ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

A SYSTEMATIC REVIEW OF PERSISTENT AI INTERFACES IN EMBEDDED SYSTEMS: BRIDGING TECHNICAL PERFORMANCE AND USER EXPERIENCE

ALAN ISAAC TRINIDAD GONZÁLEZ[1], ELENA FABIOLA RUIZ LEDESMA[2]

[1]Department of Graduate Studies, Instituto Politécnico Nacional, Escuela Superior de Cómputo, Mexico

[2*]Department of Graduate Studies, Instituto Politécnico Nacional, Escuela Superior de Cómputo, Mexico

E-mail: [1]atrinidadg1700@alumno.ipn.mx, [2]eruizl@ipn.mx (corresponding author)

ABSTRACT

This systematic review examines the impact of persistent artificial intelligence interfaces in embedded devices on the usability and user experience of adults. We start by evaluating the performance of these systems and then shift focus to subjective experiences, considering factors such as latency, autonomy, and functionality. Following PRISMA guidelines, we searched Scopus and Springer Nature Link for studies published between 2020 and 2025, focusing on a PICO-formulated question about interactions with these "always-ready" embedded devices. Only studies involving local or hybrid computing that reported on user experience or perceived performance were included. The findings consistently show that persistence is achieved through computational proximity, utilizing voice, vision, or other sensory processes for tasks like word recognition, monitoring in public spaces, and intelligent battery management. As part of our contribution, we outline several implications aimed at establishing quality criteria for embedded devices, useful for both design and evaluation. This framework extends beyond mere computational metrics; it provides guidance for enhancing everyday user experiences, addressing operational goals such as maintaining perceptible thresholds of naturalness, ensuring continuity without delays or lags, and preventing overheating during prolonged use. Overall, it proposes methods to "measure what truly matters" without requiring oversized architectures, while incorporating strategies that make persistence perceptible without intruding on or compromising usability.

Keywords: *Artificial intelligence, Embedded devices, Latency, Pervasive interfaces, User experience*.

1. INTRODUCTION

In recent years, the popularity of artificial intelligence assistants has increased significantly, showing that natural human–machine interaction is no longer an unlikely future but a quickly developing market driven by the need for immediacy and growing digital sophistication [1], [2].

One of the ongoing challenges in this field is latency and dependence on models that run entirely in the cloud. Deploying a large model on a microcontroller remains unfeasible. A promising approach involves using "lightweight" embedded devices that can record and pre-process voice locally, while offloading semantic understanding and more advanced processing to a large language model hosted on dedicated servers. Devices like Amazon's Alexa, Apple's HomePod, and mobile assistants such as Siri and Google Assistant [3], [4] exemplify this setup.

As technology continues to mature, discussions about these kinds of integrations have shifted from a primary focus on raw performance to a greater emphasis on the subjective experiences they can offer. Reviews examining proactive behaviors in embedded AI assistants show that sustained adoption depends on anticipating user needs and reducing user effort [5], [6]. Additionally, several studies have proposed design guidelines to foster user trust, facilitate interaction, and improve usability, emphasizing that individual evaluation remains essential for ongoing engagement [7].

Therefore, the current review examined research on how interactions with embedded assistants and smart devices are integrated into daily routines, with particular focus on usability, user experience (UX), acceptance, routines, and emotional aspects.

Regarding usability—specifically ease of use, control, and reliability—findings show that users prefer simple commands, instant feedback, and consistent responses [8], [9], [10]. Akob, 2025; Shade, 2025; Liu et al., 2023]. However, recognition errors and the lack of perceived control remain the primary sources of frustration [11] Becks et al., 2023].

Embedded devices gradually become part of household routines through habituation, during which users discover practical applications (e.g., music, reminders, weather queries, home control) and subsequently develop a stable set of interactions [12]. Additionally, the device's physical placement heavily influences how often it is used. In health and dependency settings, intelligent assistants help preserve independence but need initial guidance to overcome technological challenges [13].

Multiple studies agree that user experience goes beyond just technical skills. For instance, Akob et al. [8] and Oewel et al. [14] highlight that the emotional bond formed with the assistant—often referred to as "friendship," "companionship," or "social presence"—is a crucial factor in determining whether users continue to engage. These assistants can help alleviate feelings of loneliness, enhance perceptions of companionship, and provide emotional support, particularly for older adults and individuals with cognitive challenges. However, some problems have been reported, such as emotional dependence, excessive anthropomorphism, and a loss of control when systems assume too much authority.

Conceptual limitations were also identified in the reviewed literature. Most studies approach usability and UX from a functionalist perspective, emphasizing ease of use, satisfaction, and frequency of use. However, they lack strong theoretical frameworks to explain how these interactions transform daily life and social practices. To advance the field, future work should incorporate perspectives from the sociology of use, the phenomenology of technological dwelling, and critical design theory, thereby clarifying the deeper meanings of living with a persistent intelligent assistant.

