Does Artificial Intelligence Conduct Better Research Interviews than You?

Jacqueline Eaton (GBR NIC) **Stefan Olaru (ROU NIC)**
Joint Analysis and Lessons Learned Centre (JALLC)
PORTUGAL

jacqueline.eaton@jallc.nato.int stefan.olaru@jallc.nato.int

Key words: Artificial Intelligence / Machine Learning: automated interview; collaboration tools; lessons learned; JALLC; COVID-19

ABSTRACT

Interviews as a research method are a rich source of qualitative data, but they are fraught with problems. First, they are resource intensive; a single 30 minute interview may take up to 3 hours of the researcher's time when taking into account time taken for organization, transcription and follow up. Furthermore, interviews can never be truly anonymous and no matter how experienced the interviewer is, the interviewer may still ask leading questions or bias the interview. In the information age, maybe artificial intelligence (AI) offers a better solution. With an AI interviewer, research interviews can be conducted at a scale that is simply not feasible with human interviewers. Furthermore, AI can be truly objective in its line of questioning. But can AI replace humans in this task?

NATO's Joint Analysis and Lessons Learned Centre (JALLC) trialled an AI voice-bot called DUCHESS to collect staff experiences of using collaboration tools during the COVID-19 pandemic. The JALLC project will inform decisions regarding future collaboration tools for the NATO enterprise. More than 2000 staff in the NATO military structures were invited to an interview. Five seedling questions were provided and then DUCHESS asked its own intelligent follow-up questions based on the responses. The results were automatically transcribed, and presented in PowerBI dashboards[1]. This paper describes the JALLC's experience of using an AI voice-bot instead of humans to conduct research interviews for this project, and discusses the pros and cons of using this novel technology in future JALLC analysis projects.

1.0 INTRODUCTION

The purpose of this paper is to share and reflect on the JALLC's experience of employing a novel AI-driven voice-bot for Lessons Learned (LL) knowledge acquisition as part of an analysis project. The introduction provides an overview of the need for knowledge acquisition as part of a LL capability, the potential for new technologies to improve the LL knowledge acquisition capability, a summary of the use of interviews for LL knowledge acquisition and an overview of how AI and Machine Learning (ML) is being incorporated into interview and survey approaches across different sectors. The analytical methods section introduces the JALLC analysis project and explains how JALLC set up the interviews for the project. The results section discusses the JALLC's experience of using the AI voice-bot for interviews and the paper concludes by answering the question in the title of this paper: "Does AI conduct better research interviews than you?" with

[1] A Power BI dashboard is a single page, often called a canvas that tells a story through visualizations. Because it's limited to one page, a well-designed dashboard contains only the highlights of that story. Readers can view related reports for the details (Source: https://docs.microsoft.com/en-us/power-bi/create-reports/service-dashboards)

the summary of pros, cons, and considerations for NATO to take into account for future use of this new technology.

1.1 LL Capability and New Technologies

Dyson (2020) [1] identifies four key capabilities—derived from literature on organizational learning and knowledge management (Nevis et al. (1995) [2] and Zahra and George (2002) [3])—that an organization must possess in order to learn effectively. Together, these four capabilities contribute to the organization's LL capability, which in civilian terms is also known as Absorptive Capacity: "a firm's ability to recognize the value of new information, assimilate it, and apply it to commercial ends" (Wikipedia, Absorptive capacity, 2021, Absorptive capacity - Wikipedia). [4]

The four capabilities that enable Absorptive Capacity and thus form the basis of an effective LL capability are:

- Knowledge acquisition: The ability to acquire information and knowledge from the operational environment.
- Knowledge management: The ability to store and organize knowledge in a meaningful and accessible way.
- Knowledge dissemination: The ability to quickly and effectively disseminate lessons to key areas of activity in the organization and to get lessons and best practices to personnel on the front line.
- Knowledge transformation: The capability to transform knowledge by establishing and improving the processes which allow and organization to combine existing knowledge with new knowledge.

In 2018, the NATO Joint Analysis and Lessons Learned Centre (JALLC) was challenged by Supreme Allied Commander Transformation (SACT) to actively look for ways in which innovation could be incorporated into NATO's LL capability. That year, the JALLC hosted a New Technologies Event (details in the post-event report) [5] and completed an in-depth study in which the challenges associated with NATO LL were mapped to the various features that new technologies offered in order to identify the most promising areas for development [6]. Out of the four capabilities above, the JALLC identified LL knowledge acquisition as one of the key areas where new technologies have substantial potential to deliver improvements.

1.2 Current Approaches to LL Knowledge Acquisition

Weber and Aha (2001, 20-21) [7] propose that LL processes can acquire knowledge through four main activities:

- Passive collection: Non-LL staff submit knowledge directly to a LL IT system.
- Reactive collection: Knowledge is gathered from staff in the field following a serious incident.
- After-action collection: Staff provide routine reports following training or deployment in which they analyse the successes, failures, and the reasons why things went well or badly.
- Active collection: Dedicated LL staff organize knowledge collection workshops or deploy to the field to gather knowledge about ongoing military activities and problems.

Interviews are a commonly used method for reactive collection, after-action reviews and active collection across both military and non-military organizations. For example, the British Army uses post-operational interviews in both reactive and active collection, and the German Army deploys LL staff officers who can ask questions of troops in German contingents (Dyson 2020) [1]. Similarly, NATO identifies post-event interviews as a "valuable way to capture lessons" (NATO LL Handbook) [8] and the US Center for Army

Lessons Learned interviews troops returning from deployment to capture observations.

The use of interviews outside the military for the purpose of LL knowledge acquisition is also common. For example, the World Health Organization (WHO) uses a Key Informant Interview After Action Review (AAR) method which consists of a longer and more in-depth review of an event combining research of background material with a series of one-on-one interviews with people involved in a response (Ref WHO Facilitators' handbook) [9]. In the private sector, methods like the Learning History involve compiling the knowledge from a project or other piece of work through individual interviews (Ref Knoco website) [10].

1.3 Value of Interviews for LL Knowledge Acquisition

As a research method, interviews are usually used to collect qualitative data because, compared with questionnaires, they are more powerful at eliciting narrative data that allows a deeper investigation of people's views. This is particularly important in LL knowledge acquisition because context is important and while the details about 'what' happened can often be found in documentary sources or extracted from mission systems, the reasons why something happened are often not written down and may be perceived differently by different staff.

