Capstone Final Report

for

Team 11: Reliability Artificial Intelligence

Team members:
David Deltz
Maanya Gulati
Fausto Sotelo
Kyle Threatt
Jasper van Lier

Submitted to fulfill the requirement for a final report in ISEN 460
Department of Industrial and Systems Engineering
Texas A&M University

04/26/2024
Spring 2024

Course Teaching Team:
Dr. Michael Do
Mr. Jose Vazquez
Dr. JP Elizondo
Mr. Arvin Shadravan
Mr. Jackson Sanders

Project Sponsor: Dr. David Mortin, DEVCOM DAC
Advisor: Dr. Ceyhun Eksin

Table of Contents

List of Figures

List of Tables

Executive Summary

The goal of this project was to expand on the research and list of failure modes for artificial intelligence and machine learning provided by the US Army DEVCOM Data and Analysis Center which was accomplished by adding federated learning and its failure modes to the project. Federated learning is a novel way of training an artificial intelligence or machine learning model using multiple agents. The idea is to train a central algorithm or model using the outputs of running this model on several different agents and data sets. The implementation and use of this style of training opens up new possibilities of security risks and failures especially with the possibility of impacting the trained model, but could allow a model to be trained on a wide variety of data in a shorter amount of time.

1.0 Information Gathering & Problem Definition

The following subsections document the primary background of the capstone project sponsor. Furthermore, it includes a literature review of previous similar projects and provided resources from the project sponsor. Additional information like the problem statement and customer needs are also listed and defined to identify the scope of the project. The documentation also includes the project charter which includes the objectives of the research projects as well as special considerations that need to be taken into account for the results of the project.

1.1 Introduction

With the growing relevance of Artificial intelligence in our day-to-day lives, it only makes sense that the US Army has taken an interest in how they can apply AI to their day-to-day operations.

However, nothing is perfect, and therefore failure can occur. Since the US Army has a very broad reach in military operations, there are a lot of areas where AI can be applied and where it can go wrong. Therefore, we were tasked with adding to an existing list of AI failure modes and their effects within the military. Our research project aims to add to the list of existing failure modes provided by DAC, the division of the army we are helping. Since the artificial intelligence field is so broad, the scope of the project has been narrowed down to focus on failure modes associated with federated learning.

While it's important for the US Army to be aware of the failure modes, it is also important to understand the effects of failures on their programs. Therefore, we are adding additional information to these failure modes on how they can affect potential military operations.

1.2 Literature Review

DEVCOM or The U.S. Army Combat Capabilities Development Command is a government group of "engineers, scientists, analysts, technicians, and support staff" which develops and researches cutting-edge technology for Soldiers on the battlefield (U.S. Army). Technologies such as "artificial intelligence, quantum effects, autonomy, robotics, advanced energetics, and synthetic biology" are being developed for use by Soldiers to create an advantage in the field (U.S. Army). "DEVCOM's mission is to continually deliver capabilities by proactively exploring basic and foundational research." The research DEVCOM is interested in particular includes subjects such as, "RF Electronic Materials; Quantum; Hypersonic Flight; Artificial Intelligence; Autonomy; Synthetic Biology; Material by Design; and Science of Additive Manufacturing" (U.S. Army). Their research is not limited to the list but one particular research area which is relevant to the project is artificial intelligence.

The project sponsors are Dr. David Mortin, Martin Wayne, and Nathan Herbert. Dr. Mortin serves as the Chief of the Reliability Branch at the U.S. Army Materiel Systems Analysis Activity, Mr. Wayne leads the Center for Reliability Growth within DEVCOM, and Mr. Herbert is a technical advisor in the Center for Reliability Growth within DEVCOM. All three are involved in the Institute for Electrical and Electronics Engineers and have published multiple papers regarding analytical tools and failure modes- which both tie into the goals the sponsors have for the Capstone team.

The goal of the sponsor's challenge is to add to an existing list of failure modes within artificial intelligence (AI) systems. To help the team get started, the DEVCOM associates provided multiple papers

to the Capstone team. Each of our sponsors has written academic papers in the area of reliability, however, the paper titled “Reliability Assurance for AI Systems” (Blood, 2023) is especially useful by jumpstarting the team's research into AI failure modes. The academic paper provided by them and referenced above is an incredible source of information. It provides an extensive list of possible failure modes which our group intends to expand upon.

The problem-solving effort would require a combination of technical expertise, critical thinking, research skills, ethical considerations, and effective communication. Research skills involve searching databases, academic journals, and reputable online sources related to AI failure modes. Based on the literature review and domain expertise, the team would identify potential failure modes and scenarios where AI systems could malfunction or produce undesirable outcomes. Then by utilizing a framework such as Failure Mode and Effects Analysis (FMEA), the team would assess the likelihood and severity of each identified failure mode. The team will efficiently identify and address potential failure modes in AI systems by leveraging these skills and methodologies to impact AI safety and reliability technologies (Failure Modes).

1.3 Problem Statement

The US Armed Forces utilize AI for many of their processes both combat and reconnaissance related and are asking to compile a list of failure modes and the risk associated with the application of such programs within their operations.

The US Armed Forces utilize Artificial Intelligence(AI) for many of their combat and reconnaissance processes. Since Artificial Intelligence is a new technology, its failure modes must be characterized.

1.4 Initial Customer Needs Definition

Customer needs were defined as the following:

A. Expand on the collection of failure modes (provided by DAC) including examples of them and how they could apply to Army systems
B. Consider downstream impacts of the failure modes on mission success and user trust
C. Innovative characterizations of risk involving AI
D. Identify new failure modes
E. Descriptive and robust definitions of failure modes
F. Realistic applications current and future
G. Analysis of current innovations in AI
H. Terms and definitions that benefit DEVCOM operations

1.5 Project Charter

DEVCOM is the U.S. Army combat capabilities development command. They research new technologies to use in the Army. This work is being conducted for the DEVCOM Analysis Center (DAC). Our sponsor is the Chief of the DAC Reliability Branch and works with data to ensure the reliability of technology that is to be used by the Army.

The US Armed Forces are exploring Artificial Intelligence (AI) and machine learning algorithms for many of their combat and reconnaissance processes. Since Artificial Intelligence is a new technology, its failure

modes must be characterized. A current document exists detailing the different ways AI can be used and their failure modes, but as this is a developing field, this list is neither comprehensive nor complete. This list of failure modes needs to be expanded upon along with the implications of these failures down the line.

