A Pilot Study for Analyzing Systems Engineer-Conversational GenAI Interaction: A Case Study for Requirements Development & Validation

Emilie Perreau
Univ. Grenoble Alpes, CNRS, Grenoble INP, G-SCOP
emilie.perreau@grenoble-inp.fr

Romain Pinquié
Univ. Grenoble Alpes, CNRS, Grenoble INP, G-SCOP
romain.pinquie@grenoble-inp.fr

Cédric Masclet
Univ. Grenoble Alpes, CNRS, Grenoble INP, G-SCOP
cedric.masclet@gscop.fr

With the democratization of Large Language Models, academics and professionals are searching for new use cases of conversational Generative Artificial Intelligence (GenAI) for systems engineering, including requirements engineering. This paper presents a pilot study to understand the impact of guidelines and templates on the interaction between ChatGPT and a systems engineer for developing system requirements. Results show that when appropriately used, prompting guidelines and templates improve the quality of requirements. Still, without domain knowledge, the GenAI cannot generate outputs with the quality expected by requirements engineering international standards.

Introduction

With Generative Artificial Intelligence (GenAI) and Large Language Models (LLMs) becoming very popular since version 3.5 of Chat Generative Pre-trained Transformer (ChatGPT), conversational GenAIs have become up-and-coming tools for supporting engineers. Promotional materials illustrate many promises, but we do not understand how a systems engineer interacts with a conversational GenAI and miss methodological guidelines that help to specify, design, and evaluate the Human-Systems Integration of a systems engineer with a conversational GenAI.

Literature Review

This section introduces the gist of interactions between a human and a conversational GenAI before focusing on systems engineer-conversational GenAI interactions.

Human-Conversational GenAI Interactions

As defined by (Hornbæk and Oulasvirta, 2017), the interactions can represent a set of "*two entities that mutually determine their behaviors*". In Human-Computer Interaction, these entities are "*computers (ranging from input devices to systems) and humans (ranging from operators to tools users)*", which can be conceptualized by their dialogue, transmissions, behaviors, tool-use, embodiment, experience or control. Interactions involve two parts to elaborate a dialogue between those instances. Conversational GenAI caused debates around their abilities to construct "human-like" interactions, emphasizing cognitive capacities implemented by their designers. This conflict illustrates the orientation of previous research studies on the subject, with some authors acknowledging the "collaborative" performance of the system (Skov *et al.*, 2022). Meanwhile, others demonstrate their limitations

and lack of human awareness (Sarkar, 2023) as still a mirror of their designer's mind. Thus, conversational GenAI seems like a valuable tool for supporting the decision and understanding of the user (Vinchon et al., 2023). Despite those results, studies in engineering start to highlight the lack of a comprehensive view regarding the subject, expanding the gap between the experimental data and the reality of their implementation in professional activities (Rapp *et al.*, 2023).

Human-Conversational AI Interactions for Systems Engineering

Since the release of ChatGPT 3.5 and Bard, conversational GenAI based on LLM is a technology that is gaining interest to support software (Ozkaya, 2023) and systems engineering, especially requirements engineering (Arora et al., 2023; Arvidsson and Axell, 2023; Ray, 2023), safety analysis (Qi et al., 2023), and modelling tasks (Cámara et al., 2023). The application programming interfaces of LLMs are available to other developers to accelerate the customization of generic conversational GenAI for systems engineering applications. For instance, ReqGPT[1] uses prompt engineering techniques to refine, review, and improve ambiguity, consistency, completeness, and verifiability of requirements. However, we miss empirical evidence supporting existing theories of the interaction between a systems engineer and a conversational GenAI, as most case studies are communications from the marketing departments of systems engineering software editors[2] and companies that provide consulting services, making distinguishing actual capabilities from marketing rhetoric difficult.

Research Question

As claimed in the HSI field, the Human Centered Approach recognizes humans as an integral part of the system and, as such, considers the factors that influence the system of interest at the same level as the results that it produces. This strategy, which allows more consideration into the implementation of the system, ensuring the respect of the operator's constraints and capacities, seems to develop in the context of human-conversational agent interaction, where the dynamic between both is still considered with a black box perspective. However, there is a lack of research methods to study Human-Conversational GenAI Interactions, as new studies focused on the evaluation of the technology itself rather than the evolution brought by the conversation in a dynamic context. To reduce the gap, this pilot study explores the feasibility and value of a mixed approach merging qualitative and quantitative methods towards the common goal of dialogue analysis. The mixed research method is evaluated on a case study that aims **to analyze the impact of prompting methodical recommendations on the definition of requirements** as part of the research question: How do prompting guidelines and templates impact the quality of the requirements definition?