Theoretically, there are four types of interviews that can be used: structured, unstructured, semi-structured and focus group (Ref Interviewing as a Data Collection Method: A Critical Review) [11]. Structured interviews are essentially questionnaires administered verbally and in-person. The questions tend to be closed (i.e. elicit yes or no answers), and there is no scope for the interviewer to diverge from the script to explore the answers given. Unstructured interviews allow for free-form conversation and are great for getting in-depth insight, but can be challenging to analyse since there may be few similar features to compare across multiple interviews. Semi-structured interviews take a middle ground and allow for a common overarching structure to the interviews, while giving the interviewer some freedom to explore answers that seem particularly relevant or interesting, or to ask for more details when answers are incomplete or confusing. Finally, focus groups allow for interviewees to bounce ideas off each other and thus provide some insight into group behaviour which cannot be easily ascertained from one-on-one interviews.

1.4 Artificial Intelligence (AI) / Machine Learning (ML) Technologies for Interviews and Surveys

Recent increases in the accuracy and reliability of voice-to-text services, and ubiquitous and cheap access to AI/ML technologies through cloud computing services, allow innovators to imagine a plethora of new ways to make use of these technologies in interviews and surveys.

One particular growth area for AI/ML technologies in interviews is for recruitment interviews. NATO already makes use of an online video interview service to streamline its recruitment process (NCI Agency, 2016) [12], but some organizations are taking this to the next level. Natural language processing, chatbots, sentiment analysis, facial expression recognition and visual perception, speech recognition, tone analysis, and decision-making are all AI features used in recruitment interviews. For example, an AI-based behavioural insights engine can analyse the candidate's natural speech pattern to identify fake responses and assess competencies. Some benefits of such technology include being able to interview more candidates, understand their responses, less effort from human recruiters, faster recruitment timelines, increased suitability of shortlisted candidates, and reduction of racial and other biases in the hiring process (How AI Interviewing is Redefining the Way we Hire (talview.com)) [13].

Another emerging area is the use of chat-bots for candidate pre-selection and in customer service/helpdesks. A chat-bot analyses the answers given by humans and provides intelligent responses. Typically chat-bots will listen/look for certain keywords and then use case-based reasoning to decide which response is most appropriate from a pre-defined list. However, newer applications can also listen/look for sentiment in

answers and respond appropriately. The chat-bot can be implemented via a written chat interface or using voice to text technologies over the phone, for example (https://developers.google.com/chat/concepts/bots) [14].

Finally, AI/ML is starting to be used in customer feedback/insight surveys and interviews. This application is the closest in nature to the task of LL knowledge acquisition. Several tools for this purpose offer AI/ML analytics of answers provided to pre-planned question sets (e.g the SurveySparrow chat-bot) and some tools are also applying AI/ML to adapt the question set in real time during the interview (https://towardsdatascience.com/complete-guide-to-building-a-chatbot-with-spacy-and-deep-learning-d18811465876) [15].

2.0 ANALYTICAL METHODS

This section describes the JALLC analysis project which provided the context for the use of the AI voice-bot as a LL knowledge acquisition tool. The section also includes the methodology for incorporating the tool into the analysis project, starting from the procurement, to the set up and testing of the tool, to collection, extraction of results and the and analysing.

2.1 Collaboration Tools[2] Analysis Project

The COVID-19 pandemic had a wide reaching impact on work practices globally. Specifically, many nations implemented confinement plans or "lockdowns" where the population was either mandated or recommended to work from home/ telework if possible. Such policies also applied to NATO entities, and both military and civilian staff across NATO became reliant on collaboration tools to perform their functions. NATO's Communications and Information Agency (NCI Agency) was quick to respond by rolling out increased network capacity and collaboration tools across its staff. However, it was recognized that in the longer-term, enterprise-wide solutions are required and that these solutions must satisfy the needs of staff working in both the political and military part of the Alliance.

Initially, NCI Agency was able quickly to collect user requirements for collaboration tools from staff working on the political side of the Alliance since these staff are mostly co-located at NATO HQ in Brussels. Moreover, the nature of the work conducted by these staff members is broadly similar across various sections and divisions as it relates to consultation with nations, support to committee meetings, and general office work. On the other hand, NATO's military HQs and agencies are geographically distributed across the entire North Atlantic region and are therefore more difficult to reach. Furthermore, there is greater variability in the range of tasks conducted by the military HQs and agencies, including everything from running military exercises to procurement activities, to intelligence analysis, and command and control of deployed forces.

The JALLC was requested to look at staff experiences of using collaboration tools during the COVID-19 pandemic within the NATO military HQs and collect challenges and best practices in order to inform the military requirements for collaboration tools within the NATO enterprise.

2.2 Methodology

This JALLC analysis project required the JALLC to acquire LL knowledge via active collection. Typically, JALLC uses semi-structured interviews for this purpose since such interviews offer a good balance between

[2] In the context of this project, Collaboration Tools can thought as of applications/systems that allow you to work collaboratively with others in a virtual environment. For example SharePoint, Skype for Business, Google Docs, or WhatsApp.

ensuring that the relevant topics are covered, and allowing the researcher/analyst to take advantage of the rapport formed through a face-to-face interaction to dig into areas of interest. However, in this case, the number and diversity of potential interviewees led the JALLC to consider other options. One potential option was to collect the knowledge via a survey, but survey, response rate can be very low and the data collected via survey tends to lack detail, so the JALLC decided to trial an automated interview solution instead. The overall methodology was designed around this desire to incorporate an automated interview system into the project and included five steps described below.

2.2.1 Step 1: AI/ML Procurement

The first step was to engage an AI/ML provider. DIEM Innovation was selected for their tool DUCHESS, which at the time of this report, is unique in the way it combines technologies of voice-to-text, real time answer analysis to generate intelligent follow-on questions, and analytics of the results, and. Overall, DUCHESS was therefore judged to be closest way to replicate semi-structured interviews at scale using an automated tool.

The contract was set up in two parts: (1) for the use of the software platform, (2) for consultancy services to gain best advantage of the software. The inclusion of consultancy from DUCHESS experts was critical to ensure that the tool provided the necessary results that would meet the project needs. The use of consultants is a lesson the JALLC analysts learned from previous data science projects; although the costs associated with using software platforms can be very low, the true cost to gain full benefit from the systems should factor in the need for consultancy from the system experts.

