



Capstone Project Proposal

Mohamed Fazal Mustafa

Business Goals

Project Overview and Goal

What is the industry problem you are trying to solve? Why use ML/AI in solving this task? Be as specific as you can when describing how ML/AI can provide value. For example, if you're labeling images, how will this help the business?

Round-trip communication from the Earth to Mars requires approximately 20 minutes. During this time, astronauts are effectively "on their own" regarding the need for support. Astronauts on the International Space Station (ISS), while closer to home, can have communication blackouts that lead to variable delays in support. As such, having an artificial agent as a virtual assistant could greatly benefit astronauts. This project combines human factors and human-robot interaction via an artificial agent. The purpose is to save the astronauts precious time, cognitive resources, and to mitigate detrimental interpersonal communications caused by additional stressors that result in a lack of resources. The goal is to create a mixed human-robot team in which a chatbot is responsible for its own tasks allowing astronauts to focus on more important mission events.

Studies have shown that regular environmental stress leads to decreases in mood, working memory, situation awareness, and performance. Split-attention is a result of "multitasking" which directly

	<p>impacts a person's communication with other people (image looking for a key item during an experiment while being asked a question from a colleague). Decreases in mood impact human factors more drastically in isolation, causing astronauts to become irritable more frequently and more easily in events that would not usually irritate them.</p> <p>The concept is to create a chatbot that is capable of reducing the working memory and cognitive load of the astronauts via voice assistance. The chatbot would be able to aid astronauts by providing locations to useful items, reminding astronauts regarding steps in a protocol (for example, another mission protocol will have a specific sequence and the chatbot could remind the astronaut of certain steps). Furthermore, the chatbot could provide reminders regarding workflows and time (analog astronauts might need to be reminded to take a break, conduct debriefing, or other daily tasks). Finally, the chatbot could act as a virtual, temporary counselor, allowing astronauts to destress utilizing the chatbot as a digital companion.</p> <p>The results of the communication could allow for improve AI chatbot models based on labeling and classifying various usages of the chatbot (i.e. mood enhancement or protocol briefing).</p>
<p>Business Case</p> <p>Why is this an important problem to solve? Make a case for building this product in terms of its impact on recurring revenue, market share, customer happiness and/or other drivers of business success.</p>	<p>While many technological business cases emphasis the generation of revenue, aerospace applications generally attempt to reduce <i>risk</i>. Astronauts find themselves in dangerous environments and burdened with the reality that the accumulation of small errors can lead to catastrophic mission failure or death. Since human performance is a factor of cognitive load, having an artificial agent as an assistant could allow astronauts to perform their tasks with higher accuracy and less cognitive load resulting from split attention tasks, unnecessary stressors related to object searching, overloads to short-term memory utilization, and</p>

	<p>emotional pressure caused by the isolation of space.</p> <p>Improving human cognition via mood/memory/task-offloading enhancements can reduce risks to injury, deadlines, or even performance. In mission-critical environments, human errors can result in the loss of future funding or even lawsuits due to injury and death. The goal is to decrease the risks of the astronauts directly which will impact the risks of the business. Furthermore, agencies adopting artificial agents might be able to position themselves to increase their funding by demonstrating the results of effective artificial agents including chatbots. Consider the risks associated with an injured astronaut that needs to communicate with Earth from Mars and there is a 20min round-trip communication delay. The risks to the individual and the company (agency) are quite large and need to be mitigated in an intelligent manner.</p>
<p>Application of ML/AI</p> <p>What precise task will you use ML/AI to accomplish? What business outcome or objective will you achieve?</p>	<p>Chatbots are inherently a topic of artificial intelligence, as the essence of the technology is surrounding an artificial agent. The domain is within Natural Language Processing (NLP) which has become an emergent topic recently. In classical AI, NLP problems for chatbots were solved with decision trees and hidden-Markov models (HMMs). In recent times, convolutional neural networks (CNNs) and recurrent neural networks (RNNs) can be combined to create effective, fast, and knowledge agents given enough data is present.</p> <p>The objective is to utilize the advances in chatbot technologies combined with voice recognition software and NLP databases to create an effective chatbot. The outcome should reflect an agent that is suitable for assuming some of the astronaut's typical tasks while providing information and support that might require a large amount of time if contact with the Earth were required.</p>

Success Metrics

Success Metrics

What business metrics will you apply to determine the success of your product? Good metrics are clearly defined and easily measurable. Specify how you will establish a baseline value to provide a point of comparison.