Furthermore, little research has focused on designing adaptive persistent interfaces that can learn user habits without compromising privacy, as well as on creating inclusive multimodal systems that combine voice, gesture, visual, and tactile inputs without overwhelming the user.

Based on these reflections, the objective of this systematic review—conducted following the PRISMA methodology [15] and including studies published between 2020 and 2025 in Scopus and Springer Nature Link—was defined as follows: to examine the impact of interactions with embedded assistant devices on everyday practices. Specifically, this study aims to identify usability- and UX-related factors that can inform the design of future integrated interaction devices and improve their practical usefulness.

2. METHODOLOGY

This study followed a structured protocol based on the PRISMA 2020 guidelines [16] and the PICO framework. The protocol defines the research question, search strategy, inclusion and exclusion criteria, screening and selection procedures, data extraction, and synthesis methods, as described below.

2.1 Research Question Formulation

The guiding question was developed using the PICO heuristic (Population, Intervention, Comparison, Outcome):

- Population (P): Adults who use embedded or wearable devices with continuous interaction capabilities.
- Intervention (I): Persistent AI interfaces capable of local or hybrid processing through voice, vision, or sensory input.
- Comparison (C): Conventional interaction modes (manual input, non-persistent systems).
- Outcome (O): Usability, satisfaction, perceived efficiency, and subjective user experience.

The final research question was defined as follows:

In adults who use embedded devices with continuous interaction capabilities, does the integration of autonomous and persistent AI interfaces lead to significant improvements in usability, satisfaction, and efficiency compared to manual or non-persistent operation?

This research question strikes a balance between the precision required for a systematic review and the potential for novel findings that may extend beyond purely quantitative measures or confidence intervals.

2.2 Databases and Time Frame

Two academic databases were selected for their coverage and advanced filtering capabilities:

- Scopus (Elsevier)
- Springer Nature Link

The search included publications from January 2020 to March 2025, a period marked by the rapid growth of edge AI and embedded intelligent systems.

2.3 Search Strategy and Query Syntax

Searches were conducted in both English and Spanish using Boolean operators to maximize coverage. The primary query was as follows:

("embedded device" OR "edge device" OR wearable OR "IoT node" OR "single-board computer") AND ("on-device AI" OR "offline assistant" OR "persistent interface" OR "edge AI") AND (usability OR "user experience" OR satisfaction OR efficiency OR latency).

The main AND connector ensures that records address all three blocks simultaneously, while the OR connector allows flexibility if any of the conditions match.

Filters were applied for:

- Open-access documents
- Peer-reviewed journals or reviews
- English or Spanish language
- Relevance to AI, embedded systems, or human–computer interaction

2.4 Inclusion and Exclusion Criteria

During this stage, the results from each platform were filtered to quickly remove any records that were not relevant to this research.

Table 1: Initial Debugging Flow.

Inclusion Criteria	Exclusion Criteria
Studies published between 2020–2025	Papers published before 2020
Peer-reviewed journal articles or reviews	Conference abstracts, editorials, or patents
Focus on embedded or edge AI systems	Cloud-only AI models
Reports including usability, UX, latency, satisfaction, or efficiency measures	Studies lacking user-centered evaluation
Full-text open-access availability	Paywalled or inaccessible texts

2.3.1 Description of each phase

Initially, the search on Scopus yielded a total of 303 results, while the search on Springer Nature Link returned 537 results.

During filtering, the total was significantly reduced: in Scopus, only 33 results remained. This was mainly due to filters that limited results to open access, specific languages, and subject areas; in Springer Nature Link, only 79 results remained. The decline in the number of documents in this database was primarily caused by the filter restricting results to open-access documents.

Subsequently, during the eligibility phase, each study was reviewed by downloading the .csv files from each database for analysis in Excel. It is worth noting that studies lacking metrics or whose specific topic was not directly related to the research were excluded. Finally, we included a complete set of articles documenting user satisfaction and perceptions of multiple embedded projects.

2.4 Ethical and Open Access Considerations

To ensure robustness, this work is guided by two practical and ethical principles:

2.4.1 Respect for the transparency of knowledge

Only articles with text fully available under an open access license and stored in the DByP Digital Library at the BNCT were reviewed. This way, we prevent the circulation of unauthorized copies and make sure sources are accessible without barriers.

2.4.2. Reproducibility without restriction

The selection process and query syntax are documented so that those pursuing this line of research can reconstruct a path that does not infringe on copyright or closed databases. These guidelines are designed to promote good research practices and the open sharing of texts on embedded AI interfaces, in line with the responsible innovation underlying this article.

3. RESULTS

Below are the results of collecting requirements into a manageable corpus, along with the features shared by the retained studies and those anticipated for further discussion.

3.1 Description of the Selection Process

Once the search strategy was executed in the two repositories, Springer Nature Link yielded 79 results, while Scopus yielded 33, giving us a total of 112 references. The cleaning and gradual application of the PRISMA criteria that led to the final set are shown in the following flowchart (see Figure 1).