DUCHESS runs in the Microsoft Azure cloud, so a key consideration when writing the contract was to include terms and conditions regarding data collection, storage and processing in a third party system. In the case of this project, the data being collected was not sensitive or classified and therefore collection could occur on the Internet. Going forward in event that sensitive or classified data must be collected, such as in conducting post-deployment interviews, it would be worth setting up a DUCHESS instance within a NATO secure cloud environment. In addition to setting up and launching the DUCHESS software, the JALLC requested the consultancy services to deliver tailored PowerBI dashboard summaries of the results, the transcripts in human readable form, and a summary analysis report.

2.2.2 Step 2: Interviews Set Up

DUCHESS is extremely quick to set up. The system needs the user to define around five “seedling questions” which provide the common structure for each interview. Then DUCHESS decides in real time, based on the answers given by an individual interviewee, whether to ask a follow-on question, and what that follow-on question should be: either generic (e.g. please could you tell me more?”) or specific (e.g. “you mentioned ‘email attachments’ - could you tell me more about that?”). The follow-on questions enter the interview flow as presented in the figure below[3] [16]. Depending on the length and quality of the answer to previous questions, DUCHESS may ask between zero and three follow-on questions for each seedling question [16].

[3] Sarah Vicent-Major (2021), DIEM Innovation, presentation on DUCHESS during NATO Lessons Learned Conference 2021

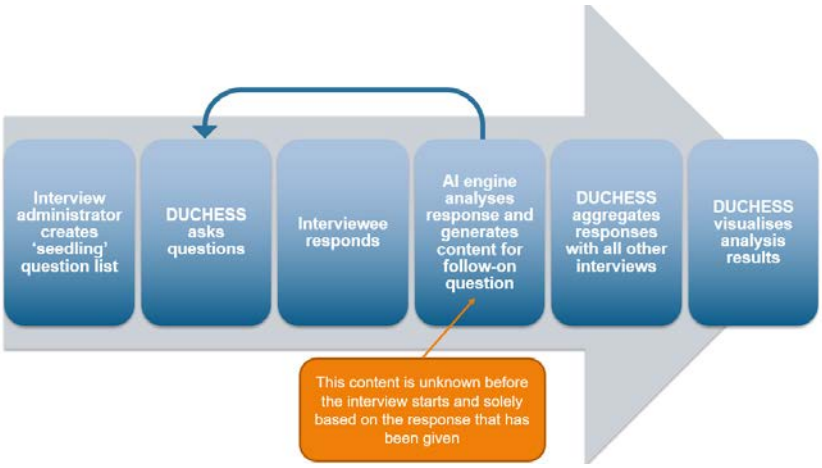

Figure 2–1: Follow-on questions within interview flow

For this project, the JALLC developed the following seedling questions in consultation with the project customer representative and the DUCHESS experts:

- Question 1: During the pandemic, when you needed to collaborate with others or access NATO information, what kind of challenges did you face in your job?
- Question 2: What was your experience of using collaboration tools to overcome your challenges?
- Question 3: Thinking about how you and others used collaboration tools for your job during the pandemic, what were your best and worst experiences?
- Question 4: Thinking about all the collaboration tools that you know about (at home and at work), which of their functions would you want for your job?
- Question 5: How do you feel about the culture of virtual collaboration at NATO?

A final Question 6 was included to give the interviewee an opportunity for final comments: This is your final interview question, is there anything else you would like to add?

The questions were designed to be as open-ended as possible and to elicit an emotional response if possible, not just a factual description, so that DUCHESS's sentiment algorithm would be able to detect anything respondents felt strongly about as a topic for a follow-up question.

An additional consideration was to collect some structured demographic data from the interviewees. This was not possible natively within DUCHESS and therefore it was decided that the JALLC would set up a post-interview questionnaire to capture a minimal set of information, such as which organization the interviewee was from, their age range, and gender. Furthermore, this post-interview survey was used to collect some information about the interviewees' English language skills and set up of the devices used during the interview as well as any feedback they wanted to give about using the DUCHESS system that could help inform this paper.

2.2.3 Step 3: Testing & Collection

The steps to launch DUCHESS for the data collection resembled the way that surveys are implemented (i.e. sending a link to audience but, when they open the link, they are talking to a bot). In this sense it was important to test the system, not just from a technical perspective, but to find out if the wording of the seedling questions was yielding the right kind of answers to support the study.

Around 250 people from a cross-section of NATO entities were invited to do a test interview and around 20 test interviews were completed. This was a disappointing response rate, but the answers provided were enough to give confidence in the system and the seedling questions and to reveal some technical challenges that would need to be overcome prior to the launch of the real interviews.

While testing was occurring, leadership in each of the target NATO HQ was approached and asked permission for JALLC to invite all staff in their entity to participate in the real interviews. Although it took the project team around one week, this approval process was seen as essential within the NATO organizational context to ensure engagement during the collection. While for some HQs this approach proved to be successful, for other HQs it ended up being more time consuming than anticipated, and in some cases resulted in substantial delays to distributing the interview link. Ultimately, the project team estimates that the interview link was made available to approximately 2000 NATO staff for a five-week period from 28 June – 02 August 2021.

In order to connect the demographic data collected through the questionnaire with the interview, DUCHESS was set up to pass a unique user identity number to the survey software. While this worked well during testing, it temporarily stopped working during the real collection. Early results indicated that while the interviewees for the test interview had all manually followed the link from DUCHESS to the survey, the interviewees in the real interview were often not clicking the link to fill the post-interview survey. As such, a few weeks into the real collection period, the DUCHESS system was reconfigured to automatically forward interviewees to the survey as soon as they finished their interview.

2.2.4 Step 4: Results & Analytics

During the testing period the DUCHESS experts developed a PowerBI dashboard to show the analytics results of the interviews. JALLC analysts worked closely with DUCHESS experts to tailor these dashboards to meet the project needs.

Initially, JALLC analysts asked for transcripts to be delivered at the end of the interview period (5 weeks). However, JALLC analysts realised that having visibility of the full transcripts as they arrived was helpful to refine the requirements for the dashboards and to be able to conduct a more in-depth manual content analysis of the data. Therefore, the JALLC analysts eventually agreed with the DUCHESS experts that the DUCHESS experts would provide transcripts to the JALLC analysts every time five interviews were completed.

Once the period for interviews had closed, the DUCHESS team developed an analysis report which was tailored to JALLC's needs in terms of a more in-depth analysis of the data collected. The report summarized the data collected and highlighted interesting results.

