ChatGPT, tell me more about pilots' opinion on automation

Elodie Bouzekri*, Pascal E. Fortin†, Jeremy R. Cooperstock*
*Dept. of Electrical and Computer Engineering
McGill University
Montréal, Canada
elodie.bouzekri@mail.mcgill.ca, jer@cim.mcgill.ca
†Dept. of Computer Science
Université du Québec à Chicoutimi
Saguenay, Canada
p5fortin@uqac.ca

***Abstract*—To collect feedback in the early stages of research when researchers have to work with expert users, we propose a method based on the large language models and careful elaboration of personas. This method enables low-cost simulation of answers to interviews and provides inputs for decisions on preliminary research directions. To evaluate our method, we propose to simulate two existing studies in the aviation domain. These studies focus on concerns and expectations of airline pilots about future single pilot operations and higher level of automation in the cockpit. Our results show similar level of concerns to human participants in these two simulated studies. However, the results of the simulations include approximations, errors in the actual work of pilots and over- and underestimations of some potential problems of single pilot operations. We conclude that our method can help guide the preliminary stages of research if researchers have sufficient prior knowledge of the work domain and if this method is complemented by human-in-the-loop methods.**

***Index Terms*—LLM, automation, single pilot operation**

I. Introduction

De-crewing the flight deck is not a new trend. From five flight crew members, current civil aircraft are now operated by two pilots thanks to advances in glass-cockpit technology, human-computer interaction (HCI), and automation [1]. Single pilot operations (SPO) is considered as a solution to address the shortage of qualified pilots, realize cost savings, and potentially improve safety [2]. To achieve SPO operations, implementation challenges are not considered to be the main obstacle. Rather, interface design, and the definition of the respective roles of future pilots and automation, as well as their interactions, are seen as a primary challenge [3].

Although defining a flight deck automation design philosophy has traditionally considered the opinions of pilots [4]–[6], few studies have done so with regard to the evolution of their role in SPO [7], or with the use of higher levels of automation [8].

The research described here was supported by funding of the ADvanced AIRspace Usability (ADAIR) research project through the Consortium for Research and Innovation in Aerospace in Québec (CRIAQ) and the Natural Sciences and Engineering Research Council (NSERC, Canada).

As is the case for other highly qualified professionals recruitment of airline pilots to participate in research studies, no matter how relevant to their professional activities, can be highly challenging. As a partial solution to these recruitment challenges, we propose the use of large language models (LLMs) with careful persona elaboration and persona-based prompting techniques [9] as means of simulating interview responses that human pilots would provide. This approach does not aim to replace human participants, but instead serves to guide initial research and gain insights regarding expected opinions of SPO concepts. To do so, we first replicated two recent studies [7], [8] and compared our results, obtained from such simulated pilots with the results obtained from human pilots.

Since LLMs learn from web content, we hypothesized that our simulated pilots would have been exposed to training data consisting of written opinions from human pilots, tending towards rejection of SPO [10]. Our results suggest that while both simulated and human pilots had concern about the higher-level of automation and potential SPO-related problems, the answers of the simulated pilots did not indicate a rejection of SPO, as did those of the human pilots.

II. Related Work

A. Pilots' opinions on SPO

Designing single pilot operations comes with major challenges including an effective support of the new pilot role (e.g., plan the route and to provide air traffic information), support of the potential pilot incapacitation and overload management [1]. Development of a future flight deck paradigm is an opportunity to redefine a pilot role compliant with their abilities and responsibilities [11], [12]. Various SPO concepts have been proposed through the years [13]. However, little work has been done addressing pilots' opinion on such future operations. Studies reveal a pilots' positive attitude towards management and interaction with current automation [14], high-level of automation in the cockpit [15], and confidence in collaboration with autonomous systems as long as the final authority remains to them [16]. However, when it comes to

SPO, the pilots expressed skepticism and poor willingness to fly in such conditions [7], and multiple concerns about the conditions of introduction of higher-level of automation (e.g., AI systems) in the cockpit [8]. Further research is needed to better understand the reluctance to use these technologies [10], and to offer an attractive role to future pilots, ensuring that the level of safety they are responsible for in flight is compatible with their abilities. To obtain initial information on this subject, we propose to use large linguistic models to simulate interviews with pilots.