1.5.1 Technical Objectives

This is a research-based project with no physical outcome. Initially, the group must grasp the current developments of AI and compare these findings with what DEVCOM DAC has covered.

TO1: Identify new failure modes. There is no set quantity for how many to define, it is dependent on the team's assessment of the research done. The sponsor will decide if the failure modes put forth are unique enough to fulfill a new characterization. Due to troubleshooting and validation from DEVCOM DAC, this portion of the project will take 4-6 weeks.

TO2: Analyze the downstream impact of identified failure modes. Once the team has new failure modes verified, considerations of real-world impacts can be made. This involves applying the failure modes to one or more military systems and then identifying the consequences of said failures. These examples will validate Technical Objective 1 by showing real-world proof. Validation from DEVCOM DAC will also be provided by the Sponsor.

TO3: Develop a visual map of the current failure mode characterizations DEVCOM DAC has which will be expanded upon once additional failure modes are identified. This is done to make the project quantifiable as well as to visually show the impact this capstone group will make in the upcoming semester. This objective will begin once TO1 is completed and all failure modes will be mapped in the diagram. Amendments can/will be made during the semester but are dependent on completion of technical objectives 1 and 2.

1.5.2 Special Considerations

A concern that the team has with this project is the amount of research that needs to occur. A lot of the available documentation regarding AI Failure Methods is not only new but very high level as well which oftentimes requires additional experience or knowledge via higher education. To bridge this gap of knowledge the team will seek knowledge from the faculty advisor's expertise. Additionally, with the military nature of DEVCOM, minimal access to the applications of AI within DEVCOM is made available to the team. The details of algorithms used by the U.S. Army are also loosely defined or unknown to the public. We must use public research information to gather intelligence on current AI technology.

2.0 Conceptual Design

Generating a design for a research-based project is a different approach than a conventional Capstone however through some innovation, the group has delivered this through flowcharts and a clear definition of the project's operation. Designing this mental model of the group will prove to show effectiveness regarding group uniformity of understanding.

2.1 As-Is Concept of Operations

With this project being research based the group has to sit on how the as-is Concept of Operations (CONOPS) could be illustrated. We compared ourselves to other groups that had physical As-Is characteristics and tried our best to represent the digital documents given to us as our As-Is Concept of Operations. The report DEVCOM DAC handed to us was of the most priority. This document was used to determine their current achievements on Failure Modes in AI. We gathered all the current definitions of DEVCOM's failure modes and mapped out their current relations to each other within the document.

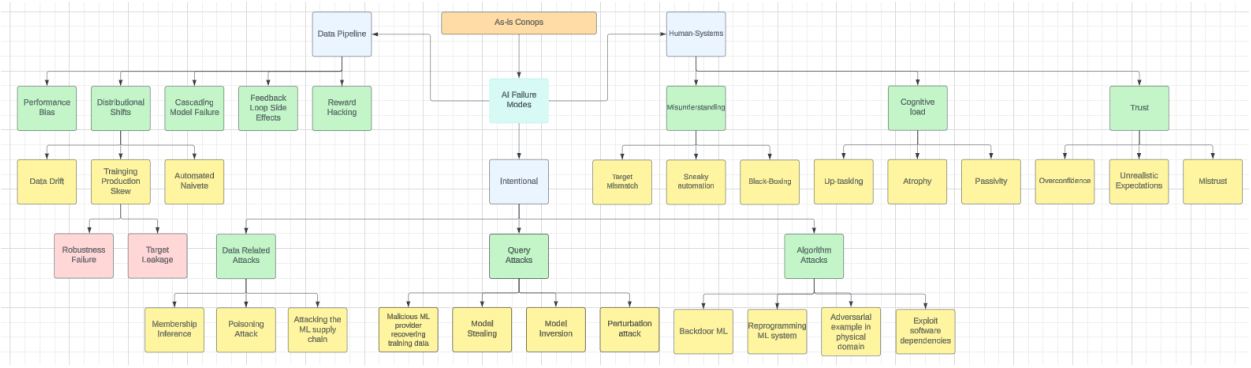

Figure 1: As-is CONOPS

This mapping of all categories, sub-categories, and Failure Modes showed our As-Is CONOPS and helped the group gain a better cohesive understanding. We achieved this because before this all we could do was ask each other if we had read the document and since reading is so subjective the process of the group gaining an even understanding was difficult. During the conception of the As-Is flowchart, the group could discuss an individual understanding of the topics and this led us to a uniform understanding of the operation as illustrated below.

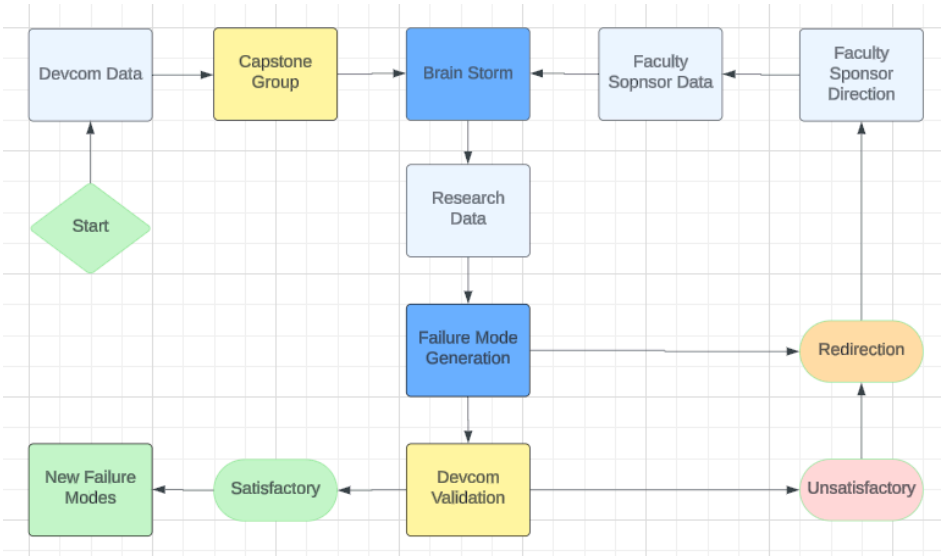

Figure 2: System of Project Operation

2.2 Refinement of Customer Needs

DEVCOM DAC asked us for new failure modes however gave us full mobility in direction when considering all the different facets of Artificial Intelligence. Through our faculty sponsor, Dr. Eksin, we encountered Federated Learning models. This topic was not something that would be categorized, but through this new AI model, we could develop a few new failure modes that were not present in the original as-is CONOPS. This direction from our faculty sponsor and members of the group's research on institutional documentation covering federated learning refined the customer's needs.