Research Method

As a pilot study, the objective is to construct and test the implementation of mixed methods, enabling a cross-comparison of subjective and objective views for human-agent interactions. The study was built around four data collection methods: questionnaires, video recording, dialogue recording and recorded semi-structured interviews. The experiment targets the sensibility of the method regarding the diversity of user-profiles accounted for in this type of activity. Three mains' personas were defined based on their experience level in Systems Engineering (SE) and Model-Based SE bodies of knowledge, as well as the use of conversational AI. To apply the pilot study and test the elaborated method, three participants matching those personas were recruited from an online list diffusion.

At the beginning of the experiment, participants were given a free consent form and a notice detailing the experiment. The volunteers were asked only to read the first part of the document, divided into

[1] https://www.opencaesar.io/projects/2023-05-23-ReqGPT.html
[2] https://www.valispace.com/ai/

three exercises. The exercises were conditioned by their completeness: the participant only acknowledged the content of the following exercise after completing the first one, with no return allowed to the previous section. The case study was limited to a maximum of two hours. The instructions were given in ten minutes, followed by a questionnaire to evaluate participants' knowledge level in requirements engineering and conversational GenAI while evaluating their ability to understand the model-based definition of the system context and functions provided as input data. Each exercise was limited to twenty minutes, with a ten-minute break, and concluded with a self-assessment questionnaire before passing to the next one.

A review of systems engineering standards and guidelines shows that several ontologies propose a different standardization of terms, definitions and quality criteria for a requirement and a set of requirements. Consequently, participants were given a glossary of keywords (system, system-of-interest, stakeholder, external system, operating environment, system function, external interface, system functional requirement, requirements, validation, validated system functional requirement, validated set of system functional requirements, requirement correctness, requirement completeness). We also provided participants with generic quality criteria of a requirement (unambiguous, consistent, complete, singular, feasible, traceable, and verifiable) and a set of requirements (complete and correct) (SAE ARP4754A, 2015) to guide non-experts in the requirements validation task.

The task concentrates on the specification of requirements, mainly their development and validation, for an Electric Toothbrush System. An electric toothbrush was chosen as the system-of-interest because it is relatively simple and does not require a domain expert. Still, it is technically relevant to follow the systems engineering approach. The development of requirements consists of deriving functional requirements at the system – seen as a black box – level from various system artifacts, such as the intended use environment, external entities, external interfaces, and system functions. After the requirements development, the validation task aims to increase confidence in the completeness and correctness of each requirement and the set of requirements (SAE ARP4754A, 2015). The selection of the requirements development task was motivated by the strong influence of natural language. Indeed, the natural language is ambiguous, especially when there is no shared conceptualization amongst the members of the domain of discourse used in requirements engineering, leading to various ambiguities in the definition of the concepts (e.g., need, function, functional requirement, capability, service, mission, use case, scenario) but also the writing of requirements statements. The second task, requirements validation, was chosen to evaluate the influence of the systems engineer's trust in the conversational GenAI. Both tasks were repeated in three exercises where the participants were free to require the assistance of ChatGPT 4.0: – 1) without prompting guidelines for the function "To transform electric power into mechanical brushing power", 2) with prompting guidelines for the function "To send last brushing duration", and 3) with prompting templates for the function "To inform on status ON/OFF/IDLE". Moreover, without any context and expected system functions, the requirements specification developed by different systems engineers would likely end up in a completely different set of system requirements. Thus, the design inputs also included the definition of the system context, including the external stakeholders, systems, and interfaces, as well as three system functions from which system functional requirements must be derived. This system context definition was presented to systems engineers using a concise SysML Internal Block Diagram (Figure 1). Finally, researchers conducted a semi-directed interview with each participant within a limit of one hour.

A quantitative study was planned to calculate quality scores for each modality based on criteria defined in SAE ARP4754A (2015), ISO/IEC/ IEEE 29148 (2018), and the INCOSE Guide to Writing Requirements (INCOSE Requirements Working Group, 2022). However, the poor quality of the generated requirements made the approach unpracticable and unusable. A qualitative analysis was also conducted on the obtained materials. The coding was elaborated from the scientific literature (cf. Table 2) and adapted with a grounded theory approach.

