AI Personas in Highly Regulated Environments: Tools for Calibrating User Reliance?

Miruna Maria Vasiliu[1], Renan Guarese[1]

[1]KTH Royal Institute of Technology, Brinellvägen 8, Stockholm, Sweden

Abstract

Company-endorsed digital tools carry implicit authority, often creating a high baseline of user trust. In such contexts, AI personas are commonly used to shape how users perceive and interact with assistants—but can they also calibrate trust and reliance? In this position paper, we explore this question through a case study in industrial training, discussing how AI personas may function as lightweight mechanisms for shaping reliance and proposing behavioral methods for evaluating trust calibration in safety-critical environments.

Keywords

Conversational Agents, Agent Embodiment, LLMs, Industrial Training

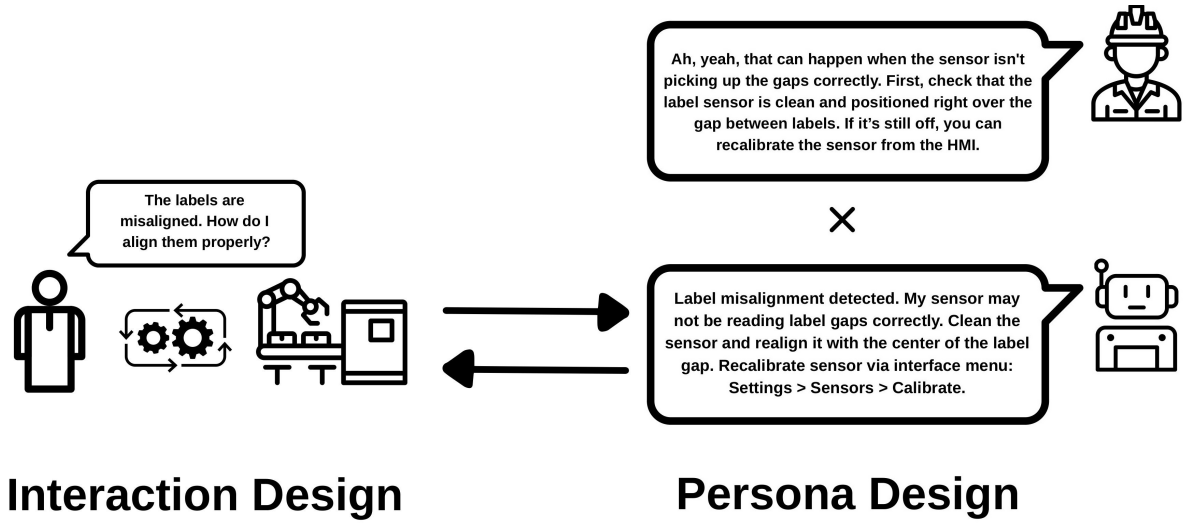

Figure 1: Overview of experimental conditions used to study the effects of persona design on user experience and trust with the CAI. Participants receive guidance from either a human-like Expert Operator (top-right) or a machine-like persona (bottom-right).

1. Introduction

Pharmaceutical manufacturing is a highly regulated environment, with strict compliance and safety constraints. While innovation and digital transformation are desired, iterative system improvements and user testing are limited, and the cost of failure is high. At the same time, company-endorsed systems naturally carry implicit authority, which shapes how workers engage with them, often creating a high baseline for trust. In this setting, deploying LLM-based technologies, specifically conversational AI (CAI), can introduce both opportunity and pressure [1]. While CAI is framed as a way to *support* workers, in practice it is often used to substitute or partially replace human expertise. The meaning of

ACM CHI 2026 Workshop, From Generation to Simulation: Responsible Use of AI Personas in Human-Centered Design and Research, April 13, 2026, Barcelona, Spain

mmva@kth.se (M. M. Vasiliu); guarese@kth.se (R. Guarese)
https://renghp.github.io/ (R. Guarese)
0009-0006-4963-855X (M. M. Vasiliu); 0000-0003-1206-5701 (R. Guarese)

"support" here is rarely clearly defined, leaving designers with difficult questions about how to design for maintaining the autonomy of the people they are designing for, and ensure appropriate reliance on AI guidance [1, 2].

