting keywords within context. LIWC aided psychological, social, and linguistic analysis.

To evaluate linguistic and psychological patterns in agency and heuristic-level communication, statistical analyses were conducted on a structured dataset of language features. Pearson's correlation measured the strength and direction of linear relationships between numeric variables. Pearson's coefficient (r) calculated correlations. These correlations were paired with p-values to assess statistical significance ($p < 0.05$).

RESULTS

Per Nielsen, **visibility of system status** (**Heuristic 1**) ensures that users receive clear, timely feedback about what the system is doing, fostering trust and informed decision-making. **Match between the system and the real world** (**Heuristic 2**) emphasizes the use of familiar language and real-world conventions to create intuitive interactions. **User control and freedom** (**Heuristic 3**) lets users undo actions, exit unwanted states, and prevent frustration. **Consistency and standards** (**Heuristic 4**) encourage designers to adhere to platform and industry norms so users can rely on familiar patterns. **Error prevention** (**Heuristic 5**) focuses on designing systems that anticipate and prevent mistakes before they occur, reducing the need for corrective actions. **Recognition rather than recall** (**Heuristic 6**) minimizes cognitive load by making relevant options and information visible, rather than requiring users to remember details. **Flexibility and efficiency of use** (**Heuristic 7**) supports novice and experienced users through shortcuts, customization, and multiple tasks. **Aesthetic and minimalist design** (**Heuristic 8**) eliminates unnecessary elements for clarity and focus**. Using plain language, visuals, and guidance** (**Heuristic 9**) helps users recognize, diagnose, and recover from errors. Finally, **help and documentation** (**Heuristic 10**) should be accessible, concise, and task-oriented for when direct interaction alone is not enough.

Per LIWC, **tone** measure emotional positivity or negativity in text. Higher scores reflect positivity and lower scores reflect negativity. **Analytic** language represents the degree of formal, logical, and hierarchical thinking. **Authenticity** estimates the perceived honesty and genuineness of the author's language. **Clout** reflects the relative social status, confidence, and leadership conveyed in language, with higher scores indicating more authoritative or dominant tone. **Cognition** is a superordinate category that reflects how often people refer to mental activities such as thinking, knowing, remembering, and reasoning. **Linguistic processes** refer to word usage that shapes sentence construction and meaning. **Function** measures sentence structure and linguistic style.

The most frequently identified usability principles were Visibility of System Status (n = 14), User Control and Freedom (n = 14), Aesthetic and Minimalist Design (n = 14), Help and Documentation (n = 14), and Consistency and Standards (n = 10). These five heuristics dominated the categorization, indicating that NYC agencies designing or procuring algorithmic tools are largely focused on user transparency, control, clean interfaces, and support documentation. Match between the system and the real world, flexibility and efficiency of use, and helping users recognize and diagnose errors, and recover from them.

Visibility of System Status tied for the highest frequency, which underscores a consistent focus on keeping users informed about system operations. This is especially evident in emergency response tools (e.g., EMS and fire services), where users need immediate, real-time feedback about system decisions or current load distributions.

User Control and Freedom appeared equally often. Tools under this category often allowed for manual input or adjustment to give end-users the

flexibility to influence outputs, which is a key feature for systems requiring human judgment or operation in dynamic environments.

Aesthetic and Minimalist Design also ranked high. Reports categorized under this heuristic frequently referenced simplified outputs and visual clarity, supporting cognitive ease and reducing distractions in decision-making contexts.

Help and Documentation featured prominently as well, showing that many systems come with guidance materials, tutorials, or embedded help features. This emphasis ensures usability across a broad range of users with varying levels of technical expertise—critical in large, decentralized city agencies.

Consistency and Standards, though slightly less frequent (10 times), was still well represented. Tools with this designation often aligned with sector norms or internal standards, promoting ease of training and smoother integration across platforms or departments.