Presumably, an effective artificial agent is one that is able to complete its tasks within acceptance criteria. Over time, the behavior of the system can be translated to success depending on system usage, adoption (by other companies/agencies), and direct mitigation of risks.

Therefore, the initial business metrics of interest include:

- A reduction of human errors of 10% within six months
- Chatbot usage increases by at least 20% within one month for support tasks (instead of a human actor)
- The number of injuries decreases by 50% within six months
- Astronaut performance as measured by the NASA-TLX, SAGAT, and NASA-PVT tests increase by at least 10% within two months

The basis for these metrics is the historical data and trends that the agency currently has.

Data

Data Acquisition

Where will you source your data from? What is the cost to acquire these data? Are there any personally identifying information (PII) or data sensitivity issues you will need to overcome? Will data become available on an ongoing basis, or will you acquire a large batch of data that will need to be refreshed?

Initially for building this model data will be sourced from 3 sources. They are listed below:-

1. Speech Accent Archive: the dataset contains 2,140 English speech samples, each from a different speaker reading the same passage. Data Type – Audio, Size – 906 mb.

2. Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): RAVDESS contains 24 professional actors (12 female and 12 male), vocalizing the same statements. Not only this, but the speech emotions captured include calm, happy, sad, angry, fearful, surprise, and disgust at two levels of intensity. Data Type – Audio, Size – 563 mb

3. Google Audioset: This dataset contains expanding ontology of 635 audio event classes and a collection of over 2 million 10-second sound clips from YouTube videos. Moreover, Google used human labelers to add metadata, context and content analysis. We will only use Human Speech class from this dataset which is approx. Data Type – Audio, 1.2 GB in size.

We will be highly representing audio mp3 data in our dataset.

All these datasets have been made open-source and free to use and train models on. So data acquisition cost will not be incurred.

Also, Data will be acquired via direct verbal communication with the chatbot. The content of the audial data will be converted to text (Speech-to-Text) and stored in a log file (or ROSBag). The initial source of the language model/corpus will come from product integration with the Google Cloud Platform (GCP). In the event that GCP is not sufficient (i.e. due to architecture or interface issues), Microsoft Cognitive Service via Azure will be utilized for the language library. The initial data from the libraries (for speech recognition) is already quite large but the libraries will be updated on a continuous integration and continuous delivery schedule. The results

	<p>from the model/system will also be used in future architectures and the data from each successful mission will be used as a source for future missions.</p> <p>PII is an intrinsic issue in privacy as personal information is being stored (spoken content and private data). To overcome this, no personal information regarding the astronaut will be supplied to or shared with the log files. The content will be filtered and metadata will not contain any superfluous information. The content (type, frequency) of sentences spoken to the chatbot will be documented but the astronaut that provided the content will not.</p> <p>Timestamps will not be used and the data (sentences, words) will be randomly sorted prior to data analysis to prevent a case in which the metadata could be extracted based on time (for example, knowing which astronaut was conducting an EVA at a particular time will expose their privacy and speech content during that period).</p>
<p>Data Source</p> <p>Consider the size and source of</p>	<p>In terms of utilizing the GPC or Azure, the maximum amount of performance for the speech recognition model is already current. Proper responses from the chatbot are where issues can arise due to misinterpretations and cases</p>

your data; what biases are built into the data and how might the data be improved?

of synonyms/homonyms. In terms of objective questions like [In what container is the 9mm wrench?] or [What are the mission protocols and sequence for the “drone” project?], this is less of an issue as there is a standard communication model and common language already established to prevent miscommunication and ambiguity. However, for conversations in which the chatbot is a mood enhancement companion, subjective questions and statements (especially containing colloquialisms) is problematic. Therefore, the “general” behavior of astronauts, including their discussion topics, will need to be documented and used in future models. However, by anonymizing the data, biases can result based on various astronaut personalities. Astronaut A might converse more freely about topics regarding emotional states than Astronaut B. Therefore, the emotion-based topics of Astronaut A will be distributed to the same agent that Astronaut B uses which results in a shifted “weight” due to cumulative usage (effectively the behavior of Astronaut A becomes applied to all other astronauts). Unfortunately, the only “easy” solution to this problem is to record and document the personal information of each individual astronaut. At this time, this topic will not be explored in more detail until the ethics board is consulted or advancements in affective computing allow for adaptive agents at real time (meaning the agent can respond naturally without prior information of the user based on the current state and “mood”).