2.2.5 Step 5: JALLC Analysis & Reporting

The last steps were to translate the summary results into a written product that would meet the needs of the customer, and to reflect on the experience of using the new tool and write this paper.

At this point, the JALLC combined/compared the data acquired through DUCHESS with its own background research on the features of different collaboration tools, the nature of the military tasks, and relevant NATO LL Portal data regarding the use of collaboration tools over the COVID-19 pandemic (i.e. knowledge acquired through passive collection).

3.0 RESULTS

The following sections discuss various aspects of the JALLC experience with using the AI voice-bot

DUCHESS and conclude with the JALLC's takeways and what to sustain or improve with regards to each of the following aspects: number and diversity of interviewees; practical implementation; quality of responses; and level of engagement.

3.1 Number and diversity of interviewees

During the five weeks that the interview was available online, a total number of 73 NATO staff accessed and responded completely or partially to the interview questions as follows: 11 interviewees responded only partially; 47 interviewees completed all of the five content-based questions; 15 interviewees completed all five content-based questions as well as the sixth question which was generic and offered the interviewees the opportunity for a last say on the topic. The distribution of the responses across the interview period is presented in Figure 3-1 below.

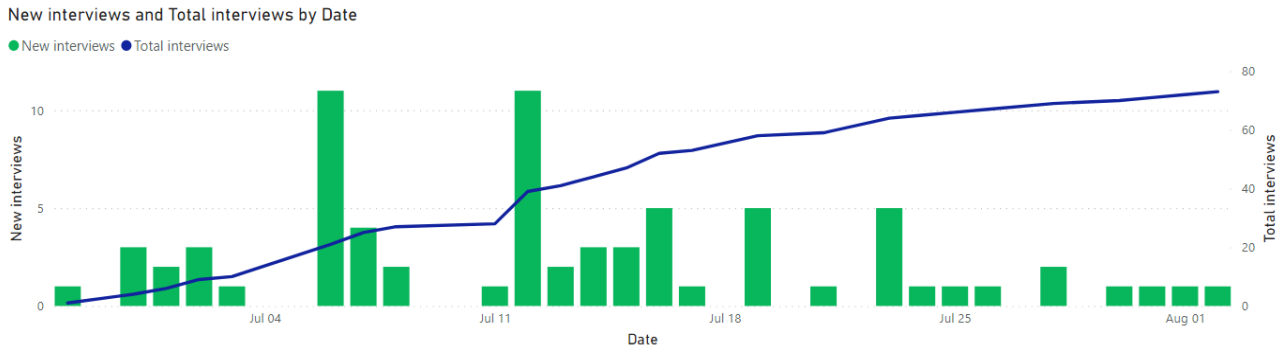

Figure 3-1: Distribution of responses across the interview period

Compared to the number of over 2000 respondents invited to take the interview, 73 represents a 3.65% return rate, which is lower than expected and lower than the 8% return rate achieved during testing. The project team identified two main reasons for this small return rate: the period for interview availability and technical limitations.

- The period for interview availability: the interview was available online during 28 June – 02 August 2021 period, which corresponds with the summer leave period in the targeted HQs. However, the risk associated with this period was assumed by the project team for the following reasons:
  - Before this period, the targeted HQs were involved in some major activities (e.g. exercises; operational assessment cycle, efforts related to operational areas). The project team assessed that there was a high risk with sending out the interview during the major activities because it would be ignored by the staff;
  - By sending the interview out after the major activities, there was potential to capture new challenges and best practices experienced by the staff during these major activities;
  - In order to allow time for transcripts review and content analysis, it was not possible to extend the period for online availability of the interview, and still meet the expected timeline for deliverables.
- Technical limitations: in many cases, the link accessibility and/or microphone availability were limited in duty networks, as further explained in the "practical implementation" section below.

However, when compared with other JALLC projects when interviews were conducted in person or written surveys were used, 73 represents a high number of returns. According to a JALLC expert[4], in a typical JALLC project around, 10 – 45 interviews will be conducted, but the project would involve more staff and would take much longer than this project. For projects where surveys are conducted, JALLC tends to target a much smaller sample than for this project and usually the return rate is around 20%, meaning at most the number of survey responses received for analysis is around 50.

Moreover, the diversity of the interviewees across the NATO Command Structure[5] suggests that even with a small number of returns, the main aspects specific to each HQ in NCS that are in the scope of this project, were likely captured. Figure 3-2 below presents the distribution of interviewees across NCS.

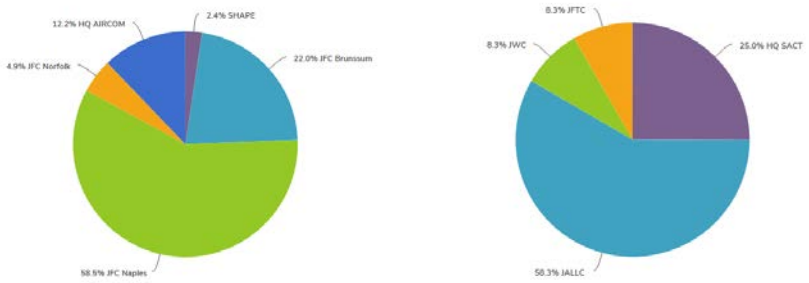

Figure 3-2: Interviewees distribution across NCS

As a note, JFCNP from ACO and the JALLC from ACT had the most staff complete interviews. The project team associates the high return rate from these HQs with the level of leadership support for staff participation in these HQs as with an active approach to engaging staff in the activity. For example, the JFCNP Information Manager personally emailed all staff the interview link emphasizing JFCNP COS support for their participation and asking staff to report to him when they had completed the interview.

Given the geographical distribution of the NCS HQs, it would have been difficult and costly for one single project team to travel to all these locations and conduct the interviews in-person in such a short period, especially while facing restrictive travel conditions due to the COVID-19 pandemic. Similarly, online (virtual) interviews as an alternative were unrealistic for at least two reasons: firstly, the project team did not know who was willing to respond in order to engage and schedule the interview and secondly, because it would have been challenging to accommodate everyone's agenda, especially during summer time.