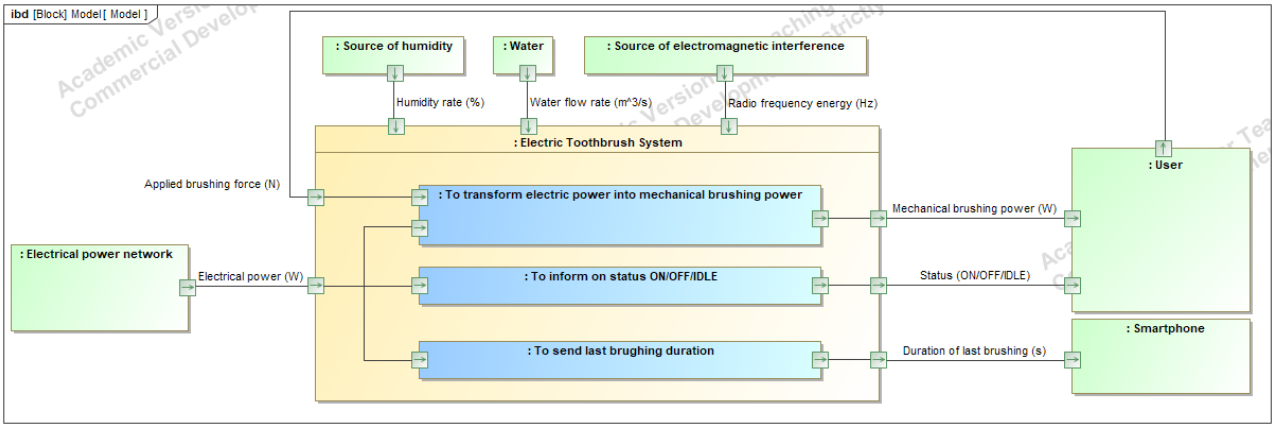


Figure 1. A SysML Internal Block Diagram defining the system-of-interest (in yellow), the external entities (in green) and interfaces, and the system functions (in blue).

Methodological results

The selection ends up with one participant per profile, leading to the following sample:

Table 1. Study sample

Participants	Experience in MBSE	Experience with a conversational AI
P1	High (PhD in MBSE, certified INCOSE Associate Systems Engineering Professional, MBSE researcher)	High, science-based (researcher in conversational AI, daily use)
P2	Medium (PhD in SE)	Low (3-4 previous uses)
P3	Low (apprentice in SE)	High, non-science-based (daily use)

The research collected nine dialogues associated with three interviews transcribed. The overall material was analyzed according to the previous interaction definition (Hornbaek and Oulasvirta, 2017) and specialized for the study (see Table 2).

Table 2. Dialogue analysis

Types of interactions	Dimensions	Observation
Tool Use	Task delegation	The study comparing different levels of supervision without analyzing the nominal situation didn't allow for observation of task delegation, leading to the creation of a new dimension called "tasks chronology".
	Chronology of the tasks	The distribution of tasks and subtasks throughout the activity was analyzed chronologically and related to the moments of use of the guidelines and the templates.
	Forms of guidance	Guidance are commonly studied in Psychology. Here, they were contextualized as behavioral (anticipatory or corrective) and tool-oriented (information management). Those dimensions demonstrate an interest in the utility perceived by participants towards templates and guidelines regarding their position and succession in the task.
	Frequency of use	The study allowed a high level of freedom concerning the use of templates and guidelines. The frequency of their apparition towards tasks and subtasks was used