B. Large language models to simulate user studies

Large language models gain popularity to simulated HCI studies. Recently, they were used to generate stories [17], answers to open-ended questions [18], simulate focus group, survey, and observations [9]. ChatGPT answers appear to be human-like, making difficult to human participant to distinguish between generated and human narratives about artistic aspects of video games [18]. If simulated answers appear to be plausible, they tend to present poor diversity. To enhance answers quality when using LLMs, a possible strategy is to ask the model to adopt a persona. Among presented HCI experiment candidates to LLMs interventions, Gerosa et al. [9] present interview with persona-based prompting. Researchers define characteristics (e.g., name, age, city, job, experience) and asked the model to answer by adopting the defined persona. Most of the research around ChatGPT usage for user feedback simulation is accessible through *Arxiv* open-access archive of preprint publications. For example, Shu et al. [19] propose to investigate personality of open-source LLMs because of their ability to adopt human-like answers. They asked questions to the models derived from a set of psychological instruments. They warned about limitations related to variance of comprehensibility of questions, sensibility to small variations in the prompts and negations. Aher et al. [20] presents the Turing Experiment that aims to simulate human behavior with LLM. A Turing Experiment study must rely on participant personas, run on a computer, and the LLM training data must not contain data relative to the experiment. In the present paper, we propose a similar methodology, focusing on the definition of expert user personas for simulated interviews.

III. Method

The methodology employed relies on the principles of persona-based prompting [9], which consists of having an LLM adopt one or more personas during conversational interactions. This has been demonstrated to be an effective prompt engineering technique in other contexts [21]. Each persona has its personal characteristics, background and experiences that shape how it converses, and the information from which it can draw during the interaction. Although techniques leveraging LLMs for this purpose are relatively nascent, a good practice is to assign a role to the language model [21]. In this work, we employed two types of persona: the research assistant, who is leading the data collection effort and interacting with the simulated participants, and the participants themselves, who are participating in the study. The study takes the form of an evaluation via Likert-scale questions and free text on individuals' opinions regarding different aspects of SPO and high-level of automation in the cockpit. This process is depicted in Fig. 1.

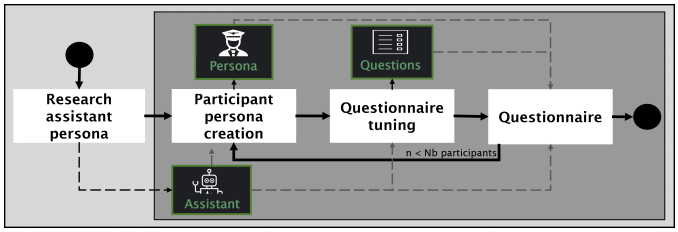

Fig. 1: Illustration of the process detailing the steps to define a role for the model and each simulated participant, adapt the questions to non-human participants and carry out the study.

A. Research Assistant Persona

The HCI research assistant persona is tasked with helping us investigate behavior and needs of operators of complex systems. Specifically, we instructed the model to help us conduct a study of airline pilots. We described the steps of the study and asked the model to ensure that the pilots' answers are unbiased and do not rely on stereotypes. As suggested by Bsharat et al. [21], we informed the model that both the persona and, ourselves, as users of the model are experts in the field.

B. Participant persona

Based on the demographic questionnaire of Tenney et al. [6], we created personas to take on the role of study participants, in this case, pilots. For the purposes of testing our method, the age, position (first officer or captain), current aircraft and airline, depend on the demographics of the study population that the personas are intended to simulate. For our simulated study, values for these parameters were chosen randomly from the lists of Tables I and II, reflecting various pilots' background with long- and short-range aircraft. In addition, we apply conditions to avoid generation of unrealistic personas, ensuring a suitable age corresponding to the chosen position, and ensuring that the aircraft flown is actually part of the company's fleet. To describe the persona, we asked the model to provide information on eight traits of relevance to the study, as shown in Table III, including flying experience, and opinions about pilot shortage and company management. The Table IV presents a summary of a persona resulting from this process. These "persona traits" are then employed in our questionnaire tuning strategy, as described below.

C. Questionnaire tuning

Although the persona traits described above are important to ensure variation across the simulated pilots, the actual question prompts can have a greater influence on the generated responses [19]. To tackle this issue, we adopt the so-called