The main takeaway relating to the number and diversity of the interviewees is that using an AI voice-bot for conducting the interview online allowed JALLC analysts to quickly reach a larger and more diverse audience than would have been possible by in-person or online interviews under the same time and budget constraints. From this perspective, the following aspects can be sustained or improved:

Sustain: request that HQ leadership support the interview request and explicitly promote participation within their HQ. Aside from the politeness of acknowledging the time that is required from their staff to participate

[4] The information was provided by JALLC Principal Operational Analyst during an interview conducted in the scope of this project.

[5] Although representatives from a wide range of NATO entities took the interview, as presented in methodology, the reference is made here only to NCS in order to reflect the perspective of military staff and the challenges and best practices with regards to collaboration tools in military environment.

in the interviews, it guides HQ leadership to stimulate their staff's participation. Additionally, ask HQs to provide updates on the levels of participation of their staff;

Improve: consider a more extended period for interview availability online with additional periodic reminders to increase the return rate.

3.2 Practical implementation

Similarly to in-person interviews, the practical implementation of this activity required planning, testing, execution, and analysis efforts, in addition to specific efforts relating to procurement and DUCHESS related coordination and deliverables. However, with respect to the subsequent analysis of the data collected, a key difference was that DUCHESS automatically analysed the data and aggregated the results, whereas JALLC analysts would normally conduct such analysis manually. Two aspects of practical implementation were worth further examination: level of effort and technical dimension.

3.2.1 Level of effort

For this project using DUCHESS, the level of effort and coordination was minimal for the project team. However, given that this was the JALLC's first experience using an AI voice-bot, the project team invested an additional analysis effort to validate individual transcripts and the findings from the automated analysis.

The table[6] below presents a comparison between the average level of effort required for a traditional semi-structured interviews versus AI voice-bot interviews.

Table 3-1: Comparison of level of effort

Interviewer task	Level of effort	
	Traditional interviews	**DUCHESS interviews**
Define open questions	3 hours	3 hours
Test the questions and update	3 hours	30 min
Plan and administrate (audience permission, communications; venue; approvals; booking; travel)	0.5 hour (30 min +) / interview	30 min
Conduct interview	0.5 hour (30 min +) / interview	0
Transcribe interview	2 hours + / interview	0
Analysis of each interview	0.25 hour (15 min +) / interview	0
Analysis of aggregated interview data	1 hour + / interview	0

Calculating the total level of effort for 73 interviewees within the context of this project, the results indicate that in the case of an in-person interviews at least 240 hours (3+3+73x0.5+73x0.5+73x2+73x0.25) would be needed, which for eight working hours / day represent about 30 working days. Additionally, in case of face-to-face interviews, travel resources (time, costs) with associated administrative efforts should be added.

Working with DUCHESS, the project team spent about 4 hours to prepare the interview for launching. As part of this effort, large scale testing proved very helpful for the project team to be able to identify and

[6] The data presented is sourced in DIEM presentation during NATO Lessons Learned Conference 2021, experience within this project, and a historical review of several JALLC analysis project where interviews were conducted for data collection.

correct some aspects of interview set-up, better define the seedling questions, and understand some of the technical challenges to be addressed prior to official launching of the interview.

With regards to the follow-on questions that the interviewer should pre-plan or formulate directly during the interview, with DUCHESS the project team did not expend any effort since these questions were developed completely by the AI tool based on previous responses provided.

Finally, another element of the practical implementation is the fact that the interviewees were not bound to a specific time (within the online availability) to respond, or a limit on how much time they had to respond, so that they can access the link as suitable for them and take as long as they wanted to complete the interview.

3.2.2 Technical dimension

Being work related, the interviews were mainly conducted during working hours and accessed via NATO networks. After identifying some technical challenges during test interviews, the project team engaged with NCI Agency and local IT staff in various HQs to include the interview site in whitelists. Unfortunately, despite these efforts, in most of the cases, the interview link was not always accessible from NATO networks. There were also situations in which the link was accessible but the staff did not have permissions to use microphones at their workstation, so they could not complete the interview. A good example of this is in the feedback provided by one interviewee who said that "I think I lost my microphone connection halfway through, so I had to reject the last couple of questions...sorry!"

Finally, some staff reported that they did not feel comfortable to speak to a computer but would have contributed more if it had been possible to type the responses. Although the figure below shows that a major part of the respondents belong to a certain age interval, after the analysis of the length of interview (in words) distributed by age, the project team was not able to identify a direct correlation between the respondent age and the lengths of the response or the preference of typing the responses.

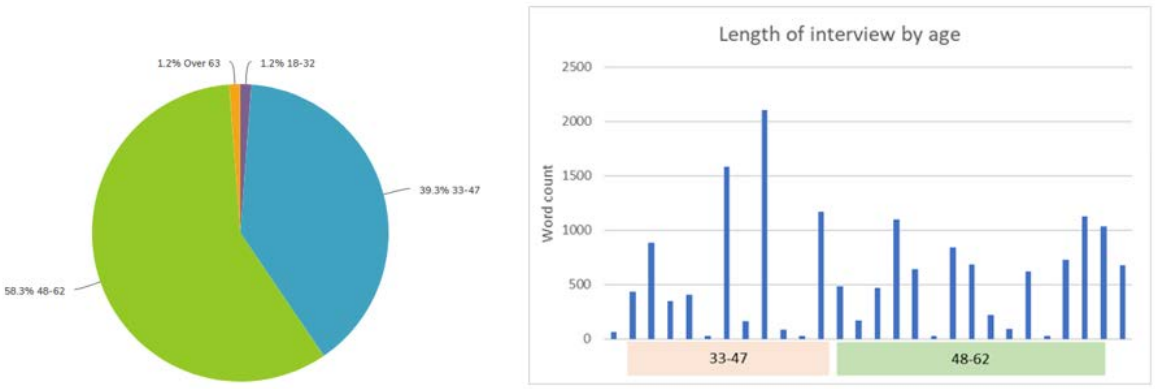

Figure 3-3: Interviewees age distribution and length of interview by age group

The main takeaways relating to the practical implementation are: the level of effort is significantly reduced for an AI voice-bot interview compared to a traditional interview; the technical set-up can be a critical barrier for responses. From these perspectives, when also considering some feedback from interviewees, the following aspects can be sustained or improved:

Sustain: conduct a pre-deployment test with a large community across different HQs to identify content and technical shortfalls. Additionally, engage with NCI Agency and local IT staff to whitelist the interview link if necessary.