Initially, the references from each repository were downloaded into a .csv file, along with the metadata provided by each platform, including authors, year, title, abstract, keywords, and DOI. These two databases compose the top part of the diagram.

The next step involved a screening process that eliminated duplicate articles, unrelated texts, and those that were not open-access. After this selection, 49 articles remained. If any were unavailable, we searched for their abstracts or keywords and evaluated their relevance to the current work. The result of this process was 24 articles.

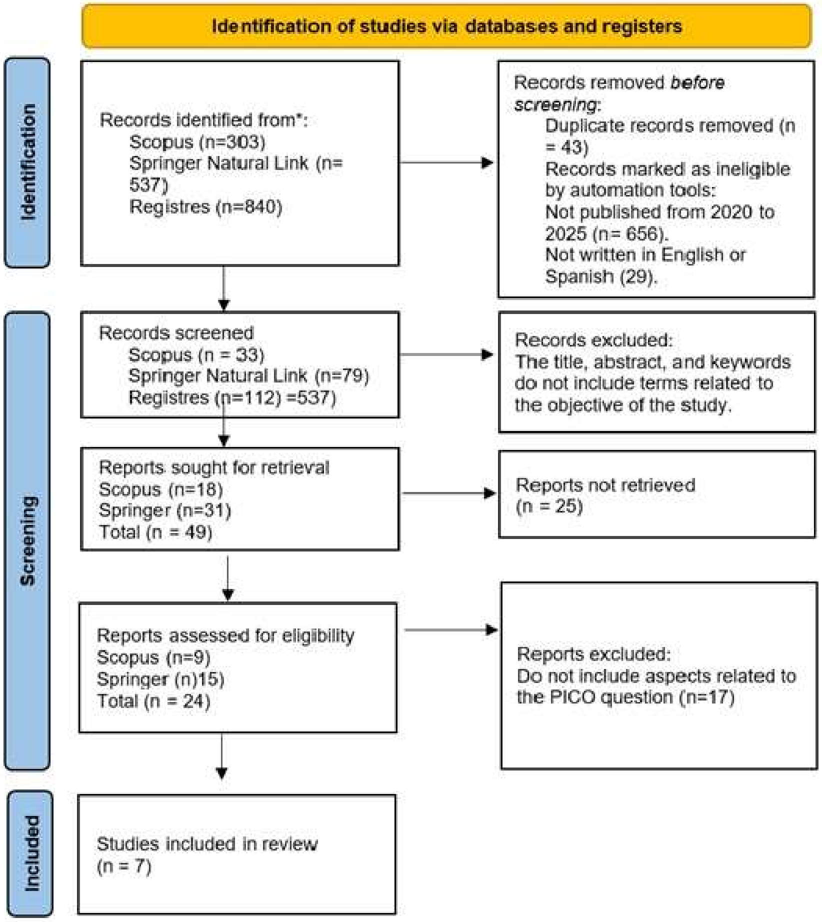
Identification of studies via databases and registers
Identification
Screening
Included
Records identified from*:
Scopus (n=303)
Springer Natural Link (n= 537)
Registres (n=840)
Records removed before screening:
Duplicate records removed (n = 43)
Records marked as ineligible by automation tools:
Not published from 2020 to 2025 (n= 656).
Not written in English or Spanish (29).
Records screened
Scopus (n = 33)
Springer Natural Link (n=79)
Registres (n=112) =537)
Records excluded:
The title, abstract, and keywords do not include terms related to the objective of the study.
Reports sought for retrieval
Scopus (n=18)
Springer (n=31)
Total (n = 49)
Reports not retrieved
(n = 25)
Reports assessed for eligibility
Scopus (n=9)
Springer (n)15)
Total (n = 24)
Reports excluded:
Do not include aspects related to the PICO question (n=17)
Studies included in review
(n = 7)

Figure 1: PRISMA Flow Diagram of the Study Selection Process.

3.2 Comparative Overview of the Included Studies

To complement the characterization of the articles that followed the PRISMA process flow diagram, a comparative table of the seven articles is provided, organized alphabetically by title. It includes relevant data and variables for future analysis, as well as the hardware used, the type of interface, the sample, and the UX method.

In the column for device or hardware type, the platform or device where the model runs is indicated; the kind of persistent AI interface shows how a standby or alert mode is maintained on the device; the samples or UX methods represent both the sample size and the metrics reflecting users' experience; the qualitative findings present the subjective perceptions gathered from the articles; and the limitations are the self-critique reported by the authors.

3.3. Global Synthesis of Findings

Cross-readings reveal three cross-cutting trends and two gaps that are worth keeping in mind before the discussion and interpretation of results:

3.3.1 Diversity of hardware.

- There is variation in the type of platform used across the different studies (Microcontrollers vs. FPGA vs. edge GPU). Still, throughout, there is the same aspiration to reduce latency and provide privacy by processing data on the devices.
- The energy factor appears as a key success metric; the articles report values ranging from 27% to 80% compared to other cloud-first implementations, while noting that the metric remains unpredictable.
- In FPGA use, reconfiguration is valued as a significant benefit because engineers report less cognitive "friction" thanks to PYNQ and overlays [17], something that was not considered possible five years ago.

