Stepford Twins and Potemkin Engineering: A Critique of Synthetic Personas in the Age of Generative AI

Michael Muller, IBM Research[1]
Katie Seaborn, Institute of Science Tokyo & University of Cambridge[2,3]

1 Introduction

Advances in generative artificial intelligence (AI) and notably large language models (LLMs) are heralding in new trajectories for designers, researchers, and practitioners in human-computer interaction (HCI) and elsewhere (Breum et al., 2024; Grossmann et al., 2023; Rossi et al., 2024). Experts are exploring the idea of representing people and populations based on existing data (Breum et al., 2024; Rossi et al., 2024) and synthesizing novel, unavailable, or costly users through the productive capacities of generative AI (Rossi et al., 2024). The idea is that ongoing challenges in research—such as small sample sizes, difficulties accessing special populations, and practical concerns like cost and time—may be alleviated or corrected through the apparent linguistic fluency and power of LLMs (Grossmann et al., 2023). These textual representations of "users" and "participants" are sometimes supplemented by human-appearing images created through large image models (LIMs). Designers may be drawn to the notion of LLM-generated personas that "mirror" real people (Rossi et al., 2024; von der Heyde et al., 2025), while scientists are exploring (and sometimes deploring) the validity and reliability of LLM-driven "participant" data (Bisbee et al., 2024; Grossmann et al., 2023). Generative AI as a research tool for user studies is tantalizing.

Yet, critical voices have urged caution in the face excitement about (Rossi et al., 2024)—and "techno-haste" towards (Seaborn, 2025)—such simulations of humans, generally termed *synthetic personas*. Evaluative work has surfaced flaws in the abilities of LLMs to accurately represent people at all scales (Bisbee et al., 2024; von der Heyde et al., 2025). Still, the conversation has tended to focus on the **LLM and its capabilities** in terms of data fidelity and representativeness (Rossi et al., 2024), stability over time (von der Heyde et al., 2025), and believability (Rossi et al., 2024; Törnberg, 2024). Rossi et al. (2024) extended the conversation to include meta-level concerns. Notably, they raised the problem of *epistemic legitimacy*: when there is no data, we should first ask why and be wary of the using generated content based on small data sets as representing some "truth" about real people. They also offered the frame of *situated knowledges*, calling on the community of experts to go beyond mere technical specs and reflectively question how our *positionality* enables us to determine what is data and what is not.

Here, we aim to advance the discourse even further by surfacing a higher-level issue: **the unacknowledged constructionism and reductionism in the very notion of synthetic personas.** Ultimately, we argue that experts are in a position of power as producers of the "who" and hence the "what" of generated "participant"-based design and research. As practitioners in HCI, we argue that synthetic user generation techniques defeats the purpose of user-centred design as it simply manifests and ventriloquizes the team's ideal users. We untangle our key points of contention here.

2 Why Use Personas?

We begin by acknowledging the rich history of persona use, especially in pursuit of user-centred design. HCI workers have described diverse motivations for the use of personas. Caballero et al. (2014) noted that personas originate in marketing practice (refer also to Cohen and Amble, 2025) and advocated for the use of personas as HCI *segmentation* aids. Acuña et al. (2012, ms p.1) used personas to avoid what they claimed were methodological and procedural uncertainties of "pure HCI techniques." Barambones et al. (2024) mentioned the difficulty of recruiting actual humans for empirical work. This problem may be acute if the end-users are distant from the design and development team (Putnam et al., 2012). Nielsen (2013, p. 8) pointed out the value of personas in supporting the design team "to maintain the users' perspectives." In a literature survey, Salminen et al. (2022) summarized eight uses of personas

[1]michael_muller@us.ibm.com
[2]katie.seaborn@cst.cam.ac.uk
[3]Katie Seaborn was partially supported by a JST PRESTO grant (#JPMJPR24I6).

This work is shared under a CC BY-SA 4.0 license unless otherwise noted

across HCI practice: Organization, Conceptualization, Ideation, Prototyping, Education, Copywriting, Prioritization, and Communication. Persona use touches on all levels and generations of HCI practice.