Improve:

- The project team to identify solutions and alternatives to overcome technical barriers in conducting the interview at early stages of the preparatory work.
- Use professional contact networks to distribute the interview link in parallel to official channels since the experience shows that using professional contact networks elicits a more personal response and motivates staff to participate in the interview.
- Better clarify to the interviewees aspects relating to data protection and privacy of both the interviewee and interview content (level of security, data storage, duration, etc).

3.3 Quality of responses

Many interview responses contained very good insights on the topic of collaboration tools in NATO. They showed understanding of the subject and personal experience in dealing with collaboration tools. In contrast, some responses were very short and of low quality aiming potentially only to run through the interview to complete a task or to try the tool rather than contribute to the data collection. It is the project team's understanding that the anonymous nature of the online interview allowed for such a situation—which is commonly encountered in surveys but is rarely encountered during the face-to-face interviews—to occur. However, in this case, in contrast to a survey, where it may be difficult to detect poor quality responses, having access to the full transcripts and word counts for the answers made it easy to detect and remove poor quality responses from the DUCHESS results.

Beside the quality of responses, the quality of transcripts is also critical for a quality analysis. Following a manual review of transcripts, the project team perceived that the quality of transcripts was sufficient to conduct the analysis. However, it was evident from this review that the transcription technology faced limitations triggered either by physical environment of the interviewee (e.g. wearing a mask; background noise; distance to microphone) or by the English language profile of the respondent. With regards to the accent, four profiles were considered by the project team as presented in Figure 3-4 below.

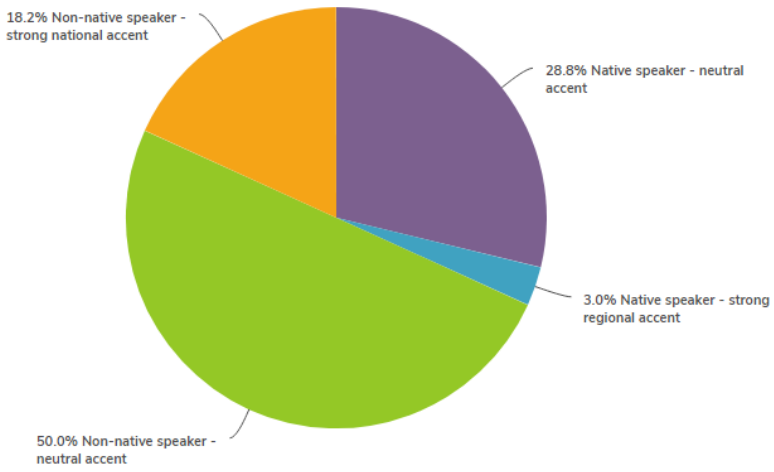

Figure 3-4: Interviewees English language profile

The project team observed that while for native speakers the transcription achieved a very high level of fidelity, for different accents the quality of the transcripts suffered to the point of distorting the meaning of the response (see Table 3-2). However, it was observed that, whenever long responses were provided, the

accuracy of the transcription increased and the follow-on questions were more appropriate for the content provided. Below are some examples of the wrong transcripts that the project team had to re-validate within the context or with re-engagement with the interviewees who provided contact details.

Table 3-2: Examples of wrong transcriptions

Intent	Transcript
..them all…	The mall
..phones and that was….	..phones. Anne. That…
..Zoom…	..Zune…
…his colleagues…	…it's colleges…
?????	...bikes account define talk made, some meetings with disease.
…everyone adhere to this culture.	…everyone are there to this culture.
….to discuss same kind of issues….	…to discuss some kind of. Shoes.
… were both for call in ….	…were about for Colin…

These types of errors in the transcripts had a direct impact on the construct of the follow-on questions and the level of engagement (more details in the next section), as well as the subsequent analysis results. In this context, the project team assessment is that the quality of transcription should be higher to be able to rely on the automated analysis results. This would be especially important in a project involving a very large number of responses such that reading and verifying each individual response manually would not be feasible.

The main takeaways relating to the quality of responses is that the quality of transcription was directly influenced by the English language profile of the interviewee and the length of the response, directly impacting the quality of the follow-on questions and automated analysis results. From this perspective, considering also some feedback received from interviewees, the following can be sustained or improved:

Sustain: include an open question that allows the more talkative interviewees to give longer responses.

Improve: depending on the AI voice-bot used, conduct some test interviews with long responses that include a wide range of topic-related vocabulary to train the AI to formulate more relevant follow-on questions;

3.4 Level of Engagement

Successful data collection via a semi-structured interview depends largely on the level of engagement of the interviewee and the way the interviewer formulates the follow-on questions.

To measure the value of using DUCHESS with its auto follow-on questions compared to using only a set of pre-planned questions, the project team analysed the level of engagement and the length of follow-on question responses. In this analysis, the project team removed the responses with less than 10 words usually given when someone does not want to answer, does not have anything more to say, or, as indicated in the interview feedback, does not feel comfortable to speak to a machine and would prefer an improved interaction interface or to type the responses.

As already mentioned in the takeaways of previous section, individual differences play a significant role in the level of engagement since the amount that each individual says while responding to a seedling question influences the quality of follow-on questions. In this respect, if a participant has given a long response to the seedling question they are less likely to give a long response to follow-ons. This is illustrated in Figure 3-5 below extracted from DUCHESS technical report [17] where the response length for the more talkative participant drops for the follow-on questions, but a less talkative participant is encouraged to increase the length of their answers by the asking of follow-on questions.

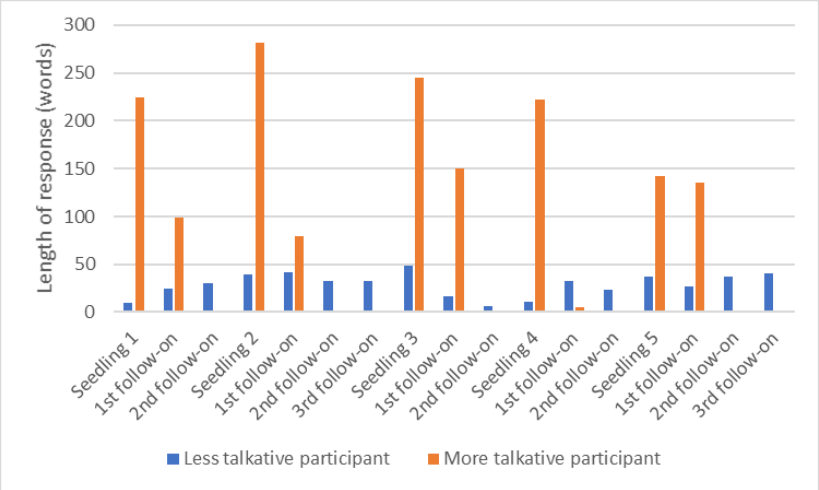

Figure 3-5: The human factor

It was also observed by the project team that the level of engagement with follow-on questions decreases as the interview goes on. Figure 3-6 below presents the responses to follow-on questions. The trend identified in the chart is expected since, in the project team's understanding, participants will have started to cover everything they want to say, given that the seedling questions are so open yet on the same topic.