3.3.2 Persistence of AI

Three works incorporate engines that remain active [18], [19], either in a fuzzy model (fuzzy-logic-based system) or an RL agent (Reinforcement Learning), which reduces manual interaction and creates an intelligent environment.

Users describe the experiences as smooth or natural as long as responses are under 300 ms and do not need corrections.

3.3.3 Maturity in usability and satisfaction metrics

Five of the seven articles include standardized user-experience tools (SUS, NASA-TLX, interviews), representing an increase from the 2020–2022 period.

The depth of the instruments varies: while some combine NASA-TLX with qualitative interviews [18] to assess trust and distraction, other studies are limited to collecting direct empirical user data [17], [20].

Table 2: Comparative Profile Of The Included Studies

No.	Author(s)	Device / Hardware	Type of Persistent AI Interface	Samples & UX Method	Qualitative Findings	Limitations
1	Orășan, et al. [21].	Overview of Cortex-M0/M4/M7 MCUs (≤ 512 kB SRAM).	Focuses on offline toolchains.	Metric extraction (latency, power consumption, footprint) across 10 articles.	Notes that "development convenience" weighs as much as throughput for embedded design teams	Lack of comparable benchmarks and of end-user studies
2	Chen, et al. [24]	Jetson Nano + CSI camera; Raspberry Pi 4 mini-cluster	Continuous AI vision for people counting; quantized model resident on edge GPU	18 librarians; post-task SUS survey	Users value the "smoothness" and zero network dependency during mass events.	Accuracy drops in areas with irregular lighting; study restricted to a single building.
3	Kalapothas, et al. [17]	Xilinx ZCU104 FPGA; PCIe interface to host	Autonomous runtime that keeps models	Laboratory test (no external sample)	Highlights the reduction of engineer-hours	Needs to evaluate developer acceptance

			loaded in on-board DDR		when porting CNNs	and learning curve
4	Wang, et al. [23]	ARM Cortex-M4 (STM32L4) and PULP SoC "Mr. Wolf" (8 RI5CY)	MLP inference resident in flash + SRAM; always-on for gestures/falls	3 apps (gestures, activity, falls); NASA-TLX with 12 volunteers	Users perceive "instant response" (< 20 ms); privacy by processing locally	Toolkit lacks CNN support; UX tests only in the lab
5	Gondi, et al. [23]	Raspberry Pi 4, Coral TPU, Jetson Nano	Local ASR pipeline; wake word active 24/7	20 participants; SUS + task time	Lower "conversational stress" by not relying on the cloud	Only available in the English language
6	Li [18]	Sports band with IMU + PPG (nRF52832 MCU)	Fuzzy engine that evaluates signals every 5 s; local history	19 athletes; interview + NASA-TLX	Coaches value the "interpretability" of fuzzy rules	Small, all-male sample; laboratory conditions
7	Suder, et al. [19]	ARM-M55 smartwatch prototype; Zephyr RTOS	Resident RL agent that adjusts power levels without intervention	Simulation + 8 users in the field	Users notice "longer battery life" without affecting latency	80% of results in the simulator

3.4 Identified Gaps

The authors acknowledge that most tests are conducted in controlled environments, such as laboratories, indoor gymnasiums, and libraries, which offer stable lighting and minimal ambient noise. At the same time, documentation on other variables such as crowds, perspiration, temperature variations, or domestic settings remains limited.
Regarding privacy, only two studies [17], [21] specifically discuss privacy concerns (voice and image capture). These concerns are essential, despite ongoing discussions about persistent models that constantly listen or record.

4. DISCUSSION

This review begins with a challenging yet straightforward idea: if AI is to coexist with us—on wearables, in schools, at workplaces, and in public spaces—it should be persistent, always ready, like a cell phone today, without being intrusive or quickly draining the battery. The seven selected works demonstrate this from different angles—voice, vision, and biosensors—highlighting a key lesson: the quality of the experience depends not only on the final interface but also on a combination of algorithms, hardware, and usage policies. When these elements are balanced, persistence lowers friction in interactions—fewer steps mean less waiting—and builds trust—fewer surprises mean more control. Conversely, when there is an imbalance, it can cause waiting fatigue, privacy issues, or a short battery life, which may weaken reliability and trust.

In [15], the recent shift toward "wake words" and decoders that operate locally is discussed, which are now possible on most devices. Low latency leads to more natural conversations, fewer repetitions, and a stronger sense of control. The "persistent" label does not mean listening without a time limit, but instead activating under specific rules, such as through visible signals (LED, brief tone) and immediate silence for proper activation.