3 Problems with Personas: Expert Self-Deception, Then and Now

Concerns about persona use were raised before the emergence of LLMs. The representativeness and thus relevance of personas for end-users was questioned early on by Chapman and Milham (2006). Matthews et al. (2012) found that HCI practitioners tended to use personas as *communication aids within the design team*, noting the risks of misleading or distracting other power-holders with irrelevant persona attributes. The inclusion of machines is also somewhat a distraction. Just as marketing firms crafted representations of consumers for the average citizen to model (Caballero et al., 2014), HCI experts, then and now, are constructing who the "user" is through the compelling persona format. Generative AI merely boosts the power of experts to "make real" their vision of the ideal user. The novel interactive fluency of LLM- and LIM-based synthetic personas may have greater persuasive influence than the preceding generation of human-made personas, which often had non-interactive forms like posters. An interactive, LLM-based persona may be mistaken for an actual person *or* a representation based on an actual person. This deception may then be practiced upon team-members, executives, customers, or the academic community. We recall the ethically-questionable formulation of the Turing test as a matter of deceiving an observer between computational entities and flesh-and-blood humans (Natale, 2021), i.e., lying (Warwick and Shah, 2016). We describe two scenarios of deception through synthetic personas.

3.1 Potemkin Engineering

Grigori Aleksandrovich Potemkin was (among other roles) an administrator for Catherine the Great of Russia (Laplante, 2005). To persuade Catherine that he had improved the economy of recently conquered lands in Ukraine—absent of real improvements—he created false villages, populated by actors, constructed along Catherine's route. She was convinced. These "Potemkin villages" were portable, and could be moved from one point to the next, so as to simulate not a single village, but many villages.

Synthetic personas may function similarly to a Potemkin villager by substituting fiction in place of reality. They can be perceived as making truth claims about users who may or may not exist, deceiving teams and perhaps the experts themselves. Rossi et al. (2024) highlighted the naive enthusiasm undergirding this, echoed in the misimagination to which experts can fall prey when it comes to the epistemological realities of LLMs (Seaborn, 2025). The time and effort to create a Potemkin persona may take away needed resources from empirical work and testing the assumptions behind the synthesized user(s). In line with the non-synthetic versions (Salminen et al., 2022)—and akin to the portable Potemkin village—a Potemkin persona may be moved from one point to the next in a design+development process. A critical mass of Potemkin personas may thus magnify the distortion and potential for deception, tapping into our propensity for big numbers, i.e., the numerosity heuristic (Pelham et al., 1994).

3.2 Stepford Twins

Ira Levin (1972) described a patriarchal dystopian village wherein wives were replaced by simulated women: robots docile and obedient to their husbands. These "Stepford wives" were designed with made-to-order appearances and behaviours that appealed to the sexual tastes of their flesh-and-blood owners.

Synthetic users (personas) can function similarly to a post-replacement Stepford wife. They are docile. We made them, so they are likely to behave as we anticipate; obedient, they will seldom surprise us. They are likely to be attractive to us—designed to be just what we think a user should be. At the same time, they masquerade as "Digital Twins" that artificially "reflect the particular and individual, the idiosyncratic" (Bruynseels et al., 2018, p. 3). These Stepford Twins displace the actual users in our development process, leading to epistemic problems (we learn only what we have put into the synthetic users) and labor justice problems (we avoid paying actual people to participate in HCI activities).

4 Beware of Producing rather than Representing People through LLMs

LLM-based synthetic personas enable HCI workers to produce well-behaved users who serve our needs and tempt us with artificial realism. Inadvertently, we take on the role of Potemkin's engineers and Levin's Stepford husbands, creating portable, docile, attractive fictions that construct rather than represent real people. These fictions can deceive our teams, our clients, and ourselves. We ventriloquize actual users.

We hope that this critique will spark attention to the epistemic and political questions about generative AI, and return us to feminist and labour concerns in how we create and enact HCI methods.

This work is shared under a CC BY-SA 4.0 license unless otherwise noted

References

Acuña, S. T., Castro, J. W., and Juristo, N. (2012). A HCI technique for improving requirements elicitation. *Information and Software Technology*, 54(12):1357–1375.