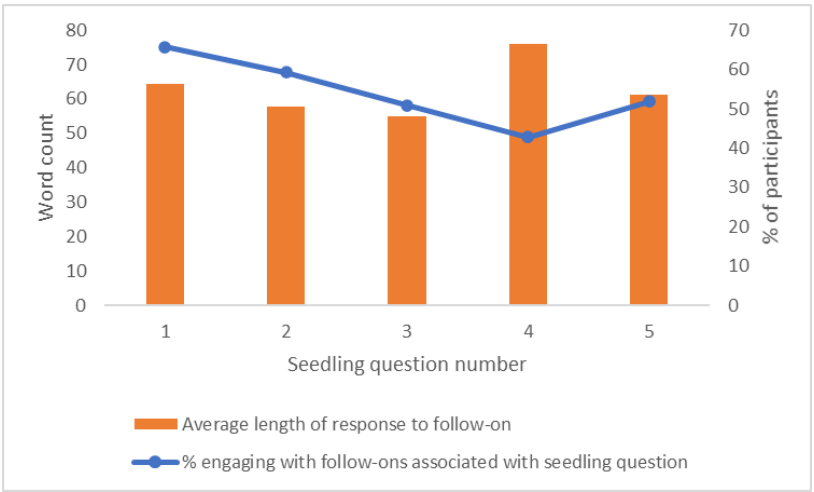

Figure 3-6: Responses to follow-ons

When interviewees did engage with the follow-on questions, the average length of responses to the follow-on questions ranged from 55 to 76 words with the overall average being 64 words. This indicates that the follow-on questions elicited a substantial amount of additional information. The following response to a

follow-on question was 63 words long and shows the insight that would not have been captured by a fixed question set.

> *"Yes, regarding interoperability. Once we understand the security implications on the enterprise. The lack of interoperability between things like Polycom and understand why we couldn't use zoom, but the lack of interoperability between Polycom and Microsoft Teams etc meant that we were very constrained in the VTC capability and we saw a rapid increase in the requirement to operate on different platforms to different users."*

The project team also observed that engagement with follow-on questions increased for the 2nd and 3rd follow-on questions to the same seedling question (see Figure 3-6 below).

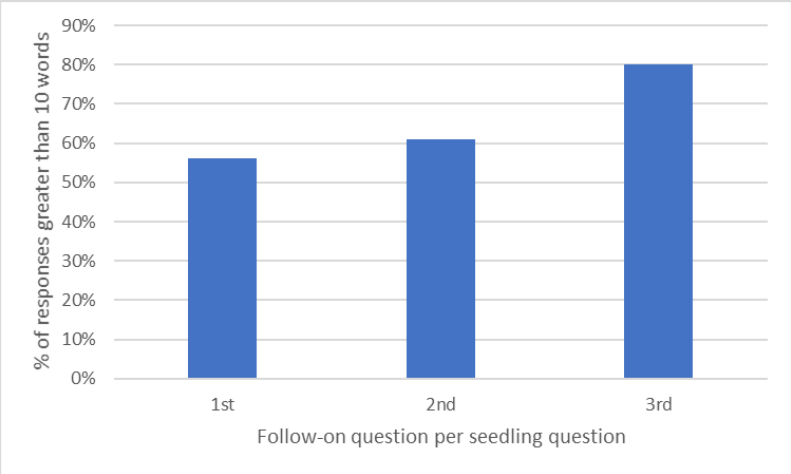

Figure 3-6: Progress through follow-ons

According to DUCHESS experts (Technical report) [17], this is to be expected as the DUCHESS algorithms improve as the participant provides more data. As such, if they have given a response to the seedling question, 1st follow-on and 2nd follow-on questions, the 3rd follow-on question is likely to be more relevant than 1st and 2nd and they are then more likely to engage.

Finally, with regards to the content of the interview, it is not clear for the project team how an AI voice-bot can ask follow-on questions in order to bring the interviewee back to the topic in case their responses are off-subject. If the answer to the seedling question is off-topic, is seems likely that an AI voice-bot could easily proceed to ask even more in-depth questions on subjects which are not in the scope of the interview.

The main takeaway relating to the interviewees' level of engagement is that the quality of responses to the seedling questions directly impacts the quality of the follow-on questions' formulation which subsequently impacts the interviewee's willingness to share more on the topic and continue with the interview. From this perspective, the following aspects can be sustained or improved:

Sustain: asking a maximum of three follow-on questions appears to be enough to capture all insights from the interviewee with regards to the respective topic.

Improve:

- configure the interview site interface to allow written responses as well as to allow the respondent to pause during recording a response;
- provide the transcript or the recording for the respondent to review and validate that the AI properly

understood the response or add some kind of visual cue (e.g. a progress bar) that can give the interviewees confidence that their response has been heard and registered in the system;

- project an avatar to give the interviewee the impression of having a conversation with a person;
- start the interview with some more generic interaction to create a better atmosphere and introduce the topic before getting to the first question;
- consider a way to formulate follow-on questions based on pre-defined topics to ensure the interview stays in the intended direction and the interviewee is not encouraged to go off-topic.

4.0 CONCLUSION

At this stage of the project, the answer to the question "Does AI conduct better research interview than you?" is: **it depends.** This answer is driven by several pros, cons, and considerations for NATO to take into account for future use of this new technology.