The study in [16] examines the use of computer vision to monitor the number of people entering a library continuously. The focus is not only on achieving high accuracy but also on implementing the solution effectively. The authors choose a lightweight backbone that maintains steady frame rates, along with scene cropping and appropriate resolutions to identify individuals as counting units without needing personal identification. This method emphasizes the social acceptability of technology. Moreover, Kalapothas et al. [17] support the technical aspect by demonstrating that, in scenarios with multiple tasks running concurrently, deploying a DPU on platforms such as Kria, Vitis, or PYNQ ensures consistent frame-by-frame processing. From the user's perspective, this leads to an improved experience, as the model runs smoothly without interruptions or noticeable delays, even though the computational workload within the BRAM remains unseen.

Regarding biosensors, [17] introduces a "user-friendly" form of continuous data collection that operates based on predefined rules. For example, when a stride pattern changes or decreases, a subtle alert is triggered. This feature is significant in systems and devices meant for everyday use, as their value lies not only in performing tasks accurately but also in providing clear and reasonable explanations for their actions.

A similar method is employed in SmartAPM [18], where persistence is achieved through intelligent policies—such as DVFS, duty cycling, and reinforcement-based adaptation—that enhance battery life without compromising system performance. From the user's point of view, this improvement is felt practically, shown by comments like "the battery lasts all day," turning autonomy from just an expectation into a real, tangible benefit.

Quantization, kernels, and toolchains like X-CUBE-AI are not just optimizations but what make the interface feel immediate, stable, and private [19]. A model that infers data in milliseconds and does not send raw data to the cloud results in shorter wait times, fewer doubts, and less user configuration [19].

From this overview, four ideas stand out:

1. Governed persistence. Being "always ready" doesn't mean always being active. Conditioned activation (voice, gesture, threshold) protects user privacy while promoting efficient resource use.
2. Energy is vital to user experience. The autonomy of current devices is more than just a key performance indicator; it lays the foundation for forming habits. SmartAPM utilizes this approach by creating a strategy to identify usage patterns based on context, ensuring persistence without relying on a potentially vulnerable "power-saving mode" that could be exploited by resource-intensive functions.
3. Operational privacy. Privacy, more than just a legal requirement, must be integrated into the system design, with edge processes that limit data retention to what is necessary. Use controls such as shutdown, local history deletion, and logs, and clearly communicate how they are implemented.
4. Error recoverability. Any system aiming for persistence is vulnerable to false positives. The goal should not be to promise perfection but to provide feedback and learn locally from indicated corrections.

4.1 Interpretation of results

The research question that guided this study examined whether persistent AI interfaces in embedded devices improve usability, satisfaction, and efficiency compared to manual interaction. Based on the review of the selected articles, the most appropriate conclusion is affirmative, but only under specific conditions, which are detailed below:

Reducing the effort needed to interact with such systems is a clear benefit. Simply using a wake word, sensor, or vision to initiate a sequence of tasks makes the interaction feel more natural—shorter and with fewer repetitions.

The emphasis is on consistency rather than perfect accuracy. In computer vision, for instance, lightweight models like YOLOv8n prioritize maintaining steady frame rates, which helps keep quality stable, even if that isn't the primary goal of the system [16].

Regarding autonomy, this directly affects the user experience. Policies such as DVFS, duty cycling, and reinforcement-based adaptation (as used in SmartAPM) build trust, as users no longer need to constantly monitor battery levels and start to view the wearable as a seamless part of their daily routines [19].

Privacy can be maintained even during continuous operation. A common concern is the amount of data recorded and how it is transmitted, particularly in voice-based systems. This issue is addressed by edge-based processing (activation control), which limits recording to only what is necessary, reduces data retention, and thereby increases public trust.

Furthermore, systems can achieve continuous error recovery through immediate feedback, undo functions, and lightweight local learning models that gradually turn errors into adjustments. An example is Real-Time Athlete Fatigue Monitoring Using Fuzzy Decision Support

Systems [17], where fuzzy logic enables graded decisions that adapt over time through use.

How does this translate into usability, satisfaction, and efficiency?

In voice assistants: fewer steps, fewer corrections, and higher System Usability Scale (SUS) scores during everyday use.

In vision applications, a stable persistence flow without requiring consumption peaks ensures reliability during continuous operation.

In biosensor systems, context-aware alerts that build trust and minimize false alarms are combined with effective battery management to ensure continuous operation throughout the day.

Together, this review highlights how persistent interfaces in embedded systems can significantly enhance the quality of life for adult users, provided they are designed with transparent governance, optimal autonomy, operational privacy, and error recovery capabilities. The seven studies examined not only support this idea but also suggest a direction for developing further technological solutions that can positively affect daily life.

4.2 Implications for the design of embedded devices

Bringing the previously discussed evidence into the design domain requires understanding "persistence" not merely as an active state, but as a continuous cycle of perception, decision-making, and action constrained by architecture, energy, and user trust. From the seven reviewed studies, concrete implications emerge that may guide hardware architects, engineers, and UX designers in the development of AI interfaces.

4.2.1. Governance of persistence

A persistent interface must clearly indicate when it remains active and provide controls to pause, mute, or turn off. In voice control, this translates to using wake words, listening confirmations (such as an LED or slight vibration) that do not interrupt the current task. In biosensing systems, persistence should be limited to a specific time window (like exercise sessions) and supported by notifications that issue alerts for rate variability, helping to reduce false positives or anxiety, and include a local history of notifications.