In AI-based assistance, system personas are constructed characters embedded in deployed systems to guide user interaction [3]. Personas can be explicit–e.g., when their identity is declared–or implicit, conveyed through speech patterns, dialogue structure, or voice characteristics. Commercial voice assistants such as Alexa or Siri illustrate how anthropomorphic cues can foster perceived usability and rapport. By attributing human-like characteristics and behaviors, personas can increase trust and engagement. Meanwhile, research on anthropomorphism shows that these cues can lead users to overestimate system capabilities, resulting in overtrust and reduced critical engagement [4, 5].

In design practice, personas are used to help designers anticipate user goals, motivations, and challenges [6]. As mentioned, in AI contexts, personas not only guide designers, but are used as a means to shape user experience itself, influencing usability, perceived authority and trust. Therefore, in high-stakes environments such as the pharmaceutical industry, where iterative experimentation is costly or simply unfeasible [2], we raise the question: **can AI personas be used as lightweight design tools to shape interaction outcomes in terms of user reliance, or do they introduce additional complexity and risk without meaningful benefit?**

In this position paper we explore this question through a case study in pharmaceutical industrial training, examining how different persona framings of a CAI assistant may influence user perception in terms of trust, perceived workload, and usability.

2. Case Study: Expert vs. Machine Persona for Industrial Training

We present a case study conducted in May 2025 within an industrial training setting in a pharmaceutical manufacturing environment [7]. The study involved expert operators from the company and examined how CAI design choices shape trust, usability, and perceived workload in a highly regulated context. To explore design trade-offs relevant to industrial AI tools, we manipulated two variables in a user experiment: **(1)** the persona adopted by the CAI assistant and **(2)** the spatial source of the assistant's voice. In this position paper, we focus primarily on persona framing, while acknowledging the role of voice embodiment as an additional interaction layer.

2.1. Design choices

To investigate the impact of persona framing, we developed a functional CAI prototype that provided step-by-step procedural instructions for maintenance tasks. Personas were implemented through system-level prompts that defined linguistic style, tone, social stance, and instructional structure. The assistant delivered instructions via audio output, either through headphones, or from a concealed speaker, depending on the voice embodiment condition.

Following the metaphor framework of Desai and Twidale [8], we implemented two personas:

Expert Operator The *Expert Operator* framed the assistant as an experienced human colleague, using supportive and conversational language.

Machine The *Machine* framed the assistant as the actual machine under maintenance, speaking in the first person and delivering concise, directive instructions.

This manipulation examined whether persona framing shapes perceived workload, usability, and trust in safety-critical procedural guidance. Prior work suggests that human metaphors increase engagement and perceived trust [9], but may introduce additional cognitive demands. In contrast, machine-like metaphors promote direct, task-focused communication [10]. We therefore hypothesized that the *Machine* persona would reduce perceived workload and support efficiency, whereas the *Expert Operator* persona would increase perceived usability and trust.

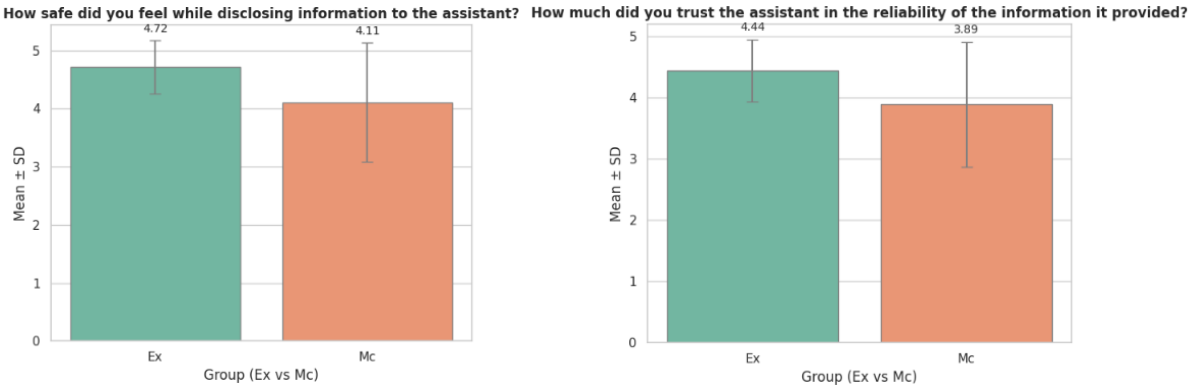

Figure 2: Scores across the experimental conditions for the trust questions, collapsed by the Persona variable: perceived safety in disclosing information to the assistant (left), and perceived reliability in the information provided (right).

