Evaluation of AI-Based Digital Assistants in Smart Manufacturing

Alexandros Bousdekis[1](✉), Gregoris Mentzas[1], Dimitris Apostolou[1,2], and Stefan Wellsandt[3]

[1] Information Management Unit (IMU), Institute of Communication and Computer Systems (ICCS), National Technical University of Athens (NTUA), Athens, Greece
{albous,gmentzas}@mail.ntua.gr
[2] Department of Informatics, University of Piraeus, Piraeus, Greece
dapost@unipi.gr
[3] BIBA - Bremer Institut für Produktion und Logistik GmbH at the University of Bremen, Bremen, Germany
wel@biba.uni-bremen.de

Abstract. Industry 5.0 complements the Industry 4.0 paradigm by highlighting research and innovation as drivers for a transition to a sustainable, human-centric and resilient industry. In this context, new types of interactions between operators and machines are facilitated, that can be realized through artificial intelligence (AI) based and voice-enabled Digital Intelligent Assistants (DIA). Apart from the existing technological challenges, this direction requires new methodologies for the evaluation of such technological solutions that will be able to treat AI in manufacturing as a socio-technical system. In this paper, we propose a framework for the evaluation of voice-enabled AI solutions in Industry 5.0, which consists of four dimensions: the trustworthiness of the AI system; the usability of the DIA; the cognitive workload of individual users; and the overall business benefits for the corporation.

Keywords: Industry 5.0 · Evaluation methodology · Trustworthy AI · Voice-enabled assistant

1 Introduction

Industry 4.0 has revolutionized the manufacturing sector by integrating several technologies, such as Artificial Intelligence (AI), the Internet of Things (IoT), cloud computing, and Cyber Physical Systems (CPS). On the other hand, Industry 5.0 complements the existing Industry 4.0 paradigm by highlighting research and innovation as drivers for a transition to a sustainable, human-centric and resilient industry [1]. In this context, new types of interactions between operators and machines are facilitated, thus fostering the hybrid-augmented intelligence paradigm [2]. This paradigm can be realized through voice-enabled Digital Intelligent Assistants (DIA), which is more than a voice-based human-computer interface; its intelligence rests with the integration of diverse AI functionalities that have the capability to interact with the user via voice [2, 3].

Published by Springer Nature Switzerland AG 2022
D. Y. Kim et al. (Eds.): APMS 2022, IFIP AICT 664, pp. 503–510, 2022.
https://doi.org/10.1007/978-3-031-16411-8_58

Although industrial applications of DIAs have emerged only recently, they are expected to play a significant role in the collaboration between humans and AI systems [4–6]. Apart from the existing technological challenges, this direction requires new methodologies for the evaluation of such technological solutions that will be able to treat AI in manufacturing as a socio-technical system.

In this paper, we propose a framework for the evaluation of voice-enabled AI solutions in Industry 5.0, which consists of four dimensions: the trustworthiness of the AI system; the usability of the DIA; the cognitive workload of individual users; and the overall business benefits for the corporation.

The rest of the paper is organized as follows. Section 2 describes the four dimensions of the proposed evaluation framework and reviews the state-of-the-art in each dimension. Section 3 presents the proposed evaluation framework. Section 4 presents three use cases in which we will apply the proposed framework, while Sect. 5 concludes the paper and outlines our plans for future work.

2 Dimensions of Evaluation for Voice-Enabled AI Solutions in Industry 5.0

In this Section, we describe the four dimensions of the proposed evaluation framework, i.e. AI trustworthiness, system usability, cognitive workload, business benefits. For each dimension, we review the governing principles, and we review the state-of-the-art of related approaches tools in the literature.

2.1 AI Trustworthiness

To maximize the benefits of AI, while at the same time mitigating its risks, the concept of Trustworthy AI (TAI) promotes the idea that individuals, organizations, and societies will only ever be able to achieve the full potential of AI if trust can be established in its development, deployment, and use [7]. The TAI concept has been studied in several works (e.g. [8–10]. The increasing literature implies that although human-centricity is an indispensable feature of AI, it has not suited to be used by data scientists in the development of AI-based services or products [9, 11].