4.1 Pros

Based on the results of this project, the project team has identified that there are aspects where AI-based interviews are definitely better than traditional interviews conducted face-to-face, either in-person or online. The positive aspects of AI-based interviews include:

- Extremely well suited for structured interviews where the questions are pre-defined and validated by the project team during tests prior to launching the interview;
- Engages a very large sample which improves completeness of data collection and increases diversity of experiences and perspectives collected. Subsequently, this allows a more substantial analysis and supports decision-making based on more and stronger evidence;
- Allows the audience to choose when to take the interview, as long as they have a minimum set-up for the devices (e.g. laptops) to use and access to the interview link;
- Empowers people to speak out given the benefit of anonymity, ensuring no response attributionto a person, organization, or nation, as such eliminating self-censure and caution, that are common in a more formal approach to LL knowledge acquisition interviews;
- Facilitates auto-transcription, saving massive amounts of time in the interview process, and recording the original words of the interviewee, rather than summaries or notes which can be subject to errors and omissions;
- Performs text analytics such as sentiment analysis, part of speech, etc., assisting researchers or analysts to identify themes or trends in the responses, and hone in on particularly emotive topics.

Additionally, although beyond the scope of this experience using an AI voice-bot, other new technologies can be used to assist in any interview process such as:

- Self-booking systems which can transfer the effort associated with scheduling interviews to the interviewee, thus freeing up some of the interviewer's time;
- Auto-translation capability which would allow the interviewee to conduct the interview in their native language, and for that transcript to be translated automatically into a common language / English.

4.2 Cons

The project team also recognizes that AI voice-bots are at their early stages of development and there are elements that require improvement. According to the results of this project, they refer mainly to:

- While for native English speakers the quality of transcripts is very good, for non-native or those speaking with an accent, the transcription does not appear to be at a suitable level to guarantee full confidence in the subsequent analysis. This may introduce bias within the subsequent analysis because responses from those with clearest accents are better represented and weighted in the analysis;
- In the case of semi-structured interviews, similar to the interview conducted within this project, the follow-on questions, although based on wording in the response for previous question, do not always appear clearly within a logical flow of the interview. Moreover, there is a significant risk that if interviewees go off-topic in their initial response, the follow-on question would lead them even more off-topic instead of directing them back in the scope of the interview.

4.3 Considerations

Finally, for quality interviews, besides the interviewer (in this case the AI voice-bot) the audience engagement is critical. From this perspective, in the project team understanding, the following aspects are future enablers and need to improve in NATO to make it possible for interviews conducted by an AI occur in a suitable environment:

- The technical dimension is critical to ensure the audience can access the interview link and their devices allow for listening to and recording responses to the questions. While this is not at all a concern if the means for the interview is a personal device (e.g. smartphone), it was a critical aspect to consider in NATO's office environment;
- For interviews on classified topics, a suitable way to enable deployment of the AI voice-bot in a classified network/environment should be identified. The interview within this project was conducted completely on a commercial platform over the Internet;
- The audience should be familiarized with verbal interaction with computers so that the level of confidence in the way the interview is conducted is high and allows for a natural approach to the interview.

5.0 REFERENCES

[1] Dyson, T., "Organisational Learning and the Modern Army (Cass Military Studies)" (1st ed.). Routledge, 2020

[2] Nevis, E., DiBella, A. & Gould, J., "Understanding organizations as learning systems", Sloan Management Review, 36(2), 73-85, 1995

[3] Zahra, S. A., & George, G., "Absorptive Capacity: A Review, Reconceptualization, and Extension", The Academy of Management Review, 27(2), 185–203, 2002, https://doi.org/10.2307/4134351

[4] Wikipedia contributors, " Absorptive capacity", Wikipedia, July 2021, Retrieved September 15, 2021, https://en.wikipedia.org/wiki/Absorptive_capacity

[5] JALLC, "Report on New Technologies Event 2018", 2018, Retrieved September 16, 2021, https://nllp.jallc.nato.int/IKS/Sharing%20Public/NTE18%20Final%20Report%2018122018.pdf

[6] JALLC, "New Technologies in support of Lessons Learned report", JALLC/CG/19/011, 31 January 2019

[7] Weber, R., Aha, D., & Becerra-Fernandez, I., "Intelligent lessons learned systems. Expert Systems with Applications", 20(1), 17–34, 2001, https://doi.org/10.1016/s0957-4174(00)00046-4

[8] JALLC, "NATO Lessons Learned Handbook", 3rd Edition, 2016, Retrieved September 15, 2021, https://www.jallc.nato.int/organization/products/books-guides

[9] WHO, "After Action Review Key Informant Interview AAR Facilitator's Manual', 2019, Retrieved September 15, 2021, https://www.who.int/publications/m/item/aar-toolkits

[10] Milton, N., "Learning History - guidance from Knoco Ltd.", Knoco, Retrieved September 15, 2021, from https://www.knoco.com/learning-history.htm

[11] Alshenqeeti, H., "Interviewing as a Data Collection Method: A Critical Review", English Linguistics Research, 3(1), 2014, https://doi.org/10.5430/elr.v3n1p39

[12] NCI Agency, "On-demand interviewing boosting recruitment", October 2016, Retrieved September 15, 2021, from https://www.ncia.nato.int/about-us/newsroom/ondemand-interviewing-boosting-recruitment.html

[13] Pattapu, C., "How AI Interviewing is Redefining the Way we Hire", Talview, June 2019, Retrieved September 15, 2021, https://blog.talview.com/how-ai-interviewing-redefining-the-way-we-hire

[14] Google Developers, "Chatbot Concepts Google Chat API", September 2021, Retrieved September 15, 2021, from https://developers.google.com/chat/concepts/bots

[15] Taruno, M. E., "The Complete Guide to Building a Chatbot with Deep Learning From Scratch. Medium", February 2021, Retrieved September 15, 2021, from https://towardsdatascience.com/complete-guide-to-building-a-chatbot-with-spacy-and-deep-learning-d18811465876

[16] Vincent-Major, S., "DIEM Innovation presentation on DUCHESS during the NATO Lessons Learned Conference 2021", March 2021

[17] DIEM Innovation, "JALLC Collaborative Working interviews: Analysis Results", Technical report, DIEM_LL_CW01 v1_0, September 2021

6.0 AKNOWLEDGEMENTS

The project team acknowledges the excellent support provided by:

- NATO HQ C3 Staff for their open-mindedness and willingness to allow JALLC to trial a new system for their analysis project;
- JALLC leadership for their support in conducting the project in this manner;
- NATO colleagues who took the interview and provided feedback for improving it;
- DUCHESS experts from DIEM Innovation for their professional support of this project;
- JFCNP Information Manager - Mr Antonio De Frutos - for his personal commitment to get more staff taking the interview;
- NEDP alumni for their support with the testing phase of the interview.