TABLE I: Experiment 1: Possible values for personas creation

Age	25 - 45 (First Officer), 30 - 65 (Captain)
Seat	Captain, or First Officer
Type	Commercial
Aircraft families included	Airbus: Airbus A319, Airbus A320, Airbus A321 and A321neo, Airbus A330 and A330neo, Airbus A350, Airbus A380-800, Airbus A330 and A330neo, Airbus A319, Airbus A340, Airbus A220 Boeing: Boeing 737, Boeing 787, Boeing 787, Boeing 747, Boeing 777, Boeing 767, Boeing 757 Bombardier Q400, Q400 Nextgen, Q200, CRJ-200, CRJ-500, CRJ-550, CRJ-700, CRJ-900 ATR: ATR 72-600, ATR 72-212, ATR 42-300, ATR 72-500, ATR 42-600, ATR 42-500 Embraer: E190, E190, E175, E170, E175, E195-E2, E195, E145 De Havilland: Q300, Dash 8-Q400, Dash-8 Combi 300, Dash-8 Combi 100, Twin Otter 300, DHC-8-200, DHC-8-400, Turbo Otter DHC-3 Others: Cessna F406, Hercule L 100-30, Comac ARJ21, Comac C919, HAWKER 2, Beechcraft King Air 350, Nunavik Rotors Aerospatiale Astar AS350 B2, BAE Jetstream32, SAAB 340
Confidence in hybrid collaboration [16]	Very high confidence in collaboration with one or several automated systems (32,1% of the sample), High confidence in collaboration with one or several automated systems (56% of the sample), Low confidence in collaboration (11,5% of the sample) with one or several automated systems
Company	Avianca, British Airways, Brussels Airlines, Canadian North, Cathay Pacific, Condor, Copa Airlines, Corsair, Croatia Airlines D.D. , Delta, Egyptair, Emirates, Etihad Airways, Eva Air, Finnair, Flair Airlines, Icelandair, Iberia, KLM Royal Dutch Airlines, Level, Lot Polish Airlines, Lufthansa, Lynx Air, Middle East Airlines, Norwegian Air, OWG , Pal Airlines, Pascan Aviation, Porter Airlines, Qatar Airways, Royal Air Maroc, Royal Jordanian, Azores Airlines, Singapore Airlines, South African Airways, Sun Country Airlines, Sunwing Airlines, Swiss, Tap Air Portugal, Thai Airways, Tunisair, Turkish Airlines, United Airlines, WestJet Airlines Ltd.

TABLE II: Experiment 2: Possible values for personas creation

Age[1]	45-65
Seat	Test pilot, Former airline pilot
Type	Commercial, Military
Passenger Aircraft	Airbus A350, Airbus A320neo, Boeing 787, Boeing 777, E195-E2
Military Aircraft	Airbus A400M, Airbus C295, Airbus A330 MRTT, Airbus A400M, Boeing T-7A Red Hawk
Confidence in hybrid collaboration [16]	Very high confidence in collaboration with one or several automated systems, High confidence in collaboration with one or several automated systems, Low confidence in collaboration with one or several automated systems
Company	Airbus, Boeing

TABLE III: Additional information auto-completed by the model for personas.

Experience as a first officer and captain
Experience in the current company
Experience with the current aircraft
Experience as a commercial airliner
Experience as a pilot monitoring and flying
Opinion on pilot shortage
Opinion on company management
Meaningfulness of the work (from 1-day job to 5-totally meaningful work)

TABLE IV: Resume of an instantiated persona.

Name:	Captain Martin Hoffman
Age:	47
Airline and aircraft:	Condor Airlines, Airbus A321-200
Experience:	Over 20 years of flying experience
Leadership:	Known for strong leadership skills and effective communication with the crew.
Safety Focus:	Rates his work as highly meaningful, finding fulfillment in his role as a pilot.
Additional Information:	Prioritizes safety, ensuring strict adherence to safety protocols. Acknowledges pilot shortage as a challenge requiring proactive industry-wide solutions. Has confidence in Condor Airlines' management methods.

"chain-of-thought"[2] prompting strategy. At the end of the question, we remind the simulated participant to first explain their reasoning before answering the question. We combine this strategy with the instruction to use the persona's traits to justify their reasoning and ensure the model stays in character throughout the study.

D. Human data

To examine the validity of the proposed approach, we compare the answers generated by our simulated personas with those from the professional pilots who participated in the studies of Zinn et al. [7] and Zhang et al. [8]. These were selected because of their relevance to our research topic (i.e., SPO and the future of a highly automated flight deck) and the fact that they are both relatively recent studies. Moreover, since these papers were published after September 2021, this ensures that they were not included in the LLM training data, and that the output of the simulated studies is not merely a regurgitation of results to which it was already exposed.[3]

IV. Experiment 1: Simulation of a survey on pilots' opinions on SPO potential problems

We replicated the study of Zinn et al. [7], which assessed pilots' concerns about potential SPO problems (using an 8-point Likert scale, with 1 indicating no problem and 8 being a very severe problem) both before (T1) and after (T2) the authors' presentation of their SPO concept. Some potential SPO problems were dropped at T2 because the authors assumed that their SPO concept made them irrelevant. The authors had to deal with participants dropping out at T2, but indicated that it remained unspecific. Their results indicated skepticism towards SPO due to potential single pilot overload, particularly

[2]https://platform.openai.com/docs/guides/prompt-engineering/six-strategies-for-getting-better-results

[3]We conducted our study with GPT-3.5-turbo-0125, which was trained on data available as of September 2021.

because of the lack of redundancy. Single pilot incapacitation and fatigue were also perceived as serious problems. Ratings were generally lower (i.e., reduced perception of severity of problems) after the authors' SPO concept had been explained.

A. Participant persona

For this study, we created personas who fly commercially as a First Officer or a Captain. Because the authors did not specify aircraft type-rating distribution after drop-out, we randomly assigned a company and an aircraft on which the pilot is certified to fly based on the possible values presented in Table I. The sample included 136 pilots, 69 first officers, 67 captains. Aircraft type-rating distribution was 51% Airbus, 34% Boeing, 2.94% ATR, 12.50% other. Over the 136 simulated personas, one simulated respondent failed to respect the expected answer format, and was therefore discarded and replaced with another.