		as an indicator to question participants about their underlying needs at the time of their utilisation. This dimension was used as a vector for exploring the participants' representations of the provided supervision tools.
Behaviors	Decision-making (Viros and Selva, 2019)	As coded by Viros and Selva (2019), the behaviors of exploration (enrichment of the TCA proposition) and exploitation (instant use of the TCA proposition) were both observed in the material. However, those propositions had been modified to analyze precisely the proportion of the proposition directly accepted by the participant until the production, those without direct feedback but treated in later turns, those directly submitted to further explorations and those denied from the following turns.
	Intentions	The intentions were only revealed during the self-confrontation interviews with the help of the How? Why? technique (Graesser et al., 1978), and were limited to the participant's level of self-consciousness.
	Intervention management	This analysis observed the capacity of the locutor to answer from a direct or an indirect request source. The coding was composed of omission, direct answers and indirect answer coding, acknowledging the position of the reply from the request source.
	Attention manipulation	Focusing on a micro level of the conversation, this dimension was analyzed with a track of the information emission source (participant, TCA as long as their place in the dialogue) and their influence in the user request (primary, secondary). This coding illustrated the impacts of templates and guidelines on dependencies inside the text.
	Social coordination	As turn-taking is already pre-defined with the request-answer format, an analysis of role allocation was carried out. Macro and micro analyses are recommended as a speaker could handle multiple roles simultaneously in the same prompt. The roles were also coded as “induced” by a direct ask or “construct”.
Dialogues	Trouble sources	Identified as the position behind a repair initiative, trouble sources were then compared to the temporalities of the utilization of guidelines and templates.
	Repairs	Repair strategies can be seen as an indicator of the quality of the interaction as a sign of adaptability to new situations. In fact, as mentioned by Corti and Gillespie (2016), a repair strategy is often awaited in the three next turns of conversation, following the process of apparition of the trouble source, repair initiation, and repair.

General Discussion

The proposed method presented a first view of the strengths and weaknesses of dialogue analysis for studying human-conversational GenAI interactions. As such, evaluating the efficiency of tools and processes seems to show some substantial limitations in time-constrained experiments for various levels of knowledge in the domain of interest. Without the factual indications of previous personal experiences, the pre-assessment questionnaire failed to differentiate the participant profiles precisely. However, it was observed in the prompting structure at the beginning of the interaction with the conversational GenAI. Although the study enabled us to understand the dynamic of the interactions, the analysis still lacks a clear view of the internal requirements expected by the participant and their evolution throughout the task itself. As a participant might not have the same representation of the prescriptive criteria, the behaviors and production couldn't be entirely related to the capacity of adaptation of the technology, questioning the user-centric or task-centric evaluation perspective. Future studies should consider that perspective by integrating auto-evaluative measures of status and completeness towards the task. Nonetheless, dialogue analysis seems valuable for the human-centric approach because it highlights phenomena that limit or enhance the impact of such supervision tools. As detailed in dialogue analysis, the temporality of the exchange should be distinctive from the temporality of the task to understand their reciprocity.

To this extent, integrating interactions recommendations should also benefit from a systemic approach by considering the dialogue as a component influenced by and influencing internal and external entities of the same environment. The subject of interest is a GenAI-driven chatbot, so future studies should integrate the degree of freedom towards the sake of the analysis, notably by observing the assertiveness of mutual propositions and their effects on the resulting behavior.

Observations show that when used, prompting guidelines and templates do impact the quality of requirements but also the dynamic of the conversation itself. Templates are even more so. Both can help to increase the completeness and correctness of an individual requirement and a set of requirements. They can also enable a requirements engineer to drastically reduce the time to get outputs with a quality they judge sufficient. Nevertheless, ChatGPT – with or without guidelines or templates – does not substitute for expert knowledge. It can help specify performance criteria and, in a broader sense, quantitative values learnt while training the LLM and suggest functions or conditions that the engineer overlooked. However, ChatGPT, like any generative algorithm that makes next-word predictions unrelated to any real understanding of language, only goes where the human guides it. If you ask for requirements, it will generate a set of broad requirements. If you ask for requirements, each as a combination of an in-out transformation function and performance interval, it will generate it for you but without the conditions of use, meaning that the system must perform the function with the prescribed performance under any condition of use, leading to an incomplete requirement. ChatGPT offers assistance when it is fed with expert knowledge (vocabulary, methodological rules, standardized processes, domain knowledge, etc.). The conversational GenAI may seem impressive for novices as it provides a profusion of useless text. A large amount of generated text can distract the user from the primary task. Without enough domain and systems engineering knowledge, he cannot react by prompting ChatGPT to refocus on the main objective. When the human cannot inject expert knowledge, or only a little, into the prompts, guidelines and templates can help him escape superficial, all-purpose, and often methodologically irrelevant outputs. Despite the guidance, the templates are still tools that require prior knowledge regarding their use. Without fulfilling this condition, participants with less experience in ChatGPT prompting disengaged from the generated outputs, the misunderstanding of template requirements leading them to generate unexpected and unsatisfactory results, discarding their utility. This situation encourages participants to experiment with prompting on their own, trying to refer to their knowledge or creativity to access their primary intentions and pursue the conversation. This lack of intention assessment from the conversational GenAI makes proficient participants feel like mentors trying to guide ChatGPT, overcoming their first objective of defining qualitative requirements: "*I mean, he is like an intern. You ask him to do something, and*