Notably, heuristics such as Error Prevention, Recognition Rather Than Recall, and others from the full Nielsen framework were underrepresented or absent in this dataset. Agencies are attentive to transparency and user support, but they can try to reduce user error and minimize cognitive effort more.

These patterns offer practical insights for technologists, designers, and procurement officers. They reveal areas where human-centered design is well-applied (e.g., visibility and support), while highlighting opportunities to further enhance usability—especially for non-expert users and high-stakes contexts. As the public sector increasingly adopts AI and algorithmic tools, such evaluations will be vital in ensuring accountability, usability, and equity in service delivery.

Table 1: Heuristics frequencies.

Heuristic	Freq.	Example
1: Visibility of System Status	14	This algorithm and the resulting output file that is used in our EMS CAD system to suggest atom order for unit search is currently provided by a vendor.
3: User Control and Freedom	14	The tool can take the total number of available staff and optimally allocate them across tours to maximize the minimum difference between supply and demand.
8: Aesthetic and Minimalist Design	14	[G]enerate a complete minimum spanning tree (MST) Used to generate the minimum spanning tree relationships which are used to rule in or out Legionella strains in outbreaks.
10: Help and Documentation	14	The tool determines the location of the sound source, and once classified as potential gunfire sends the incident to acoustic experts for additional analysis.
4: Consistency and Standards	10	The tool requires a human user to evaluate the output data to see if complaints identified as similar are, in fact, connected to a pattern.

Continued

Table 1: Continued

Heuristic	Freq.	Example
6: Recognition Rather than Recall	7	The tool analyzes an uploaded image or video and searches and compares it with lawfully possessed images to generate a pool of possible matches.
5: Error Prevention	3	Produces phylogenetic trees which are used to rule in or out bacteria such as N.

The linguistic features across usability heuristic categories reveal distinct rhetorical styles. Help and Documentation stands out with the lowest word count (108), highest punctuation use (20.6% total punctuation), and shortest sentence structure, indicating highly structured, directive language. In contrast, Consistency and Standards and User Control and Freedom have the highest word counts (384.7 and 300.6 respectively) and longer, more elaborate sentences, reflecting more descriptive and narrative text. Visibility of System Status shows a high use of personal pronouns and articles, suggesting a user-centered focus. Cognitive processing words (e.g., "because," "think") peak in Help and Documentation and Error Prevention, pointing to more analytic and instructional tone. Analytic thinking scores are high across all categories but lowest in Help and Documentation (94.6), while User Control and Freedom exhibits the highest Clout (47.7), reflecting confident, assertive language. These patterns suggest that early usability categories favor rich, user-oriented descriptions, whereas later ones use more technical, structured, and formal linguistic styles.

Table 2: LIWC linguistic features.

Heuristic	WC	Analytic	Clout	Authentic	Tone	BigWords	Linguistic	function	Cognition
10	108.21	94.63	28.17	33.00	24.34	39.93	40.22	30.51	16.25
1	273.86	97.64	39.61	30.88	42.44	32.38	53.17	40.53	9.82
3	300.57	97.71	47.68	26.61	24.28	33.64	48.22	38.04	11.59
4	384.70	96.91	41.96	19.77	37.66	35.58	49.00	38.45	10.50
5	68.67	97.74	31.57	35.43	44.70	43.36	42.02	30.06	15.17
6	247.57	96.91	44.24	27.74	44.60	39.75	48.15	35.60	16.04
8	204.86	96.37	37.78	35.93	30.36	38.05	45.29	35.02	13.82

The reports averaged 240 words. The highest word counts appear in Consistency and Standards (384.7) and User Control and Freedom (300.6). The categories have more elaborate descriptions. In contrast, Help and Documentation has the lowest word count (108.2) and concise, directive language. This is supported by lower words per sentence (WPS = 15.7) in Help and Documentation, compared to 23.1 in Visibility of System Status and 22.2 in Consistency and Standards. All categories show high levels of analytic language, with values above 94, but Help and Documentation is slightly lower (94.6) compared to others like Error Prevention (97.7). Clout, suggesting authoritative language, peaks in User Control and Freedom (47.7) and is lowest in Help and Documentation (28.2), indicating a more instructive, less commanding tone in documentation contexts. Error Prevention has the highest authenticity score (35.4), implying more personal or straightforward expression, while Consistency and Standards is lowest

(19.8). Interestingly, Error Prevention also scores highest in tone (44.7), indicating more positive emotional content, whereas User Control and Freedom (24.3) and Help and Documentation (24.3) are neutral or negative in tone.