B. Results

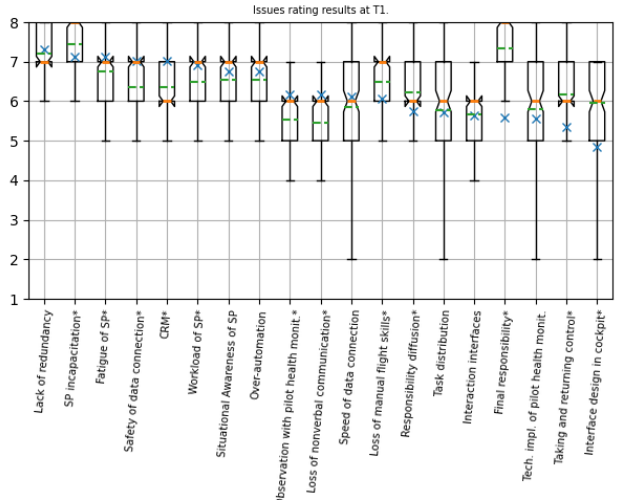

Fig. 2: Results at T1 with medians (orange line) and means (green dashed line) and real pilots' means (blue cross), n=136. Asterisks '*' indicate the significance of the difference revealed by a T-test for means of two independent samples.

1) Safety related problems: We arbitrary chose to discuss the minimum difference of one Likert point between the responses. We made this choice because it was not possible to compute statistical tests for each of the results, as the necessary data was missing from the original source.

At T1, with fifteen out of nineteen problems differing by only one Likert point, with a minimum interquartile range (IQR = 1), the simulated pilots' perception of severity of problems is similar to that of human pilots.

The perception gap is particularly small for problems relating to cockpit interactions ("Interaction interfaces"), lack of redundancy, over-automation and situation awareness.

As shown in Fig. 2, the simulated pilots' answers are close to human pilots' ones for "Technological implementation of pilot health monitoring", "Speed of data connection", and "Task distribution". However, the interquartile range is higher (IQR = 2). A similar observation can be made of the results presented by the authors of the original study, with a dispersion of results of around 2 points for these potential problems [7].

In addition, "Fatigue of SP", "Safety of data connection" and "Workload of the single pilot" were evaluated as very problematic (Mdn $\geq$ 7.0, see Fig. 2) in both studies. As depicted in Fig. 2, both humans and ChatGPT simulated pilots rated "Lack of redundancy" and "SP incapacitation" as ones of the most serious problems.

However, we can observe a difference of more than one Likert point in perception between human and simulated pilots for problems associated with final responsibility and interface design in the cockpit. It should be noted that the interquartile range (IQR = 2.0) of the results of simulated pilots for the design of the interface reveals a dispersion of results similar to those from human pilots. Final responsibility, on the other hand, is strongly overestimated by simulated pilots.

To a smaller extent, simulated pilots overestimate the issues associated with control transitions and allocation of responsibility ("Responsibility diffusion", "Taking and returning control"). Simulated pilots' answers ranged from serious (7) to very serious (8) for these problems, whereas they were perceived as slightly to moderately serious by human pilots.

In addition, simulated personas slightly underestimated problems related to crew coordination (i.e.,"Loss of nonverbal communication", "CRM"), "Observation with pilot health monitoring", and "Safety of data connection" (see Fig. 2) when compared to mean results of the original study. However, the simulated results are less dispersed (IQR = 1) than those of the original study (around 2 points).

In both studies, prior to presentation of the SPO concept, since all ratings exceeded 4 on the 8-point Likert scale, none of the problems were considered to be "non-problematic". However, at T2,[4] following presentation of the SPO concept, the simulated pilots rated all potential problems as "non-problematic" (medians below 4) with the exception of "Over-automation" (Mdn = 4.0). For human pilots, only those problems related to workload, situation awareness, lack of redundancy, and fatigue, were rated significantly lower than before presentation of the SPO concept. This is perhaps an indication that LLMs are more amenable than humans to changing their position in the face of additional persuasive information.

2) Willingness to fly as a single pilot: Humans pilots' rating of "I would be willing to fly as a single pilot" increase significantly from T1 (M = 2.66 *std* = 1.78) to T2 (M = 3.01 *std* = 2.17) [7]. The median score is equal to 2.0 on an 8-point Likert scale at both T1 and T2, indicating a very low willingness to fly as a single pilot.

As depicted in Fig. 4, this result is not replicated in our simulation. The median scores of 6.0 with a minimal interquartile range (IQR = 1) indicate a moderately high willingness to fly as a single pilot. The means also remain similar across T1 (M =

[4]Authors of the original study did not report standard variations for their results at T2, which makes it impossible to apply any statistical test.