then you refine and do it. It was like I wanted him to learn what I wanted to get. [...] So we must be careful to continuously provide the information needed for the task.".

Conclusion

This paper reports on a pilot study that investigated the relevance of mixed methods for observing the impact of supervision levels into human-TCA interactions. Methods demonstrate the value of analyzing closely the dynamic of interaction as a qualitative assessment for supervision tools, as a complementary approach for the performance analysis. In fact, the application of the proposed dimensions confirmed their potential for future studies, especially the coding grid adapted from previous researches. However, this research presented some potential of improvement, notably with the integration of a subdimension to understand the tool appropriation thought the different stages of the task. Despite the self-assessment questionnaire, the study could have evaluated the user experience with questionnaire oriented towards the acceptability and perceived utility of interaction strategies, and not only the TCA output. The skills assessment also failed to differentiate user profiles and their fitting to the elaborated personas, to reduce the hazardous choice, future assessment should implement open questions reflecting the participant's representation and knowledge for the task. Additionally, the analysis was missing the data provided from the nominal situation of the activity, which could have enabled further comprehensions towards the user's interaction strategies. The authors recommend further specification of the coding grid from ecological activity data, to improve their external validity for a given context.

References

Arora, C., Grundy, J. and Abdelrazek, M. (2023), "Advancing Requirements Engineering through Generative AI: Assessing the Role of LLMs", arXiv, 1 November.

Arvidsson, S. and Axell, J. (2023), "Prompt engineering guidelines for LLMs in Requirements Engineering".

Cámara, J., Troya, J., Burgueño, L. and Vallecillo, A. (2023), "On the assessment of generative AI in modeling tasks: an experience report with ChatGPT and UML", *Software and Systems Modeling*, Vol. 22 No. 3, pp. 781–793, doi: 10.1007/s10270-023-01105-5.

Choi, M 2008, 'Contesting Imaginaries in Death Rituals during the Northern Song Dynasty', PhD thesis, University of Chicago (Chicago, IL, US).

Corti, K., & Gillespie, A. (2016). Co-constructing intersubjectivity with artificial conversational agents: People are more likely to initiate repairs of misunderstandings with agents represented as human. *Computers in Human Behavior*, *58*, 431-442. https://doi.org/10.1016/j.chb.2015.12.039

Graesser, A., C; Robertson, S., P., Lovelace, E., R. and Swinehart D., M. (1980) Answers to Why-Questions Expose the Organisation of Story Plot and Predict Recall of Actions, *Journal of Verbal Learning and Verbal Behavior*, 19, 110-119.

Haskins, C (ed.) 2007, Systems Engineering Handbook: A Guide for System Life Cycle Processes and Activities, Version 3.1., INCOSE, San Diego, CA (US).

Hornbæk, K. and Oulasvirta, A. (2017), "What Is Interaction?", *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, presented at the CHI '17: CHI Conference on Human Factors in Computing Systems, ACM, Denver Colorado USA, pp. 5040–5052, doi: 10.1145/3025453.3025765.

INCOSE Requirements Working Group. (2022), *Guide to Writing Requirements*.

ISO/IEC/ IEEE 29148. (2018), "Systems and software engineering Life cycle processes — Requirements engineering".

Mankins, JC 1995, 'Technology Readiness Levels', White paper, NASA Office of Space Access and Technology, viewed 16 October 2010, http:/www.hq.nasa.gov/office/codeq/trl/trl.pdf.

Ozkaya, I. (2023), "Application of Large Language Models to Software Engineering Tasks: Opportunities, Risks, and Implications", *IEEE Software*, Vol. 40 No. 3, pp. 4–8, doi: 10.1109/MS.2023.3248401.

Pollan, M 2006, The Omnivore's Dilemma: A Natural History of Four Meals, Penguin, New York, NY (US).