Customer Needs
N1: Expand on the collection of failure modes (provided by DAC)
N2: Demonstrate examples of new failure modes and applications to Army systems
N3: Consider downstream impacts of the failure modes on mission success and user trust

Table 1: Customer Needs

2.3 System Requirements Definition

System Requirements	Meets Need
R1:The system shall gather relevant information from external and institutional sources	N1, N2, N3
R2:The system shall identify more than one new failure mode.	N1
R3:The system shall find an application for the new failure modes.	N2
R4:The system shall allocate the failure mode into its respective category (Existing or nonexistent).	N1
R5:The system shall define the downstream impacts of failure modes.	N3
R6:The system shall validate failure modes through the sponsor's decision.	N1

Table 2: System Requirements

2.4 Engineering Actions

Engineering Actions
EA1:Analyze the current documents provided by DEVCOM
EA2:Analyze academic documents on Artificial Intelligence
EA3:Search for state-of-the-art research on AI in databases
EA4:Build off of all given reports
EA5:Brainstorm new characteristics
EA6:Describe a real-world application of new failure modes
EA7:Amend DEVCOM-validated characteristics into the given list of failure modes
EA8:Create a flow chart of all failure modes
EA9:Amend flowchart with additional failure modes
EA10:Describe the downstream impacts of real-world application
EA11:Present identified failure modes to sponsor for validation

Table 3: Engineering Actions

2.5 Needs-Requirements-Actions-Technical Objectives Traceability

Needs	Requirements	Engineering Actions	Technical Objectives
N1	R1, R2, R4, R6	EA1, EA2, EA3, EA4, EA5, EA7, EA8, EA9, EA11	TO1
N2	R1, R3	EA1, EA2, EA3, EA6	TO2
N3	R1, R5	EA1, EA2, EA3, EA10	TO3

Table 4: Traceability Matrix

3.0 Evaluation of Concepts

It is important to clarify what solution concepts are in the scope of this project. Since this is a research project, the set of solution concepts consists of the methodology our team would use to tackle the problem presented by DAC. During the initial screening, multiple methods by which we could give DAC their desired outcome were generated. However, the solutions either demand too much time or resources that are unavailable because most cutting-edge AI research information is done via large, fortune 500 companies that wouldn't share their information with nonemployed college students.

Through the guidance of our faculty advisor and also meeting with the DAC representatives, we were able to narrow it down. After an initial screening, it was clear that there were 3 different ways that a proper solution could be proposed to DAC. The solutions must meet the needs of the customer to be considered. This will be explored and expanded upon in the analysis of alternatives.

3.1 To-Be Concept of Operations

The To-Be CONOPS addresses the system requirements outlined in Phase 2 by giving a visual representation of the failure modes that have been identified within the scope of the project. This certainly fulfills each of the requirements (R1, R2, R4, R6) for our first customer need, identifying new failure modes (N1). The To-Be CONOPS intends to visually represent the failure modes that were discovered by our team along with the previous findings of DEVCOM. This was developed and adjusted throughout the project as new failure modes were discovered by the team. Small additions were made to the intentional branch such as the byzantine attack and further specification of poisoning attacks. The major branch that has been added is the multi-agent failure modes, a topic that DAC had yet to explore for failure modes. The definitions of our discovered failure modes can be found in the Detail Design section of this report.

Figure 3: To-Be CONOPS

3.2 Identification and Discussion of Alternative Solution Concepts

The first proposed solution is one in which the team focuses on the development of very specific types of failure modes in Artificial Intelligence. By focusing on important failure modes, a very detailed analysis of application and downstream impacts can be made.

The second proposed solution is to maximize the number of failure modes discovered by staying broad. Our team, by keeping the scope wide, will be able to fulfill the first customer need more effectively by being able to explore all subjects related to AI. This also allows us to determine where the scope of our project will end and establish many new failure modes.

For the third solution concept, our team decided to establish clear bounds for the scope of the project and focus on the development within a specific branch of artificial intelligence that was not initially explored by DAC. As seen by the To-Be CONOPS shown above, the new concept we decided to explore was multi-agent failure modes or more specifically, federated learning.

3.3 Analysis of Alternatives and Selection of Solution Concept for Final Design

All of the solutions generated in the solution concept exploration satisfied the customer's needs. However, we decided to evaluate our proposed solutions in conjunction with our capabilities as students. Furthermore, an additional concept that must be considered is how the solution reflects best on us as a team and the ISEN Department as a vessel of education.

While solution 1 is great to help hone in on and understand one failure mode, the problem is that it is too specific. While we may be able to become experts on just one failure mode, there are more that can be understood and elaborated on. The first customer need, identifying new failure modes, becomes constricted in the sense that we are unable to explore other topics unrelated to that failure mode. Some other drawbacks of using this approach are that it limits the scope of the project immensely and requires a lot of time to be able to do a thorough analysis that meets project requirements. The cons only worsen if we explore the idea of missing out on other important types of failure modes that are critical to system performance.

Solution 2 serves as a great way to divide and conquer. It also has a cost of efficiency and clarity within the document, since more topics will be explored. This solution does make it difficult to analyze the downstream impacts and generate noteworthy applications. Proper analysis of application and downstream impacts may also be impacted due to the sheer number of failure modes discovered. The problem becomes the scope of the project. If we all focus on different things, tying it together in the end may become difficult. Furthermore, with different amounts of availability for this project, the team may have different grades of work on their areas of focus which can come off as inconsistent and too broad for the DAC expectations.

Overall, solution 3 seemed to serve the best to align with our goals of a final product. Solution 3 enables us to stay within a clear scope while a clear analysis of the application and downstream impacts can be

performed. The solution will create a restriction on the failure modes we can identify, but by choosing an unexplored topic, our team hopes to mitigate this issue. Furthermore, this seems the direction our project is going to remain as general as possible when defining new failure modes. Essentially, we have decided to add a new category and look at some failure modes underneath that rather than only focusing on one failure mode or multiple different categories at once as suggested by solutions 1 and 2.