An additional source of bias in the speech recognizer is related to pronunciation. Astronauts have varying pronunciations, word usage, and accents, which impact the accuracy of the speech recognizer. Ambiguity is a component of natural language and is hard to remove for subjective conversations. Fortunately, word accuracy related to accents and dialect-influenced pronunciations can be improved over time (more data from unique users).

Choice of Data Labels

What labels did you decide to add to your data? And why did you decide on these labels versus any other option?

Initially, the speech recognition analysis is concerned with two types of speech defined as “objective” and “subjective.”

Objective speech is defined as questions or statements that require a clear, factual response (like when asking the current time).

Subjective speech is defined as communication that results in responses that are “open” or “free,” meaning there is not a singular, optimal response (and in fact, there might not exist an optimal response or not providing a response is the best solution).

Objective speech has categories including objects and events. Objects are physical objects within the physical domain (tools). Events are instances as a function of time. The speech results will be analyzed regarding the frequency of objects and events, the relative accuracy of the chatbot understanding, and the effectiveness of the chatbot’s response (in line with the astronaut’s inquiry).

Subjective speech will currently monitor emotions and moods (happy, angry, etc.) and events. The chatbot will attempt to ask questions to find which events led to the astronaut’s current emotional state. Furthermore, the emotional state of the astronaut post-chatbot-usage is of interest but these will not be labeled by an artificial agent (and will be extracted via testing that resembles “user studies”).

Model

Model Building

How will you resource building the model that you need? Will you outsource model training and/or hosting to an external platform, or will you build the model using an in-house team, and why?

As with most astronautic artificial intelligence projects/products, most of the model building/training will be done with an in-house team that includes a Project Manager, Product Owner, Human-Robot Interaction Specialist, Linguist, Psychologist, Human-Factors Engineer, Lead Architect, System Engineer, and several Engineers and Designers (effectively software developers).

We will use AWS for this purpose, All the data will be stored in S3 bucket and our ML model will be trained by creating a GPU instance in AWS Sagemaker. In order to test the model a simple web-app will be built for testing purpose and it will be connected with our model which will be tested via API Gateway.

All of it will be done in-house to ensure speedy delivery, security and control over the product's lifetime.

Architecting the product first-hand allows us to have complete knowledge about it which will help us to improve the product repeatedly and update it in the future.

Evaluating Results

Which model performance metrics are appropriate to measure the success of your model? What level of performance is required?

For the labeling model, the accuracy and recall (in terms of true/false positives and negatives) will provide insights into how good the model is at classification. Naturally, the F-Score will be used as a better measure for overall performance within classification. The level of performance should be greater than 90 to relieve developers of having to sort through many misclassified labels when designing future models.

For NLP applications, the accuracy and confidence of the model is usually measured. The accuracy of correctness in word matching (keyword spotting) will be measured. The accuracy and confidence need to be as close to 100 as possible (the requirement will be 98 for feasibility) for objective speech and the requirement for subjective speech is 85 . The lower requirement for subjective speech is due to fact that lower accuracy for these topics is more of an annoyance than a cause for mission failure and due to the fact that there is more ambiguity and potential bias when discussing subjective topics.

Minimum Viable Product(MVP)

Design

What does your minimum viable product look like? Include sketches of your product.

An MVP has been created and via the GCP and Dialogflow (an intelligent natural language framework). Dialogflow includes an analytics tool that can measure the engagement or session metrics like usage patterns, latency issues, etc. Below, the images show examples of possible sentences from the astronaut (training phrases) and other plugin features (like <Get joke> to ask for a joke from the GCP to improve user mood). An example of synonyms is demonstrated for emotions, in which the response of the chatbot will be the same if the astronaut were to make two similar/equivalent statements.

Events ⓘ

Google Assistant Welcome ⓘ Welcome ⓘ Get joke ⓘ Add event

Get joke
action_intent_get_joke ⓘ by Google Assistant

PARAMETER NAME	MESSAGE TYPE	DESCRIPTION
joke	org.schema.type.Joke	—

Training phrases ⓘ

” Add user expression

” Where is the protocol list?