Therefore, AI should be: Lawful, complying with all applicable laws and regulations; Ethical, ensuring adherence to ethical principles and values; Robust, both from a technical and social perspective [7]. These requirements serve the need for trust [8, 10]. Despite their value for a realization of TAI, the outlined principles and the corresponding frameworks and guidelines face two major limitations [9]. First, several TAI principles may conflict with each other. Second, they are so general that do not provide sufficient guidance on how they can be transferred into practice. To this end, the High-Level Expert Group on Artificial Intelligence (AI-HLEG) has created the Assessment List for Trustworthy Artificial Intelligence (ALTAI) tool that helps organizations to self-assess the trustworthiness of their AI systems [7].

2.2 System Usability

Usability and user experience has become an important performance measure in the evaluation of interactive systems, since improving the end-user satisfaction leads to a greater adoption of the products [12]. A large number of usability evaluation tools are available, such as: The Usability Metric for User Experience (UMUX) [13], The Computer System Usability Questionnaire (CSUQ) [14], Software Usability Measurement Inventory (SUMI) [15], AttrakDiff [16], System Usability Scale (SUS) [17].

The SUS is the most widely used questionnaire for the assessment of perceived GUI usability that has significantly attracted the researchers and practitioners' interest [12, 18]. However, the application of SUS in voice assistants is limited [18]. Voice-interfaces face some distinct challenges, such as: the ability to understand non-conversational cues (i.e., pauses in the middle of a conversation) [18, 19], difficulty with back and forth navigation [20], absence of a visual feedback that increases the cognitive workload [21], users' pre-conceived expectations as to how a conversation should proceed [19], the effect of the quality of the synthetic voice to the perception of the users [22].

In the past, there had been an explosion of usability evaluation approaches, metrics and scales focusing on conversational interfaces, chatbots and intelligent assistants. Examples include: Subjective Assessment of Speech System Interfaces (SASSI) [23], Speech User Interface Service Quality (SUISQ) [24], SUXES [25]. However, recent literature on the usability of AI-based voice-assistants is lagging [12, 18]. To this end, an extension of SUS targeted explicitly to chatbots and voice-enabled interfaces has been developed, called Voice Usability Scale (VUS) [19].

2.3 Cognitive Workload

When attempting to solve a problem, humans call upon their limited cognitive resources. The degree of their utilization is described as cognitive load [26]. While the number of parameters to be considered and to be processed by modern-day knowledge workers increases, their cognitive resources do not [27]. The evaluation of cognitive workload is a key point in the research and development of human–machine interfaces, in search of higher levels of comfort, satisfaction, efficiency, and safety in the workplace [28]. The workload level experienced by an operator can affect task performance, since too high a load can increase stress and failure rates and decrease the work satisfaction and performance of employees [29].

The existing evaluation tools fall into three categories [30]: (a) performance-based measures, (b) subjective measures, and (c) physiological measures. The performance-based measures are grounded on the assumption that any increase in task difficulty will lead to a decrease in performance. Subjective procedures assume that an increased power expense is linked to the perceived effort. Physiological indexes assume that the mental workload can be measured by means of the level of physiological activation.

As human–machine systems have become more complex and automated, evaluations based on the operator's performance have become prohibitively difficult. To this end, subjective measures are becoming an increasingly important tool [28]. The reasons for their frequent use include their practical advantages (ease of implementation, non-intrusiveness) and current data which support their capability to provide sensitive

measures of operator load. The most well-established and widely used subjective method is the NASA-TLX, which allows a detailed analysis of the workload source (Task Load Index) (e.g., [28, 29, 31, 32], which was proposed by [33].

2.4 Business Benefits

The business benefits are defined in the form of business Key Performance Indicators (KPIs). KPIs are the quantifiable operational and strategic measurements that reflect the success factors of the manufacturing processes that adopt a technological solution. In this sense, they are used to quantify the efficiency and effectiveness of manufacturing processes [34]. Acknowledging the contributions and support of the KPIs, the decision-makers can evidence the existing gap between the before and after situation in terms of performance. According to the manufacturing process under examination and the business requirements, different KPIs can be defined and measured.

3 The Proposed Evaluation Framework

In this Section, we present our proposed evaluation framework for voice-enabled trustworthy AI solutions in Industry 5.0. As shown in Fig. 1, it is structured across the four dimensions of Sect. 3: AI trustworthiness, system usability, cognitive workload, business benefits. Below, we present the methods that address these dimensions.