4.0 Detail Design

The following sections detail the final design of our project that is being delivered to DEVCOM DAC, as well as how the project fulfilled all requirements set from the beginning in our project charter.

4.1 Final Design of Selected Solution

The goal of the final design was to create a model for DEVCOM that visually shows all the failure modes they provided us and the ones we identified in one model that is easy to understand. The diagram is split into four categories which then further specifies the types of failures. As the diagram branches out, each tier is a more specific type of failure mode. This design saves time when explaining the types of failures in Artificial Intelligence to anyone within the organization. Our tax dollars are spent on these government organizations to fulfill different needs within the nation. Creating this model could save government organizations time and therefore tax dollars on understanding the failure modes of artificial intelligence. The full detailed list of each additional failure mode our team added and what it is is listed below:

Multi-Agent:

1 Communication	Failures with communication between agents.
1.1 Eavesdropping	Outside actor listens in on communication between agents opening the alley to intentional attacks.
1.2 Network Failures	Failures dealing with the network of agents.
1.2.1 Bandwidth Failures	Bandwidth limitations limit communication between agents which can lead to bad outputs.
1.2.2 Computing Failures	Large data between many agents can lead to slow computation leading to out-of-date models.
2 Accidental Steering	When multiple agents working together leads to a shift in desired output.
2.1 Group Over-optimization	When agents working together produce an over-optimized model that doesn't reflect the desired output.
2.2 Dropout Bias	When an agent becomes compromised and supplies less data overall, the model will become biased toward agents that are supplying data.
2.3 Data Heterogeneity	If the data is too different in the distribution or other facets, model aggregation can fail.

Table 5: Multi-Agent Failure Modes

Furthermore, our team divided the given intentional failure modes into three categories, giving them the same numbering system as the others. On top of this further types of attacks were identified such as Byzantine attacks and evasion attacks. To showcase this new categorization, the list is detailed below.

Intentional Failures:

1 Data-Related Attacks	Attacks where the data going into the system is targeted.
1.1 Poisoning Attacks	Attacker contaminates the training phase of ML systems to get the intended result.
1.1.1 Model Poisoning	Attacker targets the model itself.
1.1.2 Data Poisoning	Attacker manipulates the data so it changes the result undesirably.
1.2 Membership Influence	Attackers can infer if a given data record was part of the model's training data set or not.
1.3 Byzantine Attack	Attack where a client of a multi-agent system is hijacked.
1.4 Attacking the ML supply chain	Attacker compromises the ML models as it is being downloaded for use.
2 Query Attacks	Attacks using the query of an AI model.
2.1 Malicious ML provider recovering training data	Malicious ML providers can query the model used by customers and recover customers' training data.
2.2 Model Stealing	Attacker can recover the model by constructing careful queries.
2.3 Model Inversion	Attacker recovers the secret features used in the mode through careful queries.
2.4 Perturbation Attack	Attacker modifies the query to get an appropriate response.
3 Algorithm Attacks	Attacks on the algorithm of a model.
3.1 Backdoor ML	Malicious ML provides a backdoor algorithm that does not work unless triggered.
3.2 Reprogramming ML system	Repurpose the ML system to perform an activity it was not programmed for.
3.3 Adversarial example in physical domain	Attacker brings adversarial examples into the physical domain to subvert ML systems (e.g., 3D printing special eyewear to fool facial recognition systems).
3.4 Exploit software dependencies	Attackers use traditional software exploits like buffer overflow to confuse ML systems.
3.5 Evasion Attack	Attack that causes a client to not receive data updates from the central server leading to erroneous behavior.

Table 6: Intentional Failure Modes

Additionally, one of the final design deliverables we created was a downstream impact diagram. This showcases an example of an example of real-life implementation of federated learning and how the failure modes occur in this setting. It shows a central computing system linked to multiple drones. In this environment, the central computer has the main AI model and communicates with the drones which each have their own respective model. The downstream impacts of most failure modes related to federated learning would likely require a reset of the model or cause other failure modes to occur which will lead to unpredictable or malicious behavior. The downstream impacts can be mitigated by bringing to light the potential failures which can occur in federated learning systems, since operators will be better educated and understand how the failures occur.