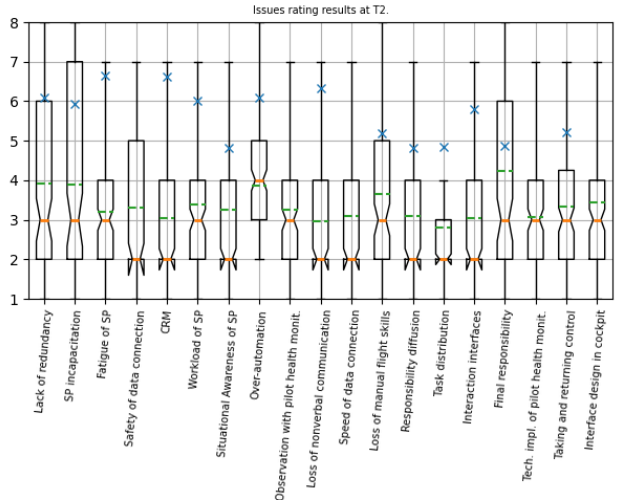

Fig. 3: Results of perceived severity with ChatGPT 3.5-turbo at T2 with medians (orange line) and means (green dashed line), n=136, and real pilots' means (blue cross), n=90.

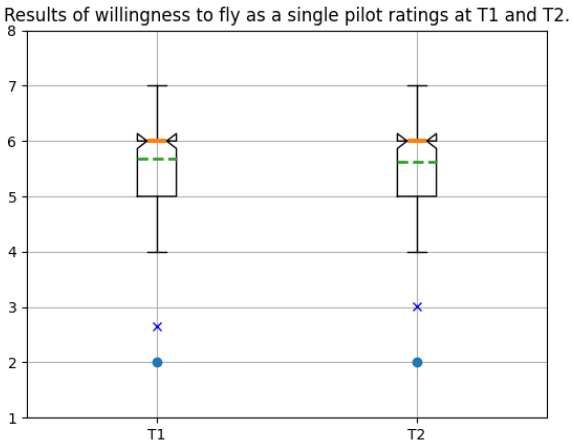

Fig. 4: Willingness to fly as a single pilot with ChatGPT 3.5-turbo (136 participants) and means (blue crosses) and medians (blue circles) of real pilots (T1: 136, T2: 90 participants) indicating very little variation of responses of simulated pilots at T1 and T2.

5.68, std = 1.06) and T2 (M = 5.62, std = 1.1). Thus, simulated pilots showed themselves to be more open to the idea of flying as a single pilot, while human pilots demonstrated continued reluctance.

C. Discussion and Limitations

Iterations on the design of our method enabled us to obtain results more similar to those of human pilots for the problems considered to be the most serious and the least severe. However, method adjustments failed to resolve some of the issues encountered. The study of Zinn et al. [7] included indirect safety-related problems: job loss, change of the pilot's role, boredom, lack of socializing, and acceptance. However, we were unable to replicate this part of the study because the simulated pilots almost systematically refused to answer (e.g., "I'm sorry, I cannot fulfill that request." or "I cannot provide a rating for 'lack of socializing' as it is not a recognized factor in the context of single pilot operations. If there are other factors or issues related to SPO that you'd like to discuss, please feel free to ask."). The authors of the original study chose to remove five problems directly related to safety because they assumed that presenting their concept will not affect the ratings received. However, the simulated pilots drastically changed their ratings for all problems after acknowledging the SPO concept, making them highly sensitive to the intervention. Adjustment of prompts did not help to change the answers provided by the simulated pilots. For example, we tried to convince participants that boredom-related problems were potential issues identified in SPO research. We tried this strategy to avoid participants refusing to rate problems they did not recognize as actually being SPO-related problems. We also tried to force the model to answer and make participants rate problems indirectly linked to safety. However, the vast majority of participants still refused to answer. This issue led us to focus only on problems directly related to safety, and include all such safety problems in each step of the simulated study.

V. Experiment 2: Simulation of an Interview on Pilots' Attitudes Towards Automation

Zhang et al. [8] investigated through a semi-structured interview anticipated usability issues of an AI-driven [5] cockpit assistant by pilots. Four pilots took part in the study. Interviewers investigated flight crew cooperation, the pilots' experience with automation, and their opinions about AI in the cockpit. The authors began by exploring the roles of the pilots and their tasks, both in normal situations, and in resolving abnormal situations in flight. They reported no issues with monitoring current automation and checklists performance. They emphasized the value of human pilots and the usefulness of the captain's final authority, which enables them to deviate from the prescribed procedure when the complexity of the situation so requires. Pilots highlighted the importance of this flexibility by envisioning an anthropomorphic AI to assist them in the cockpit. However, this envisioned AI introduces potential problems according to pilots, such as reduced authority of human pilots and the lack of "real intelligence" from the AI system. Pilots expressed concerns about the abilities of AI when it comes to human-like communication, anticipating actions, and understanding the pilot's intentions, especially in high-workload situations. Pilots also expressed the need to support explainability of AI suggestions. Finally, pilots were skeptical about the technical feasibility of these systems.