Table 3: LIWC variables correlation to heuristics.

Variable	Correlation	p-value
WC	–0.508	2.80E-06
Analytic	–0.411	2.25E-04
BigWords	0.446	5.49E-05
Dic	–0.605	6.88E-09
Linguistic	–0.735	4.31E-14
Cognition	0.503	3.55E-06
Cogproc	0.498	4.66E-06
Discrep	0.41	2.39E-04

Regarding correlations between linguistic features and heuristics, word count ($r = -0.508$, $p < .001$) and analytic language ($r = -0.411$, $p < .001$) both show moderate negative correlations. Higher-numbered categories (such as help and documentation) tend to use fewer words and slightly less analytical expression. Similarly, strong negative correlations for dictionary words ($r = -0.605$, $p < .001$) and linguistic function words ($r = -0.735$, $p < .001$) suggest a decrease in standard and grammatical word usage in later categories. Longer words (BigWords) increased with higher category numbers ($r = 0.446$, $p < .001$), as did the use of cognitive language (Cognition: $r = 0.503$, $p < .001$; cogproc: $r = 0.498$, $p < .001$) such as abstract and reflective thinking. Additionally, words that indicate uncertainty or contrast (e.g., "should," "would") were more common in higher categories ($r = 0.410$, $p < .001$). These patterns suggest a linguistic shift from structured, analytical expression in earlier categories to more cognitively complex and abstract language in higher-numbered ones. As heuristic categories increase, (e.g., from visibility to help), there is a rise in language related to thinking processes such as reasoning, insight, or certainty. Higher-numbered heuristics may thus involve more analytical or explanatory text.

CONCLUSION

The study reveals an uneven application of user-centered design principles. Reports emphasized principles such as visibility of system status, user control and freedom, minimalist design, and accessible help documentation. However, other heuristics like error prevention, recognition rather than recall, and flexibility received far less attention. Linguistic patterns also revealed that while early usability categories relied on direct, user-focused language, later categories showed more structured and abstract expression, which may limit accessibility.

These findings extend prior work by offering empirical support for concerns raised in the literature. Pubic trust in AI systems depends on transparency, interpretability, and participatory design (Roundtree, 2023;

Roundtree, 2024). The limited presence of certain heuristics shows that public AI tools often prioritize technical and compliance-based goals over inclusive user experience design (Lu et al., 2024; McKenna-Aspell et al., 2022). This study confirms that current implementations confirm some prior findings—especially about transparency and control—but also challenge the assumption that all usability dimensions are equally considered. Using heuristic-based NLP classification to assess usability language in compliance documents introduces a replicable, scalable way to evaluate human-centered principles. Bridging computational modeling with human-centered evaluation offers a way for public servants, technologists, UX designers, and policymakers to evaluate usability.

This work has potential to improve accountability and fairness in public AI systems. By identifying where usability principles are applied or overlooked, we can inform decisions, tool development and oversight. The findings support the case for leveraging human-centered design to ensure systems are not only functional but also equitable and intelligible to the public.

This study has several limitations. It is based on a relatively small dataset (76 reports) with constrained detail, so underreporting may have skewed the frequency of heuristics. The model's performance was also limited by the sample size. Logistic regression also may not have captured all of the patterns that more complex models could detect. Future research should expand the dataset across different municipalities, government levels, and sectors to validate finding. Incorporating interviews or surveys with system users and developers could enrich understanding of real-world settings. Additionally, future work might explore how participatory and inclusive design can be integrated to study overlooked usability elements such error prevention and cognitive support.