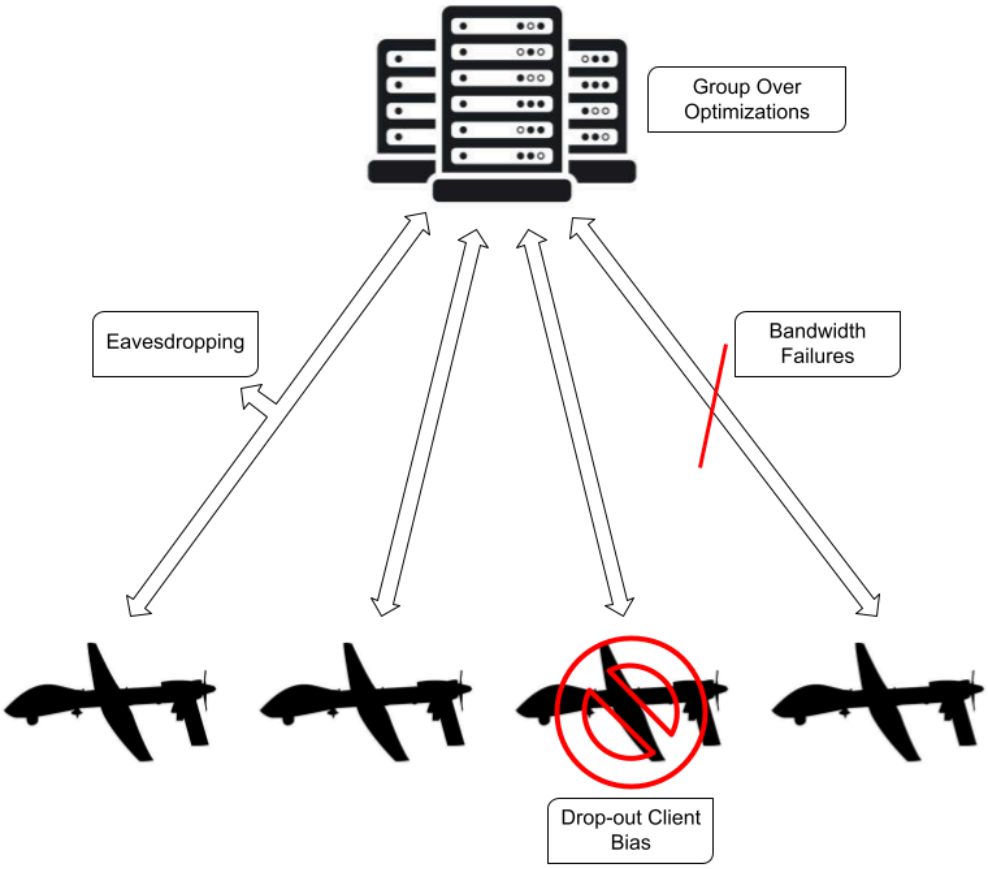

Figure 4: Downstream Impact Diagram

4.1.1 Detail Design Architecture

We went with the visual design of our diagram to easily show the map of failure modes. Initially all of the data was in tables and serialized but it seemed abstract to us. The intention is to make it digestible and

understandable by anyone, even those without technical knowledge of A.I. and its systems. The flow chart shows how the failure modes are divided into categories and how we generalize failure modes from broad subjects to a specific way the system can fail. This speeds up getting people up to date on what failure modes have been identified and hopefully also help people using A.I. systems identify failure modes that are actively occurring. This visualization saves time by making this complicated topic digestible for anyone.

The downstream impact diagram was created to go in pair with the Failure Mode Diagram. It gives an example of a federated learning system that showcases the failures and where they occur in the system. These two diagrams paired together make a comprehensive and understandable view of our projects final design solution given to DEVCOM DAC.

4.1.2 Requirement Traceability

By developing this model, all requirements were either met or had progress made. The diagram contains all relevant institutional information gathered in a well-suited illustration. Our primary customer need was to identify new failure modes which we fulfilled through research (R1, R2). Overall the new failure modes we identified came through specifically researching federated learning, a multi-agent application of AI that is used to build predictive models (R3). We learned how these types of AI systems work and then identified failure modes from that knowledge. Then, using DEVCOM's current structure we put the failure modes into the three existing categories and added our new category "Multi-Agent Failures," which was the main focus of research in our project (R4). To add a failure mode to the diagram we presented the failure mode to DEVCOM by explaining an example of an AI system that uses our application of focus (federated learning), describing how the system could fail in that way, and the downstream impact of what that failure could lead to (R5, R6). If it checked every box it was validated by DEVCOM and was added to the chart as a new failure mode. Overall this diagram represents all six system requirements being fulfilled.

Needs	Requirements	Engineering Actions	Solution Concepts	Final Design Architecture
N1	R1, R2, R4, R6	EA1, EA2, EA3, EA4, EA5, EA7, EA8, EA9, EA11	S3	New Failure Modes have been identified and added to Final Diagram.
N2	R1, R3	EA1, EA2, EA3, EA6		Example provided in Downstream Impact Diagram, and during process of validation.
N3	R1, R5	EA1, EA2, EA3, EA10		Downstream Impact Diagram showcases impact of failure modes and again the process of validation included this requirement.

Table 7: Final Design Traceability Matrix

4.2 Final design Cost Analysis

When looking at the decided solution from the lens of cost, both money and time can be used as metrics that result in a cost to DEVCOM. When taking money into consideration, by giving DEVCOM a subcategory that's been fully explored, it is less money they have to spend on resources to look into that

subcategory altogether. Furthermore, by also expanding on downstream impacts DEVCOM can use less money and time that can be spent on training and other projects. By understanding the failure modes involved in federated learning and how they can affect military operations, DEVCOM can allocate their resources more efficiently thus reducing cost through wasted time and money.

4.3 Final Design Performance Analysis

From a performance lens, DEVCOM will be furthering its work by using our project to explain common failures in AI and machine learning algorithms. This framework has not yet been used by DEVCOM for evaluating currently used AI systems but is more of an aid to further the general knowledge of failure modes in AI systems. So far, we have fulfilled our original goals and identified multiple new failure modes and presented them to our project sponsor. They responded well and liked the direction we went with the research we have done in federated learning. In addition to the failure methods our project has produced, a diagram showcasing all the failure modes we started with and added is included. The review of the diagram was received positively by DEVCOM as they liked the visual aspect of this design.

4.4 Final Design Risk Analysis

Lastly, there are also some risks associated with our work and what we put forward. A big part of it is applicability. While we can put forth work that is relevant to DEVCOM, we do not have the full scope due to our lack of access to classified information. In our eyes, what we are putting forward might apply to our sponsor's needs, however, it truly might not be. Due to the nature of security related to the work DEVCOM does, the project sponsors were not able to explicitly define the needs and goals of the project. Furthermore, there are more AI failure modes that we have not touched on. To mitigate the number of failure modes that have not been explored, we attempted to be as comprehensive with the ones we were able to put forward so that future efforts to expand the research can be more focused on new research and failure modes and the ones we put forward less.

5.0 Conclusions and Recommendations

This section goes into detail on the conclusions, recommendations, and potential future of this project.

5.1 Conclusions

In the dynamic field of artificial intelligence (AI) and machine learning (ML) algorithms, understanding the failure modes of these models is of great importance. As the use cases of AI and ML models continue to increase, so does the need for robust and reliable systems of validity.

We started with analyzing the current research that was provided by DEVCOM. After an initial analysis, we were recommended to continue the work made by previous capstone projects and to extend the current list of failure modes that were provided. Our only alternative to this very direct need for research was the group decision's strategy on how to adequately cover new or existing categories of failures. We sought direction from our project sponsor and faculty sponsor, Dr. Eksin.

The area that was recommended to us by our faculty sponsor to extend the existing research was federated learning. Federated learning is a method of training an algorithm in which there is a central model that is updated with training from many devices or servers holding local data samples. This training method presents its own unique set of failure modes which became the primary focus of the project.

In addition to federated learning, the group spent time on dedicating and amending previously discovered failure modes- ensuring they were up to date with the current novelties of Artificial Intelligence. Our method was simple as the group moved through each characterization together, only moving forward once it was all complete. This method made us all in tune with the current state of the project and democratized the process of failure mode generation.