4.2.2. Natural interaction.

It is important to note that unjustified persistence can erode trust. Measures should be implemented to allow for gradual interaction, management of preferences, and advisable pauses or silent periods to emulate natural conversation. In public-space monitoring, counts, density, and trends should be prioritized through the use of privacy controls.

4.2.3. Privacy from the architecture.

Keep local processing of wake phrases or use anonymous aggregates if sending data to the cloud is necessary. In vision devices, apply masking techniques before transmission or retain only minimal data. The goal is to treat privacy as a functional requirement.

4.2.4 Energy equals experience.

Autonomy is a usability metric; proper energy management maintains consistent performance during continuous use without the need for improvised recharges. Generally, systems with fewer shutdowns due to power issues tend to generate more trust.

4.2.5. Graceful degradation.

User satisfaction remains high, even in the face of failures, as long as the system does not crash. This is accomplished through simple rules (thresholds) or by disabling certain functions without a complete shutdown. Always strive to include a manual shortcut so users can regain control if a failure occurs; this method will be seen as trustworthy, even if it does not fully succeed, because it fails gracefully.

4.2.6 Deployments and updates

For end users, maintaining safe and reversible updates improves usability because, in the event of failures or incompatibilities, they do not need to reinstall everything. Additionally, communicate clearly what improvements are included with each update.

4.2.7 Accessibility first

Optimizing the system can become complex when adapting to different environments that may be noisy, unclear, lacking visibility, or involve diverse accents. Offering multiple input options—such as text, gestures, or buttons—can enhance usability across various situations.

These implications, viewed through the lens of usability, experience, and functionality, serve as guidelines for product architecture that are currently feasible based on the evidence from the seven analyzed studies.

4.3 Classification of Contribution

Compared to recent studies on trust, privacy, and usability in conversational and embedded AI systems [5]–[7], this work mainly contributes at the methodological and integrative levels. Methodologically, it broadens the use of the PRISMA protocol—traditionally applied in medical and behavioral sciences—to the field of persistent AI interfaces in embedded devices. This adaptation establishes a reproducible framework for evaluating usability and user experience (UX) metrics across diverse technological contexts, ranging from wearables and edge GPUs to FPGA-based systems.

From an integrative perspective, this research consolidates empirical findings from isolated subfields—voice, vision, and biosensing—into a unified analytical framework that connects latency, autonomy, and persistence to experiential outcomes such as trust, perceived control, and satisfaction. Unlike earlier reviews that narrowly focused on user trust [5] or privacy perceptions [6], this work highlights the interdependence between technical persistence and experiential continuity, arguing that energy management, local processing, and transparency are not just engineering concerns but key factors shaping usability.

In this context, the contribution of the current study lies in bridging the gap between hardware efficiency and human-centered evaluation, providing design-focused insights that can guide both AI system developers and human–computer interaction (HCI) researchers. Therefore, the research can be viewed as a hybrid contribution—both methodological, through its systematic review approach and analytical synthesis, and practical/theoretical, by proposing design principles for future persistent embedded systems.

4.4 Differences with Respect to Previous Research

While earlier studies on conversational and embedded AI systems mainly focused on user trust, privacy perceptions, or single-domain usability [5]–[7], this review differs in both scope and analytical approach. Previous reviews, like those by Bach et al. [5] and Leschanowsky et al. [6], highlighted psychological factors influencing user trust and risk perception but did not include technical performance aspects such as latency, autonomy, or local processing efficiency. In contrast, this research combines these technical elements with experiential factors to analyze how persistence—defined as continuous and context-aware operation—affects usability and user satisfaction.

Furthermore, unlike Gebru et al. [1], who examined human–machine trust from a conceptual standpoint, the current study systematically reviews empirical evidence across various modalities (voice, vision, and biosensing) using the PRISMA protocol [15]. This approach enables cross-domain comparisons that connect hardware configurations with measurable user experience outcomes, such as perceived control, reduced effort, and sustained satisfaction.

Another key difference lies in treating energy management and autonomy as essential parts of the user experience. Previous works often viewed energy efficiency as a technical limit; here, it is redefined as a factor that influences usability, shaping user trust and routine use. Additionally, a few past reviews proposed design implications that connect persistence, transparency, and ethical governance in embedded settings. This study, therefore, moves beyond simple summaries to provide design-focused insights, bridging the gap between hardware efficiency and human-centered interaction design.

Finally, this review introduces the concept of governed persistence, which views persistent AI not as continuous activity but as conditioned readiness controlled by user input, contextual awareness, and privacy boundaries. This idea represents a theoretical advancement over earlier studies that linked persistence to continuous operation, providing a more nuanced understanding of how ongoing AI can coexist with human agency in everyday settings.