REFERENCES

Bienvenido, H. P., Barinaga, B., & Mora-Fernandez, J. (2021). A Historical Review of Immersive Storytelling Technologies: New Uses of AI, Data Science, qnd User Experience in Virtual Worlds. In Handbook of research on applied data science and artificial intelligence in business and industry (pp. 1–29). IGI Global.

Chanchamnan, P., Ho, C., & San, S. (2022). Design in the age of Artificial Intelligence: A literature review on the enhancement of User Experience Design with AI. IEEE Access.

Gonçalves, M., & Oliveira, A. G. N. A. (2023). The Impacts of AI on Creative Processes in UX/UI Project Development: A Comprehensive Review. IX SIINTEC. IX International Symposium on Innovation and Technology Engineering and the Future of the Industry. https://bit.ly/4k6TZyh

Isaac, S., Phillips, M. R., Chen, K. A., Carlson, R., Greenberg, C. C., & Khairat, S. (2024). Usability, acceptability, and implementation of artificial intelligence (AI) and machine learning (ML) techniques in surgical coaching and training: A scoping review. Journal of Surgical Education.

Kendall Roundtree, A. (2024). Public Perception of AI: A Review. In International Conference on Human-Computer Interaction (pp. 72–87). Springer, Cham.

Lu, Y., Yang, Y., Zhao, Q., Zhang, C., & Li, T. J. J. (2024). AI assistance for UX: A literature review through human-centered AI. arXiv preprint arXiv:2402.06089.

Paracolli, A., & Arquilla, V. (2024). UX Sustainability in AI-infused Objects: A systematic literature review of available tools for Designers. Human Interaction and Emerging Technologies (IHIET 2024), 157, 406–417.

Ramrakhiani, N., & Kalbande, D. (2024). A comprehensive review of AI-powered skincare product recommendation systems: From data collection to user experience. E-Learning and Digital Media, 20427530241304073.

Roundtree, A. (2021, October). Facial Recognition UX: A Case Study of Industry-Academic Partnerships to Promote User-Centered Ethics in Facial Recognition. In Proceedings of the 39th ACM International Conference on Design of Communication (pp. 240–246).

Roundtree, A. K. (2020). ANT Ethics in Professional Communication: An Integrative Review. American Communication Journal, 22(1).

Roundtree, A. K. (2021, June). Ethics and facial recognition technology: An integrative review. In 2021 3rd World Symposium on Artificial Intelligence (WSAI) (pp. 10–19). IEEE.

Roundtree, A. K. (2022, July). Facial recognition technology codes of ethics: Content analysis and review. In 2022 IEEE International Professional Communication Conference (ProComm) (pp. 211–220). IEEE.

Roundtree, A. K. (2023). AI explainability, interpretability, fairness, and privacy: An integrative review of reviews. In International Conference on Human-Computer Interaction (pp. 305–317). Springer, Cham.

Sepanloo, K., Ahmadi Gharehtoragh, M., & Duffy, V. G. (2024, June). Transforming User Experience Through Extended Reality and Conversational AI: A Systematic Review. In International Conference on Human-Computer Interaction (pp.183–194). Cham: Springer Nature Switzerland.

Stige, Å., Zamani, E. D., Mikalef, P., & Zhu, Y. (2024). Artificial intelligence (AI) for user experience (UX) design: a systematic literature review and future research agenda. Information Technology & People, 37(6), 2324–2352.

Zeng, E., Liu, H., Wang, L., Zhao, W., & Feng, Y. (2025). Literature Review: A Study of XAI User Experience in Healthcare: Transparency and Doctor-Patient Trust Construction Based on AI-assisted Diagnosis. Frontiers in Interdisciplinary Applied Science, 2(01), 21–33.