5.2 Recommendations

Moving forward, we recommend that DEVCOM continue this research on failure modes and effects of AI and ML with context to US Army applications. Since AI is a field that is constantly growing and changing, new developments in all of the different identified subcategories will continue to happen. Therefore, we also recommend that DEVCOM revisit new literature on federated learning next year. This field of technology is rapidly evolving and some innovations can be added to this research. We have created a package of documents for the continuation of this research which includes: infographics, research papers, technical descriptions, and an updated failure mode and effect analysis.

5.3 Future Work Considerations

If this project were to be continued by another capstone team, it would be recommended to expand the list of failure modes after reading through the previous research. Areas that this team has brainstormed but did not pursue include side-channel attacks and failures that are involved with other methods of training an artificial intelligence or machine learning algorithm.

With new developments and innovations in the use of artificial intelligence and machine learning, there will be new research that will need to be analyzed. A good journal to look for new research in is IEEE.

Articles that are related to new methods of training an algorithm as well as applications to artificial intelligence should be considered with dates published in 2023 and after should be considered. Novel applications of machine learning and artificial intelligence should also be considered as they may be of interest to US Army DEVCOM.

6.0 References

This is a list of sources for new research that were found and included in the project by our team.

[1] N. Bouacida and P. Mohapatra, "Vulnerabilities in Federated Learning," in IEEE Access, vol. 9, pp. 63229-63249, 2021, doi: 10.1109/ACCESS.2021.3075203.

[2] J. Martinez, A. Eguia, I. Urretavizcaya, E. Amparan and P. L. Negro, "Fault Tree Analysis and Failure Modes and Effects Analysis for Systems with Artificial Intelligence: A Mapping Study," 2023 7th International Conference on System Reliability and Safety (ICSRS), Bologna, Italy, 2023, pp. 464-473, doi: 10.1109/ICSRS59833.2023.10381456.

[3] Manheim D. Multiparty Dynamics and Failure Modes for Machine Learning and Artificial Intelligence. Big Data and Cognitive Computing. 2019; 3(2):21.

[4] S. Tyagi, I. S. Rajput and R. Pandey, "Federated learning: Applications, Security hazards and Defense measures," 2023 International Conference on Device Intelligence, Computing and Communication Technologies, (DICCT), Dehradun, India, 2023, pp. 477-482, doi: 10.1109/DICCT56244.2023.10110075.

7.0 Appendix

This section provides lists of all documentation that was previously included in the latest iteration of this capstone project and documents provided by US Army DEVCOM. The list of references in section 7.2 is a compilation of references that are included in documents 2, 3, and 4 of section 7.1. These documents are the foundation of the research for this project.

7.1 Supporting Data & Documentation

[1] Herbert, Hum, et al. “Ensured Reliability for Artificial Intelligence Systems”. DEVCOM DAC-TR-2023-084 (December 2023)

[2] Cantrell, Guzman, et al. “Final Report for Team 7: Reliability Artificial Intelligence”. (December 2023)

[3] Brannon, Criscione, et al. “Final Report for Team 18: Towards Reliable Artificial Intelligence”. (May 2023)

[4] Drake, Ramirez, et al. “Final Report for Team 26: U.S Army Reliability Artificial Intelligence”. (December 2022)

7.2 Peer Provided References

[1] NIST AI RMF Playbook

[2] Soldati, Pablo, et al. "Design Principles for Generalization and Scalability of AI in Communication Systems." arXiv preprint arXiv:2306.06251 (2023).

[3] Chassang, Gauthier, et al. "An interdisciplinary conceptual study of Artificial Intelligence (AI) for helping benefit-risk assessment practices: Towards a comprehensive qualification matrix of AI programs and devices (pre-print 2020)." arXiv preprint arXiv:2105.03192 (2021).

[4] Siebert, Julien, et al. "Towards guidelines for assessing qualities of machine learning systems." Quality of Information and Communications Technology: 13th International Conference, QUATIC 2020, Faro, Portugal, September 9–11, 2020, Proceedings 13. Springer International Publishing, 2020.

[5] Soldati, Pablo, et al. "Design Principles for Generalization and Scalability of AI in Communication Systems." arXiv preprint arXiv:2306.06251 (2023).

[6] Piorkowski, David, Michael Hind, and John Richards. "Quantitative ai risk assessments: Opportunities and challenges." arXiv preprint arXiv:2209.06317 (2022).

[7] Hoffman, Robert R., et al. "Metrics for explainable AI: Challenges and prospects." arXiv preprint arXiv:1812.04608 (2018).

[8] Lohn, Andrew J. "Estimating the brittleness of AI: Safety integrity levels and the need for testing out-of-distribution performance." arXiv preprint arXiv:2009.00802 (2020).

[9] Ish, Daniel, Jared Ettinger, and Christopher Ferris, Evaluating the Effectiveness of Artificial Intelligence Systems in Intelligence Analysis, RAND Corporation, RR-A464-1, 2021. As of November 4, 2023: https://www.rand.org/pubs/research_reports/RRA464-1.html

[10] Siebert, Julien & Joeckel, Lisa & Heidrich, Jens & Trendowicz, Adam & Nakamichi, Koji & Ohashi, Kyoko & Namba, Isao & Yamamoto, Rieko & Aoyama, Mikio. (2022). Construction of a quality model for machine learning systems. Software Quality Journal. 30. 1-29. 10.1007/s11219-021-09557-y.

[11] Munoz, Cristian, et al. "Local and Global Explainability Metrics for Machine Learning Predictions." arXiv preprint arXiv:2302.12094 (2023).

[12] Hong, Yili, et al. "Statistical perspectives on reliability of artificial intelligence systems." Quality Engineering 35.1 (2023): 56-78.

[13] Nakamichi, K., Ohashi, K., Namba, I., Yamamoto, R., Aoyama, M., Joeckel, L., ... Heidrich, J. (2020). Requirements-Driven Method to Determine Quality Characteristics and Measurements for Machine Learning Software and Its Evaluation. Los Alamitos, Calif.: IEEE.

[14] Lee, Eungyu, Lee, Yongsoo, & Lee, Taejin. (2023). Adversarial Attack-Based Robustness Evaluation for Trustworthy AI. Computer Systems Science & Engineering, 47(2), 1919–1935. https://doi.org/10.32604/csse.2023.039599

[15] W. Wan, Y. Meng, B. Shang, X. Li, B. Mo and S. Rong, "Reliability Assessment Scheme for Intelligent Autonomous System," 2022 13th International Conference on Reliability, Maintainability, and Safety (ICRMS), Kowloon, Hong Kong, 2022, pp. 285-289, doi: 10.1109/ICRMS55680.2022.9944598.