Barambones, J., Moral, C., de Antonio, A., Imbert, R., Martínez-Normand, L., and Villalba-Mora, E. (2024). ChatGPT for learning HCI techniques: A case study on interviews for personas. *IEEE Transactions on Learning Technologies*, 17:1460–1475.

Bisbee, J., Clinton, J. D., Dorff, C., Kenkel, B., and Larson, J. M. (2024). Synthetic replacements for human survey data? The perils of large language models. *Political Analysis*, 32(4):401–416.

Breum, S. M., Egdal, D. V., Gram Mortensen, V., Møller, A. G., and Aiello, L. M. (2024). The persuasive power of large language models. *Proceedings of the International AAAI Conference on Web and Social Media*, 18:152–163.

Bruynseels, K., Santoni de Sio, F., and van den Hoven, J. (2018). Digital Twins in health care: Ethical implications of an emerging engineering paradigm. *Frontiers in Genetics*, 9.

Caballero, L., Moreno, A. M., and Seffah, A. (2014). Persona as a tool to involving human in agile methods: Contributions from HCI and marketing. In *Human-Centered Software Engineering: 5th IFIP WG 13.2 International Conference, HCSE 2014, Paderborn, Germany, September 16-18, 2014. Proceedings 5*, pages 283–290. Springer.

Chapman, C. N. and Milham, R. P. (2006). The personas' new clothes: Methodological and practical arguments against a popular method. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 50(5):634–636.

Cohen, Z. and Amble, S. (2025). Faster, smarter, cheaper: AI is reinventing market research. Last accessed 2025-06-06.

Grossmann, I., Feinberg, M., Parker, D. C., Christakis, N. A., Tetlock, P. E., and Cunningham, W. A. (2023). AI and the transformation of social science research. *Science*, 380(6650):1108–1109.

Laplante, P. A. (2005). The Potemkin village and the art of deception. *IT Professional*, 7(1):62–64.

Levin, I. (1972). The stepford wives. *Random House*.

Matthews, T., Judge, T., and Whittaker, S. (2012). How do designers and user experience professionals actually perceive and use personas? In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 1219–1228.

Natale, S. (2021). *Deceitful media: Artificial intelligence and social life after the Turing test*. Oxford University Press.

Nielsen, L. (2013). *Personas-user focused design*, volume 15. Springer.

Pelham, B., Sumarta, T., and Myaskovsky, L. (1994). The easy path from many to much: The numerosity heuristic. *Cognitive Psychology*, 26(2):103–133.

Putnam, C., Kolko, B., and Wood, S. (2012). Communicating about users in ICTD: leveraging HCI personas. In *Proceedings of the Fifth International Conference on Information and Communication Technologies and Development*, pages 338–349.

Rossi, L., Harrison, K., and Shklovski, I. (2024). The problems of LLM-generated data in social science research. *Sociologica*.

Salminen, J., Wenyun Guan, K., Jung, S.-G., and Jansen, B. (2022). Use cases for design personas: A systematic review and new frontiers. In *Proceedings of the 2022 CHI Conference on human factors in computing systems*, pages 1–21.

Seaborn, K. (2025). Insidious by design: Time and deception in human-LLM interactions. In Gellert, R., Schraffenberger, H., and Santos, C., editors, *Dark Patterns and Deceptive Design: Conceptualising and Systematising a Key Contemporary Phenomenon from a Legal Perspective and Beyond*. Edward Elgar.

This work is shared under a CC BY-SA 4.0 license unless otherwise noted

Törnberg, P. (2024). Best practices for text annotation with large language models. *Sociologica*.

von der Heyde, L., Haensch, A.-C., and Wenz, A. (2025). Vox populi, vox AI? Using large language models to estimate German vote choice. *Social Science Computer Review*.

Warwick, K. and Shah, H. (2016). Effects of lying in practical turing tests. *AI & society*, 31:5–15.

This work is shared under a CC BY-SA 4.0 license unless otherwise noted