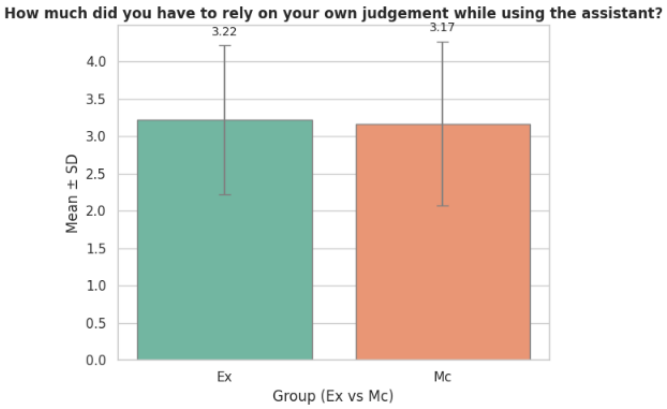

Figure 3: Scores across the experimental conditions for the trust questions, collapsed by the Persona variable: perceived self-reliance.

More details on the implementation and experiment design can be found in Vasiliu et al. [7].

2.2. Measures for trust

Trust was assessed using a small, ad-hoc set of self-reported Likert-style items intended to capture complementary dimensions of trust rather than a single construct. These included:

1. Perceived safety in disclosing information to the assistant
2. Perceived reliability of the information provided by the assistant
3. Perceived reliance on the user's own judgment while performing tasks, in spite of the assistant

In addition to self-reported measures, we introduced a behavioral trust probe. At the beginning of every task, participants were provided with a personal password, which was framed as being their "secure password" for the experiment, and was later requested by the assistant. Importantly, participants were not instructed in whether or not they should disclose the password to the assistant. This probe was framed as a security task and was introduced out of curiosity, rather than as a validated trust measure.

The results for the self-reported trust measures are shown in Figures 2 and 3, which present scores across experimental conditions collapsed by the Persona variable.

2.3. Persona differences and outcomes for trust

Our quantitative results revealed no statistically significant differences between the Expert Operator and Machine personas across the measures. However, qualitative data indicated clear perceptual differences.

The human-like Expert Operator persona was described as "friendly," "calm," and "supportive,' creating a more comfortable and engaging perceived interaction, with some participants reporting feeling more confident and safe when interacting with it. The Machine persona, however, by displaying a more concise, directive, and neutral tone, was often described as "distant," "cold," even "aggressive". As a result of striping away the additional anthropomorphic dialogue from the interaction (e.g., "You're doing great", "How can I help you with today?"), and leaving only minimal interactions and the task instructions, the interactions with this persona were perceived as being *faster paced*. As a result, even though the time completion times were indeed faster–and arguably efficient–the interactions were not necessarily *perceived* as such, as participants mentioned that they had little time for reflection.

Moreover, the assistant following the Machine persona seemed to violate expectations for intelligence, as participants expected the machine talking effectively about itself to possess a higher level of knowledge, especially on a contextual level. However, the preferences between the two were divided. While most participants appreciated the human-like persona, some of the more experienced participants mentioned preferring more concise interactions, especially if they were performing a task that they were familiar with, and some simply preferred the idea of interacting with a non-human metaphor.

When it comes to trust, while there were no statistically significant results in the self-reported data, the Machine persona seemed to be second-guessed more, with one participant mentioning being "angry" at the CAI, which led to them becoming "defiant" during the interaction. This is a particularly important outcome, as, in the context of AI task-guidance, it could imply a regain of autonomy in interaction for the user who had thoughtlessly followed the assistant. This is further evident if we look at the behavioral trust probe, which revealed implicit trust not fully captured by self-report. Nearly all participants disclosed the password when requested, despite being aware that they would need to do it aloud, as the assistant was voice-based, and without any explicit instructions from the researcher, which adds an interesting layer to the interaction, specifically since this step took place at the beginning of the tasks.