[16] Chih-Ling Chang, Jui-Lung Hung, Chin-Wei Tien, Chia-Wei Tien, and Sy-Yen Kuo. 2020. Evaluating Robustness of AI Models against Adversarial Attacks. In Proceedings of the 1st ACM Workshop on Security and Privacy on Artificial Intelligence (SPAI '20). Association for Computing Machinery, New York, NY, USA, 47–54. https://doi.org/10.1145/3385003.3410920

[17] Cai, B., Sheng, C., Gao, C., Liu, Y., Shi, M., Liu, Z., ... Liu, G. (2023). Artificial Intelligence Enhanced Reliability Assessment Methodology With Small Samples. Ieee Transactions on Neural Networks and Learning Systems, 34(9), 6578-6590.

[18] McCall, R., McGee, F., Mirnig, A., Meschtscherjakov, A., Louveton, N., Engel, T., & Tscheligi, M. (2019). A taxonomy of autonomous vehicle handover situations. Transportation Research Part A-Policy and Practice, 124, 507-522.

[19] Barocas, S., Guo, A., Kamar, E., Krones, J. Ringel Morris, M., Wortman Vaughan, J. Wadsworth, D., Wallach, H., (2021). Designing Disaggregated Evaluations of AI Systems: Choices, Considerations, and Tradeoffs, 2103.06076

[20] Cowley, Hannah & Natter, Mandy & Gray-Roncal, Karla & Rhodes, Rebecca & Johnson, Erik & Drenkow, Nathan & Shead, Timothy & Chance, Frances & Wester, Brock & Gray-Roncal, William. (2022). A framework for rigorous evaluation of human performance in human and machine learning comparison studies. Scientific Reports. 12. 10.1038/s41598-022-08078-3.

[21] Yili Hong, Jiayi Lian, Li Xu, Jie Min, Yueyao Wang, Laura J. Freeman & Xinwei Deng (2023). Statistical perspectives on reliability of artificial intelligence systems, Quality Engineering, 35:1, 56-78, DOI: 10.1080/08982112.2022.2089854

[22] Principles of Human Reliability Analysis. (2010). Joint RES/EPRI Fire PRA Workshop. Washington, DC. Retrieved from https://www.nrc.gov/docs/ML1025/ML102560372.pdf.

[23] Anthony Corso, David Karamadian, Romeo Valentin, Mary Cooper, Mykel J. Kochenderfer. (2023) A Holistic Assessment of the Reliability of Machine Learning Systems, arXiv:2307.10586.

[24] Jha, Susmit and Raj, Sunny and Fernandes, Steven and Jha, Sumit K and Jha, Somesh and Jalaian, Brian and Verma, Gunjan and Swami, Ananthram (2019). Attribution-Based Confidence Metric for Deep Neural Networks, Advances in Neural Information Processing Systems, 32, https://proceedings.neurips.cc/paper_files/paper/2019/file/bc1ad6e8f86c42a371aff945535baebb-Paper.pdf

[25] Failure mode and effects analysis (FMEA). ASQ. (n.d.). Retrieved November 15, 2022, from https://asq.org/quality-resources/fmea

[26] Hessing, T. (2021, June 2). Fault Tree Analysis. Six Sigma Study Guide. Retrieved November 16, 2022, from https://sixsigmastudyguide.com/fault-tree-analysis/

[27] Etkin, McCay, Horn, Landquist, Hassellov, Wolfword. (n.d.). Fault tree diagram. Fault Tree Diagram - an overview | ScienceDirect Topics. Retrieved November 16, 2022, from https://www.sciencedirect.com/topics/engineering/fault-tree-diagram

[28] The bowtie method. Wolters Kluwer. (n.d.). Retrieved November 15, 2022, from https://www.wolterskluwer.com/en/solutions/enablon/bowtie/expert-insights/barrier-based-risk-management-knowledge-base/the-bowtie-method

[29] Van de Schoot, R., Depaoli, S., King, R. et al. Bayesian statistics and modeling. Nat Rev Methods Primers 1, 1 (2021). https://doi.org/10.1038/s43586-020-00001-2

[30] Artificial Intelligence Risk Management Framework, NIST, 2023

[31] Information Technology – Artificial Intelligence – Guidance on risk management, ISO/IEC 23894:2023

[32] Proposed EU AI Act, EU, 2023

[33] Recommendation of the Council on Artificial Intelligence, OECD, 2019

[34] Quality Management Systems - Fundamentals and vocabulary, ISO 9000:2015

[35] Trustworthiness – Vocabulary, ISO/IEC TS 5723:2022

[36] Dr. Michael Gaither, SME interview, April 12, 2023

[37] Robert Brydia PMP, SME interview, March 27, 2023

[38] E. Blasch, J. Sung, and T. Nguyen, “Multisource AI scorecard table for system evaluation”, 2021

8.0 ABET Outcomes

8.1 Student Outcomes

This section includes a table of student outcomes as well as two essays associated with the project. The table of student outcomes outlines the learning objectives associated with the project. The essays go over the ethical and professional responsibilities as well as the broader implications of this project.