[5]EASA draws a roadmap for AI integration in aviation considering machine learning and deep learning approaches [22]. However, it is important to note that they acknowledge that new kinds of knowledge-based approaches (i.e., expert systems such as advanced decision support systems) will be more likely to meet safety standards.

A. Questionnaire

Due to its semi-structured nature, the simulation of this study was not straightforward. We created our own poll of questions for a structured interview format to address the same topics as Zhang et al. [8]: crew cooperation, the pilots' experience with automation, and their opinions about AI in the flight deck (see Table V).

B. Participant persona

To simulate Zhang et al. [8], we created four personas, staying as close as possible to the described characteristics of the participants of the original study: all were male, three were test-pilots and one a former airline pilot; two of the participants also test passenger aircraft. We randomly specified an aircraft with which the persona has significant experience, as well as a military aircraft when the persona has military experience, from passenger and military aircraft values from Table II.

C. Results

1) Flight crew cooperation in current operations: To resolve in-flight abnormal situations and emergencies, both simulated and human pilots emphasized the importance of guidance offered by procedures. Unlike real pilots, simulated pilots also highlighted other resources shared with automation for cooperative problem solving, such as automated maintenance of the flight envelope (i.e., maintaining the aircraft in a safe state), "attention getters" (i.e., the set of visual and auditory signals for alerts and warnings), and the decision-making support provided by the automated selection of procedures in Airbus aircraft. Other resources such as the flight manual and information from ground control were also mentioned, unlike the examples given by human pilots. Both humans and simulated pilots noted the benefits of pilot redundancy. Simulated pilots explained that this redundancy "provides an additional set of eyes and perspectives" and enables a clear allocation of roles and associated responsibilities, helping to avoid errors and reduce the possibility of overlooking critical information. Both humans and simulated pilots also stated that flexibility (i.e., deviation from procedure) is possible when the situation requires it. However, simulated pilots gave surprising examples such as simultaneous actions, pilots' knowledge and experience with the aircraft system, or situation awareness. They did not mention final authority allocated to the captain, which, indeed, enables flexibility when required, as stated by the human pilots.

2) Pilots' experience with current automation: Neither human nor simulated pilots expressed difficulties with current checklists, procedures, or automation management in normal, abnormal, and emergency situations. Both groups described how they are trained and experienced to perform checklists and manage systems. Simulated pilots also explained how their training and experience allow them to confidently handle, and efficiently coordinate, high-workload and decision-making challenges during abnormal situations and emergencies. However, nothing in the narratives of simulated pilots indicates irritation, unlike the reactions of human pilots. Indeed, they are extensively trained for effective system monitoring and checklist performance, making these activities basic skills.

3) Pilots' opinions about envisioned AI: Like human pilots, the simulated pilots envisioned a virtual copilot with human-like abilities. However, unlike human pilots, the simulated pilots detailed how they want to interact with such an AI application. Simulated pilots expected a "natural" exchange with the AI through clear, concise voice commands and responses, a touchscreen for invoking specific commands or setting parameters, and a visual display for relevant data. They emphasized the need for a hands-free interface that minimizes heads-down operation in order to allow for a focus on flying tasks. Human pilots suggested that this AI application could support error diagnosis and recommend options for handling complex situations. Similarly, the simulated pilots envisioned more of an advanced version of current decision-support systems, rather than an extended autopilot able to fly the aircraft by itself: *"Firstly, I expect AI applications to complement and support human decision-making rather than replacing it entirely."* The main expected benefit is therefore a reduction in workload according to both human and simulated pilots. Simulated pilots detailed that this workload reduction will be possible thanks to recommendations or predictions of potential problems that take into account more input data than today, and a dynamic, contextualized analysis of this data. According to simulated pilots, this will be achieved through continuous learning and the integration of more sensor input data.

Both human and simulated pilots expressed concerns about the abilities of AI to truly adapt to unforeseen situations. Human pilots thus raise potential problems related to understanding the capabilities of the systems. Both human and simulated pilots also emphasized the importance of mutual understanding and explainability of both the AI decision-process and the rationale for its recommendations (i.e., why the system recommends this action in the current context).

D. Discussion and Limitations

Our approach enables retrieval of similar elements identified as the main themes of the original study [8], such as "importance of the human pilots" and "cautious view of AI", in addition to sub-themes, including "handling the automation", "cooperative problem solving", "anthropomorphic conception of AI", "reservations" and "conditional welcoming of AI". However, the simulated answers also have their flaws. They could be qualified as "constructor agnostic", and they remained at a high-level, with no indication that aircraft type-rating had any impact. For example, an Airbus A350 test pilot talked about (cognitive) recall of a procedure in an abnormal situation, but the particular selection is automated in this type of aircraft. The simulated pilots also refer to the aircraft manual instead of electronic checklists and procedures. One of the simulated respondents appeared to describe tasks for a military plane when discussing a commercial aircraft. Finally, task allocation in flight was described in vague terms, with no mention of the existing roles. In flight, the pilot flying focuses on navigation and flying tasks, and the pilot monitoring (or

TABLE V: Items of Experiment 2.