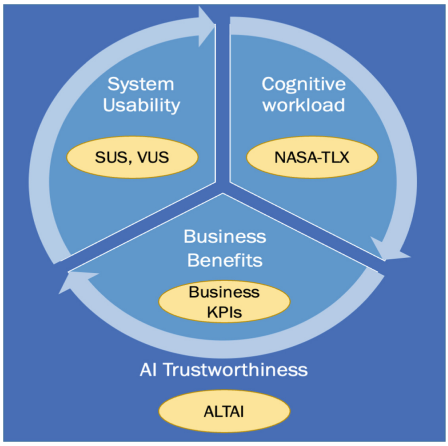

Fig. 1. The proposed evaluation framework for AI-based digital assistants in manufacturing.

AI Trustworthiness: In the proposed framework, AI trustworthiness is addressed by the ALTAI questionnaire which adopts the Ethics Guidelines for Trustworthy Artificial Intelligence proposed by the AI HLEG in order to self-assess its compliance to the seven requirements of Trustworthy Artificial Intelligence (TAI) [7]: Human agency

and oversight; Technical robustness and safety; Privacy and data governance; Transparency; Diversity, non-discrimination and fairness; Societal and environmental well-being; Accountability. Interdisciplinary expertise is required to answer since the very first questions.

System Usability: Voice-enabled AI solutions are implemented with a Digital Intelligent Assistant (DIA) as an interface, while they are usually accompanied by a GUI for visualization. The GUI can be evaluated through the well-established SUS which has proved to have high reliability, validity, while it is sensitive to a wide variety of independent variables. Drawing parallels to SUS, the VUS scale is also a 10-item one, a 7-point Likert scale having declarative statements of opinion to which the participants will respond with their rate of agreement, allowing meaningful comparisons to be made between the two [18]. The problem of developing a usability measure for voice-assistants is that there are no commonly accepted usability dimensions.

Cognitive Workload: This dimension is addressed by the NASA-TLX, a well-established and widely used subjective method, which includes six dimensions: Mental Demand; Physical Demand; Temporal Demand; Overall Performance; Effort; Frustration Level. The adoption of emerging AI technologies poses new challenges to both employers and employees of manufacturing companies who need to adapt to new processes requiring an efficient management of workload [35, 36]. Virtual assistants provide opportunities to reduce the workload by assisting in the execution of repetitive tasks that require the fast retrieval and processing of data [27, 37]. It was only recently that there is some preliminary evidence that virtual assistants are able to reduce the cognitive load when performing tasks [27]; however, evaluation in real manufacturing environments is still at its early stages [38].

Business Benefits: The business benefits are defined in the form of business KPIs by the use cases in which the voice-enabled AI solution under evaluation is deployed. Various categories of KPIs can be examined according to the scope of the technological solution, the manufacturing processes under consideration, and the business goals, such as: organizational, financial, business, operational, technology, health & safety, environmental sustainability.

4 Use Cases

In this Section, we briefly describe three use cases in which we will apply the proposed framework.

On-the-Job Training in Textile Production: This scenario addresses the shortage of qualified labor force in processes from raw materials to fabric delivery and, to clothing sale to consumers. A key goal is to maintain the worker's autonomy instead of promoting the unquestioned execution of instructions. A voice-enabled AI solution will identify the worker's current skill level and will adapt the advising behavior according to the learning experience, accompanied by explainability functionalities. In this way, the training support will contribute to the defects reduction that are caused by human errors.

End-of-Line Quality Control in White Goods Production: This scenario addresses the support of operators at the end-of-line quality control through a Digital Intelligent Assistant (DIA) in order to adopt a predictive quality strategy that will link the quality control of the finished product with the design stage and the shop floor. By integrating all available information sources (e.g. sensor data, historical operational data, and expert knowledge), it will be able to predict low-quality products and to plan mitigating actions in order to proactively identify their root causes in order to, among others, reduce organization, warranty but also reputation costs.

Line Re-configuration in Hygiene Products Manufacturing: This scenario addresses the setup and change-over of production lines that require trained workers capable of (re)configure machines, align production speeds, and adjust machine settings within a given amount of time. To address these problems, the company aims to standardize the reconfiguration process by capturing the best practices and by sharing them through a digital intelligent assistant which will guide the workers towards optimum configuration of the production line. This will reduce the change-over time, time pressure caused by downtime, and lessen the cognitive workload of workers in solving unpredictable complex tasks.

5 Conclusion and Future Work

Although industrial applications of DIAs have emerged only recently, they are expected to play a significant role in the collaboration between human and AI. This direction requires new methodologies for the evaluation of such AI solutions.

In this paper, we proposed a framework for the evaluation of AI-based Digital Assistants in smart manufacturing, which consists of four dimensions: AI trustworthiness, system usability, cognitive workload, business benefits.