” What is the location of the 9mm wrench?

PARAMETER NAME	ENTITY	RESOLVED VALUE
unit-length	@sys.unit-length	9mm
tools	@tools	wrench

Training phrases ⓘ

” Add user expression

” This space chatbot makes me irate.

” The experiment was annoying.

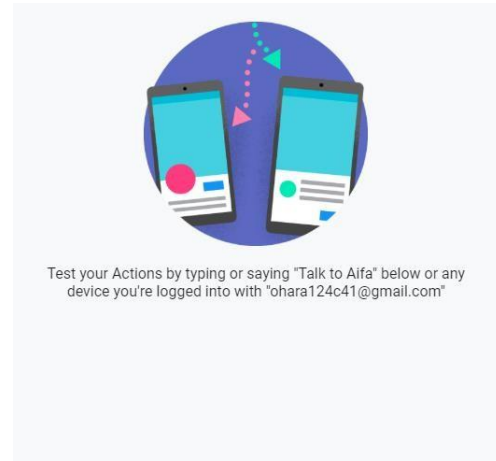
” I feel sad.

emotion

☒ Define synonyms ⓘ ☐ Regexp entity ⓘ ☒ Allow automated expansion ☒ Fuzzy matching ⓘ

sad	sad, down, merose, unhappy, dejected, regretful, depressed
happy	happy, contented, content, cheerful, cheery, merry, joyful, jovial
angry	angry, mad, irate, cross, annoyed, vexed, irritated, exasperated

The images below show examples of the demo chatbot in action with Dialogflow. The first image is the UI that the user directly interacts with (via voice or keyboard).



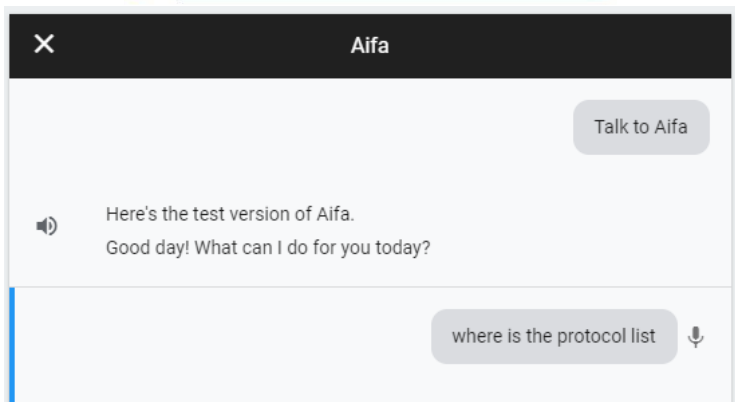
Suggested input

Talk to Aifa

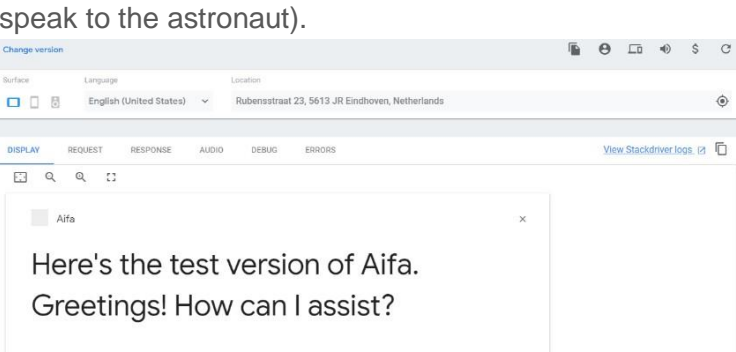
Input



Talk to Aifa



The next image shows a text-based version of the chatbot speaking (the chatbot also uses verbal communication to