Qi, Y., Zhao, X., Khastgir, S. and Huang, X. (2023), "Safety Analysis in the Era of Large Language Models: A Case Study of STPA using ChatGPT", arXiv, doi: 10.48550/ARXIV.2304.01246.

Rapp, A., Boldi, A., Curti, L., Perrucci, A. and Simeoni, R. (2023), "Collaborating with a Text-Based Chatbot: An Exploration of Real-World Collaboration Strategies Enacted during Human-Chatbot Interactions", *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, presented at the CHI '23: CHI Conference on Human Factors in Computing Systems, ACM, Hamburg Germany, pp. 1–17, doi: 10.1145/3544548.3580995.

Ray, A.T. (2023), *Standardization of Engineering Requirements Using LLM*, May.

SAE ARP4754A. (2015), *ARP4754A Guidelines For Development Of Civil Aircraft and Systems*.

Sarkar, A. (2023), "Enough With 'Human-AI Collaboration'", *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, presented at the CHI '23: CHI Conference on Human Factors in Computing Systems, ACM, Hamburg Germany, pp. 1–8, doi: 10.1145/3544549.3582735.

Sheard, SA 2006, 'Definition of the Sciences of Complex Systems', INSIGHT 9 (1), 25.

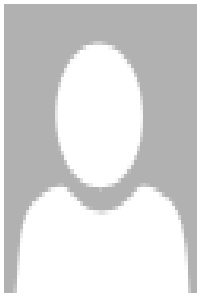
US Department of Defense 2003, Department of Defense Directive 5000.1. The Defense Acquisition System, Office of the Under Secretary of Defense for Acquisition, Technology, and Logistics. Washington, DC (US).

Skov, S.S., Andersen, J.R., Lauridsen, S., Bab, M., Bundsbæk, M. and Nielsen, M.B.D. (2022), "Designing a conversational agent to promote teamwork and collaborative practices using design thinking: An explorative study on user experiences", *Frontiers in Psychology*, Vol. 13, p. 903715, doi: 10.3389/fpsyg.2022.903715.

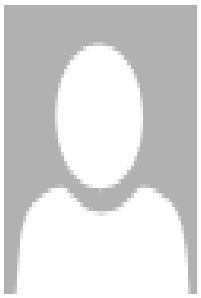
——— 2010, Systems Engineering Handbook: A Guide for System Life Cycle Processes and Activities, Version 3.2., INCOSE, San Diego, CA (US).

Viros Martin, A., & Selva, D. (2019, janvier 7). From Design Assistants to Design Peers : Turning Daphne into an AI Companion for Mission Designers. *AIAA Scitech 2019 Forum*. AIAA Scitech 2019 Forum, San Diego, California. https://doi.org/10.2514/6.2019-0402

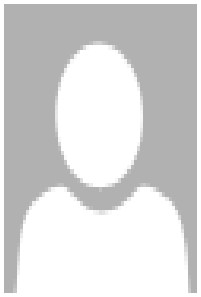
Biography



Emilie Perreau. Certified European Junior Ergonomist and Psychologist. Emilie Perreau is a first year Ph.D. student in Industrial Engineering at the School of Industrial Engineering of Grenoble Institute of Technology within the G-SCOP UMR CNRS Laboratory. Her research focused on the study of human-natural language technology interactions from a human-centered perspective, in an ecological-based context.



Romain Pinquié. Ph.D.-Ing. Romain Pinquié is an Associate Professor in Digitally Mediated Collaborative Engineering Systems Design at the School of Industrial Engineering of Grenoble Institute of Technology and a research fellow at the G-SCOP UMR CNRS Laboratory of Design, Optimisation and Production. His research tries to understand how humans represent and interact with different types of design information, and to develop new human-computer interaction for advancing human-centred computing in engineered systems architecting.



Cédric Masclet. Cédric Masclet got a PhD in mechanical engineering from INSA Toulouse , France in 2002. He joined Grenoble Alpes University (formerly Joseph Fourier University) in 2003 as an associate professor. He successively developed research on collaborative design in 3S and G-SCOP laboratories. He has been involved in several European projects (Visionair, SPARK) dealing with augmented and virtual reality systems for supporting multi-expertise collaboration. He is the head of the Integrated and Collaborative Design team in G-SCOP laboratory and deputy director of Innovacs research federation.