We are currently in the process of applying the evaluation framework in the three aforementioned real-life scenarios. An early application of the ALTAI framework with a first demonstration version of the DIA has already demonstrated the benefits, and also some limitations, of the approach. In the near future we will apply all four dimensions and examine their suitability and usefulness for digital intelligent assistants in manufacturing use cases.

Acknowledgements. This work is partly funded by the European Union's Horizon 2020 project COALA "COgnitive Assisted agile manufacturing for a LAbor force supported by trustworthy Artificial Intelligence" (Grant agreement No 957296). The work presented here reflects only the authors' view and the European Commission is not responsible for any use that may be made of the information it contains.

References

1. Maddikunta, P.K.R., et al.: Industry 5.0: a survey on enabling technologies and potential applications. J. Ind. Inf. Integr. **26**, 100257 (2021)

2. Wellsandt, S., et al.: Hybrid-augmented intelligence in predictive maintenance with digital intelligent assistants. In: Annual Reviews in Control (In Press, Corrected Proof) (2022)
3. Dhiman, H., Wächter, C., Fellmann, M., Röcker, C.: Intelligent assistants. Bus. Inf. Syst. Eng. 1–21 (2022)
4. Rabelo, R.J., Romero, D., Zambiasi, S.P.: Softbots supporting the operator 4.0 at smart factory environments. In: Moon, I., Lee, G., Park, J., Kiritsis, D., Von Cieminski, G. (eds.) Advances in Production Management Systems. Smart Manufacturing for Industry 4.0. APMS 2018. IFIP Advances in Information and Communication Technology, vol. 536, pp. 456–464. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-99707-0_57
5. Bousdekis, A., et al.: Human-AI collaboration in quality control with augmented manufacturing analytics. In: Dolgui, A., Bernard, A., Lemoine, D., von Cieminski, G., Romero, D. (eds.) Advances in Production Management Systems. Artificial Intelligence for Sustainable and Resilient Production Systems. APMS 2021. IFIP Advances in Information and Communication Technology, vol. 633, pp.303–310. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-85910-7_32
6. Wellsandt, S., Hribernik, K., Thoben, K.D.: Anatomy of a digital assistant. In: Dolgui, A., Bernard, A., Lemoine, D., von Cieminski, G., Romero, D. (eds.) Advances in Production Management Systems. Artificial Intelligence for Sustainable and Resilient Production Systems. APMS 2021. IFIP Advances in Information and Communication Technology, vol. 633, pp. 321–330. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-85910-7_34
7. High-Level Independent Group on Artificial Intelligence (AI HLEG). Ethics Guidelines for Trustworthy AI. https://ec.europa.eu/digital
8. Floridi, L.: Establishing the rules for building trustworthy AI. Nat. Mach. Intell. **1**(6), 261–262 (2019)
9. Baneres, D., Guerrero-Roldán, A.E., Rodríguez-González, M.E., Karadeniz, A.: A predictive analytics infrastructure to support a trustworthy early warning system. Appl. Sci. **11**(13), 5781 (2021)
10. Thiebes, S., Lins, S., Sunyaev, A.: Trustworthy artificial intelligence. Electron. Mark. **31**(2), 447–464 (2020). https://doi.org/10.1007/s12525-020-00441-4
11. Georgieva, I., Lazo, C., Timan, T., Van Veenstra, A.F.: From AI ethics principles to data science practice: a reflection and a gap analysis based on recent frameworks and practical experience. AI Ethics 1–15 (2022)
12. Kocaballi, A.B., Laranjo, L., Coiera, E.: Understanding and measuring user experience in conversational interfaces. Interact. Comput. **31**(2), 192–207 (2019)
13. Finstad, K.: The usability metric for user experience. Interact. Comput. **22**(5), 323–327 (2010)
14. Lewis, J.R.: IBM computer usability satisfaction questionnaires: psychometric evaluation and instructions for use. Int. J. Hum. Comput. Interact. **7**(1), 57–78 (1995)
15. Kirakowski, J.: Software usability measurement inventory SUMI. SUMI (2011). http://sumi.uxp.ie/en/index.php
16. Hassenzahl, M., Burmester, M., Koller, F.: AttrakDiff: a questionnaire to measure perceived hedonic and pragmatic quality. Mensch Comput. **57**, 187–196 (2003)
17. Brooke, J.: SUS-A quick and dirty usability scale. Usability Eval. Ind. **189**(194), 4–7 (1996)
18. Zwakman, D.S., Pal, D., Arpnikanondt, C.: Usability evaluation of artificial intelligence-based voice assistants: the case of Amazon Alexa. SN Comput. Sci. **2**(1), 1–16 (2021). https://doi.org/10.1007/s42979-020-00424-4
19. Murad, C., Munteanu, C., Cowan, B.R., Clark, L.: Revolution or evolution? Speech interaction and HCI design guidelines. IEEE Pervasive Comput. **18**(2), 33–45 (2019)
20. Holmes, S., Moorhead, A., Bond, R., Zheng, H., Coates, V., McTear, M.: Usability testing of a healthcare chatbot: can we use conventional methods to assess conversational user interfaces? In: Proceedings of the 31st European Conference on Cognitive Ergonomics, pp. 207–214 (2019)