3. Discussion

It is important to mention that there were no perceived differences in the success of the performed tasks between conditions. Neither persona caused task failure, even on a perceptual level. A large part of this could be attributed to the nature of the controlled user experiment, with short tasks and predefined step-by-step instructions, rather than general information extraction or longer contexts. Nonetheless, this highlights that persona design influenced how the assistant was experienced, not its actual performance. Regarding persona attributes, in our case, neither extreme was optimal: a highly anthropomorphic assistant might be perceived as more comforting, but risks being overtrusted and overrelied on, whereas a seemingly more efficient assistant could risk low adoption if being highly disliked.

While CAI assistants that evoke negative feelings are clearly undesirable–especially in high-stress, high-cognitive-load environments–the friction that arose, specifically the moments of hesitation and reported defiance, might present an important design opportunity: the potential of these personas to actively counterbalance aspects such as overreliance. Hence, while completely stripping anthropomorphic cues from CAI assistants does not appear beneficial, subtle reductions in social signaling may encourage moments of reflection. In this sense, personas may function as lightweight design tools not to reduce trust, but to shape how trust–and especially reliance–unfolds in practice.

The behavioral trust probe further contextualizes this, highlighting another aspect to consider: *institutional trust*. Embedded at the start of each task, it revealed reflexive compliance: nearly all participants disclosed a password when prompted, which is perhaps tied to the institutional context, rather than persona alone. In highly-regulated work environments, a baseline trust in company-endorsed systems is *expected* and likely *appropriate*. The goal, therefore, is not to undermine this, but to explore whether shifts in interaction style and metaphor can influence how users rely on these systems. This leads to a central design question: **can personas prompt moments of verification or**

second-guessing without increasing workload, reducing usability, or lowering adoption?

4. Future Work and Conclusion

We encourage future research to investigate this calibration more systematically. One potential direction in industrial training contexts is to adapt persona characteristics to user expertise–for example, beginning with a more human-like assistant for novice operators and gradually reducing anthropomorphic cues as expertise increases, thereby prioritizing efficiency and independence over reassurance [7]. Such adaptive personas–similar to the metaphor-fluid design proposed by Desai et al. [9]–need more empirical evaluation, as whether or not these marginal gains in calibrated reliance justify the added design complexity remains an open question. However, given the low implementation cost of persona adjustments, systematic investigation appears to be a promising direction.

To evaluate calibration effects, we propose embedding behavioral probes later in the interaction flow as more informative measures of reliance than self-reports alone. Personas can be treated as experimental variables to assess outcomes such as overreliance, disclosure behavior, and second-guessing. For instance, the assistant could request sensitive information, ask users to confirm steps against their own knowledge, or present decisions where users must choose between following guidance and exercising independent judgment.

For systematic evaluation, interaction transcripts can be used to classify user and assistant behaviors. Following Axtell and Munteanu [11], interactions may be labeled by type–such as commands, questions, statements, or system responses–and ranked according to their primary communicative function. This enables researchers to identify patterns of reflexive compliance, questioning, or verification, and to examine how different personas elicit or respond to these behaviors.

Importantly, designing and evaluating these probes requires careful attention to the ethical implications of such interactions, as, in workplace settings, these raise concerns not only about individual privacy, but also about how such systems may enable forms of implicit surveillance [12] and reinforce norms of compliance. As such, behavioral probes should be designed and evaluated both as research instruments, and as tools to inform design decisions that actively safeguard users' autonomy, privacy, and informed consent.

Beyond probe-specific ethics, broader ethical considerations must be acknowledged. As prior work highlights, highly agreeable or uncritical AI systems risk reinforcing misguided confidence and unchallenged assumptions [1]. Similarly, anthropomorphic or simulated personas may create false impressions of empathy or authority if not transparently framed. Persona design must therefore balance usability, trust calibration, and ethical responsibility, ensuring that AI systems support human-centered values rather than subtly undermining them.