	<p>Speak to the astronaut).</p> 
<p>Use Cases</p> <p>What persona are you designing for? Can you describe the major epic-level use cases your product addresses? How will users access this product?</p>	<p>Astronauts are like-minded people that put the mission first and have similar tasks throughout the mission regardless of their backgrounds (education, career, and personal life). As such, general use cases are easier to formulate in this domain (though specific use cases do not follow a persona). The typical astronaut is highly educated, results-oriented, goal-driven, and well trained physically and skillfully. Some example use case questions for the astronaut are:</p> <p>As an astronaut, how can I more easily find tools to complete tasks faster?</p> <p>As an astronaut, how can I discuss negative topics to improve my mood without upsetting others?</p> <p>Mission operatives in the Ground Control Center also have use cases that are tightly coupled with the astronauts (so it is important to briefly discuss them). Some examples include:</p> <p>As a capsule commander (CAPCOM), how can I ensure that astronauts have support during a communication blackout to improve their safety?</p> <p>As a medical officers (MEDOPS), how can ensure that injured astronauts can be attended to quickly to minimize their injuries and risk of death?</p> <p>At the agency level, the use cases are essentially:</p> <p>As a president, how can I provide confidence to investors/funders that result in future donations/contributions for missions?</p>

	<p>As the CTO, how can encourage and inspire a new generation of space enthusiasts to become astronauts and space engineers?</p> <p>For user access, the initial (beta) version will be similar to the MVP with a GUI for astronaut communication. Once the results of astronaut usage/frequency have been analyzed, a version of the chatbot will be created that operates through the communication channels directly across all users (similar to Siri or Alexa in which a “wake-word” activates the chatbot and it provides cross-functional, real-time assistance to all of the astronauts simultaneously.</p>
<p>Roll-out</p> <p>How will this be adopted? What does the go-to-market plan look like?</p>	<p>The first phase will include beta-testing within the laboratory with a small set of users to gain insights in the appropriate direction for the chatbot.</p> <p>The second phase will utilize the chatbot during analog missions (basically space simulations) with the performance and accuracy of the artificial agent being analyzed as a result of near real-life situations.</p> <p>The third phase will incorporate the chatbot being installed in the ISS permanently.</p> <p>The fourth phase will include the development of the chatbot (based on previous results) for long missions to Mars (and beyond) include any “permanent” habitats built on the moon, Mars, or other planetary bodies/infrastructures.</p> <p>After each phase, the results will be provided to the public to raise awareness. If all goes well, chatbots will be adapted by all major space agencies. Finally, the results will hopefully spread to other safety-critical domains in which a chatbot in a mixed human-robot team could decrease cognitive load and improve human capabilities (i.e. special forces, firefighters, military, etc.).</p>

Post-MVP-Deployment

<p>Designing for Longevity</p> <p>How might you improve your product in the long-term? How might real-world data be different from the training data? How will your product learn from new data? How might you employ A/B testing to improve your product?</p>	<p>To design for long-term, the model needs to be adaptive in allowing for improvement based on previous results. The results of the previous models, including voice recognition if possible, will be fed into future models. Furthermore, keyword spotting and automatic responses will be improved based on how the initial response were received by the astronauts. A variety of quantitative and qualitative tests will report the results of the chatbot on the astronaut's performance, mood, behavior, and attitudes. Performance here can be thought of as response time, task completion time, task completion accuracy, or correct prioritization. Mood/behavior/attitude factors include an overall reduction of stress and better emotional states for the astronaut as a result of using the chatbot.</p> <p>Real-world data will likely be different than any initial user testing or laboratory testing as people (including astronauts) behave differently during safety-critical missions or in extreme environmental conditions (including isolation in which the astronaut team cannot possibly be in physical contact with others). The stress load will increase greatly which will impact astronaut usage with the chatbot. As such, the roll-out plan (as previously described) has already addressed these issues by incorporating various phases to improve the chatbot. A/B testing could take two groups in similar conditions or the same group in noticeably different conditions. This could allow for changes in astronaut behavior and usage to be analyzed yielding some insights into how mission specifications impact astronaut communications with the chatbot.</p>
<p>Monitor Bias</p> <p>How do you plan to monitor or</p>	<p>To monitor bias in accent or usage (synonyms/homonyms), the accuracy (NLP, see above) could be monitored and keywords that cause issues can be isolated (regularly</p>

mitigate unwanted bias in your model?

misunderstood words). Astronauts could be asked to use different vocabulary in these cases.

In terms of designing responses to subjective, emotion-based topics, the easier solution is to assign conversation results and trends directly to the user. However, this can be an issue with privacy (IIP) initially. Generally, astronauts signed an NDA/waiver that provides the agency/scientists to full access to their personal data as long as it is not leaked to the public or third-party agencies. Therefore, it should not be too much effort to gain approval for assigning personal data to discrete instance (profile) of the chatbot. The log files will be monitored overtime to improve the results and behavior of the chatbot.