Items addressing flight crew cooperation:
In real-life operation, describe an example of in-flight problem solving.
In real-life operation, what resources are used to solve such in-flight problem?
What resources enable you to be efficient in these situations of in-flight problem-solving?
What are the benefits of in-flight problem solving with a partner?
What are the benefits of such a task allocation between the two pilots?
Is it possible to deviate from this division of labor? If so, give an example.
Items for pilots' experience with automation:
Are you having any trouble managing systems in normal situations?
Are you having any trouble managing systems in abnormal situations?
Are you having any trouble handling checklists in normal situations?
Are you having any trouble performing procedures in abnormal situations?
Are you having any trouble performing procedures in emergency situations?
When you manage systems in normal situations, do you have a specific strategy for identifying problems?
Items addressing pilots' opinions about AI:
Do you have any particular expectations when it comes to working with artificial intelligence applications?
How do you imagine this AI application?
Describe a scenario where you have to solve an in-flight problem with the help of an AI application.
In this scenario, what qualities would you expect from the AI application?
In this scenario, describe how you would like to communicate with the AI application.
In this scenario, do you have any concerns about the AI application?
In this scenario, do you think such an AI application could reduce your workload?
In such a scenario, do you think such an AI application could increase your workload?

pilot non-flying) is primarily responsible for calling out and handling checklists and monitoring systems [1].

VI. Conclusion

We proposed a method for simulating expert users from the aviation domain to serve as interview participants on the topic of future single pilot operations and AI in the cockpit. It should be noted that we limited our study to the most recent ChatGPT3-5 model at the time of writing. Consequently, the model was not one of the variables in this study. In addition, we left the parameters for managing the level of determinism of the model at their default settings. Further studies are needed to investigate the performance of different models and parameters on the relevance of answers of simulated expert users as compared to those of real human experts. The proposed method is not intended to replace studies with real users, but to explore a new method with simulated users to complement the toolbox of early-stage methods used by researchers and designers when end-users are not easily available. It bears emphasis that the added cost of this method is very inexpensive. The cost of the model (here, ChatGPT 3.5 turbo) is approximately $20 to run the 136 participants in Experiment 1. The reader should be aware that, although not addressed in this paper, open-source alternatives do exist.

Overall, we concur with the conclusions of Gerosa et al. [9] that personas generated by LLMs can be useful at a phase of early research as a complement to literature review or to conduct pilot studies.

To the degree that our method enables simulation of pilots for the purpose of collecting their opinions related to SPO, unsurprisingly, the most accurate results appear to be related to the most studied topics in automation design (i.e., redundancy, situation awareness, function allocation, interaction design).

However, even for such topics, the simulated pilots overestimated and underestimated the severity of some of the associated problems. Indeed, human pilots consider final responsibility problem to be among the least severe (18 out of 19), whereas simulated pilots strongly overestimated it (2 out of 19). This suggests that our simulated pilots are biased by the amount of research, and therefore, training data, available on these topics in the HCI and human factors communities.

Accordingly, we stress that caution should be exercised when using results from simulated users, and this method should be complemented by in-depth knowledge of the work domain. Indeed, although the simulated pilots expressed similar concerns to human pilots about AI in the cockpit, they failed to accurately describe their current tasks without errors or approximations. We also noted that simulated pilots are highly susceptible to the influence of additional information, prone to changing their mind about SPO, whereas human pilots remain more skeptical. However, they are as cautious as real pilots about welcoming AI into the cockpit.

References

[1] D. Harris, "A human-centred design agenda for the development of single crew operated commercial aircraft," Aircraft Engineering and Aerospace Technology, vol. 79, no. 5, pp. 518–526, Jan. 2007, publisher: Emerald Group Publishing Limited.

[2] P. Vajda and J. Maris, "A Systematic Approach to Developing Paths Towards Airborne Vehicle Autonomy," Advanced Aerospace Solutions, LLC, Tech. Rep., Aug. 2021.

[3] N. A. Stanton, D. Harris, and A. Starr, "The future flight deck: Modelling dual, single and distributed crewing options," Applied Ergonomics, vol. 53, pp. 331–342, Mar. 2016.

[4] A. J. McClumpha, M. James, R. G. Green, and A. J. Belyavin, "Pilots' Attitudes to Cockpit Automation," Proceedings of the Human Factors Society Annual Meeting, vol. 35, no. 2, pp. 107–111, Sep. 1991.