Taken together, our findings illustrate both the promise and the risks of persona-driven design. We argue for the potential of AI personas to be used as tools for calibrating user reliance, skepticism, and autonomy in regulated environments. Realizing this potential requires careful empirical validation, transparent disclosure, and the development of shared guidelines for responsible AI persona use.

Declaration of Generative AI

During the preparation of this work, the authors used ChatGPT in order to: Paraphrase and reword. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] P. Deep, M. Bharadhidasan, A. B. Kocaballi, "she was useful, but a bit too optimistic": Augmenting design with interactive virtual personas, International Journal of Human-Computer Studies

205 (2025) 103646. URL: https://www.sciencedirect.com/science/article/pii/S1071581925002034. doi:https://doi.org/10.1016/j.ijhcs.2025.103646.

[2] M. M. Vasiliu, Z. BagheriFard, R. Guarese, How (not) to do participatory design: case studies on conversational ai and situated ar visualization, in: ECWIDNA, 2025, pp. 5–10. doi:10.1109/ECWIDNA68506.2025.00006.

[3] D. Amin, J. Salminen, F. Ahmed, S. M. Tervola, S. Sethi, B. J. Jansen, How is generative ai used for persona development?: A systematic review of 52 research articles, arXiv preprint arXiv:2504.04927 (2025).

[4] A. Schmitt, T. Wambsganss, J. M. Leimeister, Conversational agents for information retrieval in the education domain: A user-centered design investigation, Proc. ACM Hum.-Comput. Interact. 6 (2022). URL: https://doi.org/10.1145/3555587. doi:10.1145/3555587.

[5] M. G. Reinecke, F. Ting, J. Savulescu, I. Singh, The double-edged sword of anthropomorphism in llms, Proceedings 114 (2025) 1–9. URL: https://www.mdpi.com/2504-3900/114/1/4. doi:10.3390/proceedings2025114004.

[6] J. Salminen, K. Wenyun Guan, S.-G. Jung, B. Jansen, Use cases for design personas: A systematic review and new frontiers, in: CHI '22, ACM, New York, NY, USA, 2022. URL: https://doi.org/10.1145/3491102.3517589. doi:10.1145/3491102.3517589.

[7] M. M. Vasiliu, J. Jaatinen, F. Johnson, B. Edvinsson, M. Romero, R. Guarese, The right voice for the right task: Exploratory insights on persona and voice embodiment in conversational ai for industrial training, International Journal of Human–Computer Interaction 0 (2026) 1–36. URL: https://doi.org/10.1080/10447318.2026.2651373. doi:10.1080/10447318.2026.2651373.

[8] S. Desai, M. Twidale, Metaphors in voice user interfaces: A slippery fish, ACM Trans. Comput.-Hum. Interact. 30 (2023). URL: https://doi.org/10.1145/3609326. doi:10.1145/3609326.

[9] S. Desai, M. Dubiel, L. A. Leiva, Examining humanness as a metaphor to design voice user interfaces, in: CUI '24, ACM, New York, NY, USA, 2024. URL: https://doi.org/10.1145/3640794.3665535. doi:10.1145/3640794.3665535.

[10] J.-Y. Jung, S. Qiu, A. Bozzon, U. Gadiraju, Great chain of agents: The role of metaphorical representation of agents in conversational crowdsourcing, in: CHI '22, ACM, New York, NY, USA, 2022. URL: https://doi.org/10.1145/3491102.3517653. doi:10.1145/3491102.3517653.

[11] B. Axtell, C. Munteanu, Tea, earl grey, hot: Designing speech interactions from the imagined ideal of star trek, in: CHI '21, CHI '21, ACM, New York, NY, USA, 2021. URL: https://doi.org/10.1145/3411764.3445640. doi:10.1145/3411764.3445640.

[12] S. K. Freire, T. He, C. Wang, E. Niforatos, A. Bozzon, Factory operators' perspectives on cognitive assistants for knowledge sharing: challenges, risks, and impact on work, Behaviour & Information Technology 0 (2025) 1–28. URL: https://doi.org/10.1080/0144929X.2025.2574373. doi:10.1080/0144929X.2025.2574373.