This research presents an integrative perspective that links technical performance with human experience in persistent AI interfaces. It introduces the original idea of "governed persistence," which has not been explored in earlier studies, defining persistence as a contextually controlled state of readiness rather than continuous activity. Methodologically, it advances by adapting the PRISMA protocol to analyze embedded AI systems, allowing a structured and reproducible synthesis across various technologies.

However, the study also acknowledges several limitations: the small number of empirical works available, the heterogeneity of methodological approaches, the absence of long-term or longitudinal evaluations, and a geographical bias toward studies conducted in North America, Europe, and East Asia. These constraints underscore the need for more comprehensive, standardized, and cross-cultural investigations to enhance future research on usability and user experience in persistent embedded AI systems.

5. CONCLUSION

The primary goal of this systematic review was to examine how interactions with embedded assistant devices influence daily routines, with an emphasis on usability and user experience (UX) factors that can inform the design of future persistent AI systems. This goal has mainly been achieved through a structured analysis of seven empirical studies published between 2020 and 2025, analyzed using the PRISMA methodology.

The review demonstrates that persistent AI interfaces—those capable of remaining ready and aware of context without requiring constant user input—enhance usability and satisfaction when managed with transparent governance, optimized energy use, and local processing that safeguards privacy. In particular, data from speech, vision, and biosensing systems demonstrates clear improvements in perceived control, effort reduction, and trust when latency remains below perceptual thresholds and when persistence is managed rather than unconditional.

The findings also extend previous research by framing autonomy and energy efficiency as experiential factors, not just technical performance measures. From a design perspective, the proposed framework of governed persistence offers a practical approach to incorporating persistent AI into daily life without compromising user agency or comfort.

However, several limitations should be acknowledged. First, the set of reviewed studies is small (n = 7) and uses different methods, which restricts the ability to generalize the results. Many experiments were conducted in controlled lab environments—quiet settings, stable lighting, and short testing periods—so their relevance to real-world situations remains uncertain. Additionally, not all studies employed standardized measures of usability or cognitive load, making it hard to compare results across different fields.

In response to the question posed in the introduction to this document, the following can be said:

The integration of autonomous and persistent AI interfaces in embedded devices greatly enhances usability, user satisfaction, and efficiency compared to manual or non-persistent operation. These benefits are achieved when persistence is managed rather than continuous, balancing autonomy with user control, privacy, and energy management.

However, the effect is context-dependent: results were strongest in controlled or semi-structured environments, while long-term, real-world evaluations remain scarce. Future work should validate these findings through longitudinal studies, standardized UX metrics, and cross-cultural analyses.

Future research should focus on expanding long-term and real-world assessments of persistent AI systems, exploring cross-cultural differences in user perception, and creating comprehensive UX metrics that include emotional, ethical, and contextual factors. Addressing these gaps will enhance the external validity of upcoming studies and promote the development of embedded AI technologies that are not only efficient and autonomous but also socially and psychologically sustainable.

6. REFERENCES

[1] B. Gebru, L. Zeleke, D. Blankson, M. Nabil, S. Nateghi, A. Homaifar, E. Tunstel, "A review on human–machine trust evaluation: Human-centric and machine-centric perspectives", *IEEE Transactions on Human-Machine Systems*, Vol. 52, No. 5, 2022, pp. 1–11. https://doi.org/10.1109/THMS.2022.3144956

[2] S. Reddy, D. Chen, and C. D. Manning, "Coqa: A conversational question answering challenge. Transactions of the Association for Computational Linguistics", Vol. 7, 2019, pp. 249–266.

[3] S. Yang, J. Lee, E. Sezgin, B. Jeffrey, S. Lin, "Clinical advice by voice assistants on postpartum depression: cross-sectional investigation using Apple Siri, Amazon Alexa, Google Assistant, and Microsoft Cortana", *JMIR mHealth and uHealth*, Vol. 9, No. 1 2021, e24045.

[4] Z. Xiao, S.Mennicken, B. Huber, A. Shonkoff, and J. Thom, “Let Me Ask You This: How Can a Voice Assistant Elicit Explicit User Feedback?”, *Proceedings ACM Human-Computer Interaction,* University of Wisconsin-Madison (United States), October 18-21, 2021, pp. 1-24. https://doi.org/10.1145/3479532

[5] T. A. Bach, A. Khan, H. Hallock, G. Beltrão, and S. Sousa, “A systematic literature review of user trust in AI-enabled systems: An HCI perspective”, *International Journal of Human–Computer Interaction*, Vol. 40, No. 5, 2024, pp. 1251-1266. https://doi.org/10.48550/arXiv.2304.08795

[6] A. Leschanowsky, S. Rech, B. Popp, and T. Bäckström, “Evaluating Privacy, Security, and Trust Perceptions in Conversational AI: A Systematic Review”, *Comput. Hum. Behav.* Vol. 159, No. 1, 2024, pp. 1-39.

[7] A. Baughan, X. Wang, A. Liu, A. Mercurio, J. Chen, and X. Ma, “A mixed-methods approach to understanding user trust after voice assistant failures”, *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, (Hamburg, Germany), April 19-22, 2023, pp. 1-16.