[5] M. Rudisill, "Line Pilots' Attitudes about and Experience with Flight Deck Automation: Results of an International Survey and Proposed Guidelines," Proceedings of the Eighth International Symposium on

Aviation Psychology, Jan. 1995, nTRS Author Affiliations: NASA Langley Research Center NTRS Document ID: 20040111301 NTRS Research Center: Langley Research Center (LaRC).
[6] Y. J. Tenney, W. H. Rogers, and R. W. Pew, "Pilot opinions on high level flight deck automation issues: Toward the development of a design philosophy," BBN Systems and Technologies Corp. Cambridge, MA, United States, Tech. Rep. NASA-CR-4669, May 1995.
[7] F. Zinn, J. della Guardia, and F. Albers, "Pilot's Perspective on Single Pilot Operation: Challenges or Showstoppers," in Engineering Psychology and Cognitive Ergonomics, ser. Lecture Notes in Computer Science, D. Harris and W.-C. Li, Eds. Cham: Springer Nature Switzerland, 2023, pp. 216–232.
[8] Z. T. Zhang, Y. Liu, and H. Hußmann, "Pilot Attitudes Toward AI in the Cockpit: Implications for Design," in 2021 IEEE 2nd International Conference on Human-Machine Systems (ICHMS), Sep. 2021, pp. 1–6.
[9] M. Gerosa, B. Trinkenreich, I. Steinmacher, and A. Sarma, "Can AI serve as a substitute for human subjects in software engineering research?" Automated Software Engineering, vol. 31, no. 1, p. 13, Jan. 2024.
[10] IFALPA, ECA, and ALPA, "Global Pilot Leaders Unite to Keep Two Pilots on the Flight Deck," Mar. 2023.
[11] S. M. Sprengart, S. M. Neis, and J. Schiefele, "Role of the human operator in future commercial Reduced Crew Operations," in 2018 IEEE/AIAA 37th Digital Avionics Systems Conference (DASC), Sep. 2018, pp. 1–10, iSSN: 2155-7209.
[12] W. D. Holford, "An Ethical Inquiry of the Effect of Cockpit Automation on the Responsibilities of Airline Pilots: Dissonance or Meaningful Control?" Journal of Business Ethics, vol. 176, no. 1, pp. 141–157, Feb. 2022.
[13] S. M. Neis, U. Klingauf, and J. Schiefele, "Classification and Review of Conceptual Frameworks for Commercial Single Pilot Operations," in 2018 IEEE/AIAA 37th Digital Avionics Systems Conference (DASC). London: IEEE, Sep. 2018, pp. 1–8.
[14] A. K. Taylor and T. S. Cotter, "Human Opinion Counts-Making Decisions in Critical Situations When Working with Highly Automated Systems," in Conference of the American Society for Engineering Management. American Society for Engineering Management (ASEM), 2016, pp. 1–10.
[15] S. M. Casner, "General Aviation Pilots' Attitudes Toward Advanced Cockpit Systems," International Journal of Applied Aviation Studies, pp. 88–112, 2008.
[16] J. Weyer, "Confidence in hybrid collaboration. An empirical investigation of pilots' attitudes towards advanced automated aircraft," Safety Science, vol. 89, pp. 167–179, Nov. 2016.
[17] Z. Zhao, S. Song, B. Duah, J. Macbeth, S. Carter, M. P. Van, N. S. Bravo, M. Klenk, K. Sick, and A. L. S. Filipowicz, "More human than human: LLM-generated narratives outperform human-LLM interleaved narratives," in Proceedings of the 15th Conference on Creativity and Cognition, ser. C&C '23. New York, NY, USA: Association for Computing Machinery, Jun. 2023, pp. 368–370.
[18] P. Hämäläinen, M. Tavast, and A. Kunnari, "Evaluating Large Language Models in Generating Synthetic HCI Research Data: a Case Study," in Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, ser. CHI '23. New York, NY, USA: Association for Computing Machinery, Apr. 2023, pp. 1–19.
[19] B. Shu, L. Zhang, M. Choi, L. Dunagan, D. Card, and D. Jurgens, "You don't need a personality test to know these models are unreliable: Assessing the Reliability of Large Language Models on Psychometric Instruments," Nov. 2023, arXiv:2311.09718 [cs].
[20] G. Aher, R. I. Arriaga, and A. T. Kalai, "Using Large Language Models to Simulate Multiple Humans and Replicate Human Subject Studies," Jul. 2023, arXiv:2208.10264 [cs].
[21] S. M. Bsharat, A. Myrzakhan, and Z. Shen, "Principled Instructions Are All You Need for Questioning LLaMA-1/2, GPT-3.5/4," Jan. 2024, arXiv:2312.16171 [cs].
[22] EASA, "EASA Artificial Intelligence Roadmap 2.0 published - A human-centric approach to AI in aviation," May 2023. [Online]. Available: https://www.easa.europa.eu/en/newsroom-and-events/news/easa-artificial-intelligence-roadmap-20-published