21. Cowan, B.R., et al.: What can i help you with? Infrequent users' experiences of intelligent personal assistants. In: Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services, pp. 1–12 (2017)
22. Babel, M., McGuire, G., King, J.: Towards a more nuanced view of vocal attractiveness. PLoS ONE **9**(2), e88616 (2014)
23. Hone, K.S., Graham, R.: Towards a tool for the subjective assessment of speech system interfaces (SASSI). Nat. Lang. Eng. **6**(3–4), 287–303 (2000)
24. Polkosky, M.D.: Machines as mediators: the challenge of technology for interpersonal communication theory and research, pp. 48–71. Routledge (2008)
25. Turunen, M., Hakulinen, J., Melto, A., Heimonen, T., Laivo, T., Hella, J.: SUXES-user experience evaluation method for spoken and multimodal interaction. In: Tenth Annual Conference of the International Speech Communication Association (2009)
26. Sweller, J.: Cognitive load during problem solving: effects on learning. Cognit. Sci. **12**(2), 257–285 (1988)
27. Brachten, F., Brünker, F., Frick, N.R., Ross, B., Stieglitz, S.: On the ability of virtual agents to decrease cognitive load: an experimental study. Inf. Syst. e-Bus. Manag. **18**(2), 187–207 (2020)
28. Rubio, S., Díaz, E., Martín, J., Puente, J.M.: Evaluation of subjective mental workload: a comparison of SWAT, NASA-TLX, and workload profile methods. Appl. Psychol. **53**(1), 61–86 (2004)
29. Cao, A., Chintamani, K.K., Pandya, A.K., Ellis, R.D.: NASA TLX: software for assessing subjective mental workload. Behav. Res. Methods **41**(1), 113–117 (2009). https://doi.org/10.3758/BRM.41.1.113
30. Meshkati, N., Hancock, P.A., Rahimi, M., Dawes, S.M.: Techniques in mental workload assessment (1995)
31. Hart, S.G.: NASA-task load index (NASA-TLX); 20 years later. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting, vol. 50, no. 9, pp. 904–908. Sage publications, Sage CA: Los Angeles, CA (2006)
32. Castro, S.C., Quinan, P.S., Hosseinpour, H., Padilla, L.: Examining effort in 1d uncertainty communication using individual differences in working memory and NASA-TLX. IEEE Trans. Vis. Comput. Graph. **28**(1), 411–421 (2021)
33. Hart, S.G., Staveland, L.E.: Development of NASA-TLX (Task Load Index): results of empirical and theoretical research. In: Advances in psychology, vol. 52, pp. 139–183. North-Holland (1988)
34. Zhu, L., Johnsson, C., Varisco, M., Schiraldi, M.M.: Key performance indicators for manufacturing operations management–gap analysis between process industrial needs and ISO 22400 standard. Procedia Manuf. **25**, 82–88 (2018)
35. Galy, E., Cariou, M., Mélan, C.: What is the relationship between mental workload factors and cognitive load types? Int. J. Psychophysiol. **83**(3), 269–275 (2012)
36. Matt, C., Hess, T., Benlian, A.: Digital transformation strategies. Bus. Inf. Syst. Eng. **57**(5), 339–343 (2015)
37. Dellermann, D., Ebel, P., Söllner, M., Leimeister, J.M.: Hybrid intelligence. Bus. Inf. Syst. Eng. 61(5), 637–643 (2019)
38. Mirbabaie, M., Stieglitz, S., Brünker, F., Hofeditz, L., Ross, B., Frick, N.R.: Understanding collaboration with virtual assistants–the role of social identity and the extended self. Bus. Inf. Syst. Eng. **63**(1), 21–37 (2021)