[8] Akob, D., Wilhelm, S., Gerl, A., Ahrens, D. und Wahl, F. (2025). Adapting Voice Assistant Technology for Older Adults: A Comprehensive Study on Usability, Learning Patterns, and Acceptance. Digital , 5 (1), 4. https://doi.org/10.3390/digital5010004

[9] Shade M, Yan C, Jones VK, Boron J. (2025). Evaluating Older Adults' Engagement and Usability With AI-Driven Interventions: Randomized Pilot Study, 9, e64763. doi: 10.2196/64763.

[10] Liu, M., Wang, C. & Hu, J. (2023) Older adults’ intention to use voice assistants: Usability and emotional needs. Heliyon 9, 1-17. https://doi.org/10.1016/j.heliyon.2023.e21932

[11] Becks, E.; Zdankin, P.; Matkovic, V.; Weis, T. Complexity of Smart Home Setups: A Qualitative User Study on Smart Home Assistance and Implications on Technical Requirements. Technologies 2023, 11, 9. https://doi.org/10.3390/technologies11010009

[12] F. Bentley, CH, Luvogt; M. Silverman, W. Rushani, B. White & D. Lottridge, Understanding the Long-Term Use of Smart Speaker Assistants Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, Volume 2, Issue 3 Article No.: 91, Pages 1 – 24 https://doi.org/10.1145/3264901

[13] M. Nunez; P. Patel; L. Ulin; L. Kian; M. Cominsky; J. Burnett, & J.L. Lee, Feasibility and Usage of a Virtual Assistant Device in Cognitively Impaired Homebound Older Adults. Journal of Applied Gerontology 2025, Vol. 44(10) 1651–1660.

[14] B. Oewel; T. Ammari; R.N. Brewer, Voice Assistant Use in Long-Term Care, Conference: CUI '23: ACM conference on Conversational User Interfaces, DOI:10.1145/3571884.3597135

[15] M. J. Page, J. E. McKenzie, P. M. Bossuyt, I. Boutron, T. C. Hoffmann, C. D. Mulrow, L. Shamseer, J. M. Tetzlaff, E. A. Akl, S. E. Brennan, R. Chou, J. Glanville, J. M. Grimshaw, A. Hróbjartsson, M. M. Lalu, T. Li, E. W. Loder, E. Mayo-Wilson, S. McDonald, L. A. McGuinness, L. A. Stewart, J. Thomas, A. C. Tricco, V. A. Welch, P. Whiting, and D. Moher, “The PRISMA 2020 statement: an updated guideline for reporting systematic reviews,” *BMJ*, Vol. 372, 2021, p.1. 1-71. doi: 10.1136/bmj.n71.

[16] E. Schardt, M. B. Adams, T. Owens, S. Keitz, and P. Fontelo, “Utilization of the PICO framework to improve searching PubMed for clinical questions,” *BMC Medical Informatics and Decision Making*, Vol. 7, No. 1, 2007, pp. 1-16. doi: 10.1186/1472-6947-7-16.

[17] S. Kalapothas, G. Flamis, and P. Kitsos, “Efficient Edge-AI Application Deployment for FPGAs”, Information, Vol. 13, No. 6, pp. https://doi.org/10.3390/info13060279

[18] A. Li, “Real-Time Athlete Fatigue Monitoring Using Fuzzy Decision Support Systems”, International Journal of Computational Intelligence Systems, Vol. 18, No. 1, 2025, pp. 1-23. https://doi.org/10.1007/s44196-025-00732-8

[19] R. Sunder, U. K. Lilhore, A. K. Rai, E. Ghith, M. Tlija, S. Simaiya, and A. H. Majeed, “SmartAPM framework for adaptive power management in wearable devices using deep reinforcement learning”, Scientific Reports, Vol. 15, No. 1, pp. 6911. https://doi.org/10.1038/s41598-025-89709-3

[21] I. Orăşan, C. Seiculescu, and C. D. Căleanu, “A Brief Review of Deep Neural Network Implementations for ARM Cortex-M Processor. Electronics, Vol. 11, No. 16, 2022, pp. https://doi.org/10.3390/electronics11162545

[22] X. Wang, “FANN-onMCU: An Open-Source Toolkit for Energy-Efficient Neural Network Inference at the Edge of the Internet of Things.

(s. f.). ResearchGate. https://doi.org/10.48550/arXiv.1911.03314

[23] S. Gondi, and V. Pratap, "Performance Evaluation of Offline Speech Recognition on Edge Devices", *Electronics*, Vol. 10, No. 21, 2021, https://doi.org/10.3390/electronics10212697

[24] Z. Chen, X. Xie, T. Qiu, and L. Yao, "Dense-stream YOLOv8n: A lightweight framework for real-time crowd monitoring in smart libraries", *Scientific Reports*, Vol. 15, No. 1, 2025, pp. 116-128. https://doi.org/10.1038/s41598-025-94659-x