8.1.1 Student Outcomes Table

Student Outcome		Outcome Refinement		Capstone Application	Rationale
1	An ability to identify, formulate, and solve complex engineering problems by applying the principles of engineering, science, and mathematics	1-a	An ability to identify, formulate, and solve complex engineering problems by applying principles of engineering	Creation of Final Diagram	We had to understand AI systems through research in order to identify failure modes often applying principles of engineering.
		1-b	An ability to identify, formulate, and solve complex engineering problems by applying principles of science	Creation of Final Diagram	We had to understand AI systems through research in order to identify failure modes often applying principles of science.
		1-c	An ability to identify, formulate, and solve complex engineering problems by applying principles of mathematics	Cost Analysis	We considered the cost impact analysis of our project for DEVCOM.
2	An ability to apply engineering design to produce solutions that meet specified needs with consideration of public health, safety, and welfare, as well as global, cultural, social, environmental, and economic factors	2-a	An ability to apply engineering design to produce solutions that meet specified needs with consideration of public, health, safety, and welfare	• Applies to As-Is CONOPS	When designing the As-Is-CONOPS, engineering design techniques were used to visualize the current state of the system. Additional failure modes found intend to improve the safety and reliability of AI Systems.
		2-b	An ability to apply engineering design to produce solutions that meet specified needs with consideration of global, cultural, social, environmental, and economic factors	• Applies to As-Is CONOPS Project Flowchart	When designing the As-Is-CONOPS, engineering design techniques were used to visualize the current state of the system. The rearrangements intend to improve knowledge on failure modes within AI systems, ultimately increasing productivity of society
3	An ability to communicate effectively with a range of audiences	3-a	An ability to communicate effectively in writing with internal and external audiences	• Capstone deliverables and communicating with DEVCOM	• We have communicated details about our project in many different deliverables.
		3-b	An ability to communicate effectively orally with internal and external audiences	• Presentations and meetings with sponsor.	• We have communicated our project to the class effectively and communicated our research with our sponsor.
4	An ability to recognize ethical and professional responsibilities in engineering situations and make informed judgements, which must consider the impact of engineering solutions in global, economic, environmental, and societal contexts	4-a	An ability to recognize ethical and professional responsibilities in engineering situations and make informed judgements, which must consider the impact of engineering solutions in global and societal contexts	Requirements of the Project	It was an ethical requirement for us to perform good research for DEVCOM and provide accurate information.
		4-b	An ability to recognize ethical and professional responsibilities in engineering situations and make informed judgements, which must consider the impact of engineering solutions in economic and environmental contexts	Requirements of the project	This project considered economic impact and made sure to help DEVCOM with research saving the time and money.
5	An ability to function effectively on a team whose members together provide leadership, create a collaborative and inclusive environment, establish goals, plan tasks, and meet objectives	5-a	An ability to function effectively on a team whose members together provide leadership, create a collaborative and inclusive environment	• Code of Conduct • Weekly Meetings	The enforcement of a code of conduct for our meetings and presentation has allowed us to create an inclusive environment for all team members.
		5-b	An ability to function effectively on a team whose members together provide leadership, create goals, plan tasks, and meet objectives	• Weekly Meetings • Presentations and Practice	Each week, meetings and practice for presentations are held to keep the team on track and establish responsibilities.
6	An ability to develop and conduct appropriate experimentation, analyze and interpret data, and use engineering judgement to draw conclusions	6-a	An ability to develop and conduct appropriate experimentation, analyze and interpret data	• Research Failure modes • Brainstorming	• We analyzed many research papers to develop our new failure modes
		6-b	An ability to develop and conduct appropriate experimentation, using engineering judgement to draw conclusions	• Validation process with DEVCOM	• The validation process involved trial and error with additional potential failure modes.
7	An ability to acquire and apply new knowledge as needed, using appropriate learning strategies	7-a	An ability to acquire and apply new knowledge as needed, using appropriate learning strategies	• We have read technical papers and used the knowledge to define new failure modes in accordance with the customer requirements.	• This fulfills this outcome because we used new knowledge and applied it to our project.

8.1.2 Ethical and Professional Responsibilities

1. Background:

The capstone project that we are involved with is to perform research on behalf of the United States Army DEVCOM group on the topic of failure modes associated with different types of artificial intelligence and machine learning algorithms. The algorithms that the military uses have different potential sources of failure depending on the context of the intended use. The project sponsors have tasked my team to expand an existing list of failure modes and to update a risk assessment model with

the research we perform. We are then asked to perform an example risk analysis on a system that is publicly known.

2. Global Implications:

The global implications of this project are involved in many industries. From transportation to nuclear weapons and national security, algorithms are not involved in assisting humans with doing work more efficiently and quickly. These technologies can have both positive and negative benefits to society depending on their implementation and use. Currently, only the most technologically advanced nations and militaries are using artificial intelligence to assist with their operations, but the use of these technologies is expanding rapidly, impacting more people each day.

3. Economic Impact:

The economic impact of artificial intelligence can have some great impacts on the cost of some operations. Algorithms can be very beneficial in reducing routing time and costs associated with moving materials and people as well as predicting expected needs of resources. Some commonly used algorithms involve routing vehicles through public roads and research is being done with self-driving vehicles. This capstone project will produce a risk analysis model that will be able to assist with risk assessment of certain types of algorithms involved in autonomous vehicles and devices.

4. Environmental Impact:

Machine learning and artificial intelligence can be used with environmental data to model changes in terrain over some time. Many factors are involved with working with environmental data which creates a very complicated model. A machine learning algorithm is ideal for a case with many input predictors. A case of this project goes over types of failures associated with data pipeline errors. An example that was introduced by our project sponsors was that a snow-covered terrain and a sand-covered terrain can appear like a computer when it is given an aerial photo. This creates challenges in which false classification and identification can occur. In this project, we aim to categorize the severity of types of failures like this to determine the possible impact this could have on an overall system.

5. Societal Considerations:

In the past few years, artificial intelligence and machine learning algorithms have become increasingly more implanted in many aspects of society. Large language models have been developing at a rapid rate which is enhancing the ability for people to query information out of a computer. We are investigating failure modes and attacks on these systems as well as other systems that are expected to receive common use by the public. The development of autonomous vehicles opens the possibility of many risks to the public with many possible methods of failure.

8.1.3 Broader Implications of Engineering Work

1. Background:

Our assignment within the Capstone group has been to research new and relevant findings while creating a model that can help DEVCOM DAC characterize failure modes in Artificial Intelligence. DEVCOM DAC is a military research group heavily involved with data analytics. As Artificial Intelligence is on the rise in this sector of technology, they have been tasked by the military to assess its current state

as it transpires in the years to come. This is no easy task and with it being on the forefront of technology in both the public and private sectors of the country DEVCOM is looking for innovation from younger minds that are more in tune with the technology capable of hosting Artificial Intelligence.

2. Public Health:

Finding these failure modes is vital to public health as this technology is currently in use while the public has little to no understanding of how it works or its implications. This technology cannot be deemed to be fully harmless to the public due to external and internal errors that occur naturally in our society. We agree with DEVCOM that the implications this technology has on the public are drastic for better or worse. Our work and design of these new characterizations of failure modes get us more involved with its negative aspects of operation. This is because the military and government organizations want to be prepared for those situations more than they would need to be since part of governing is mostly related to when things are far beyond one person or organization's control.

3. Safety:

Human Safety is the priority within this design and development. The belief that knowledge is power is suitable to this situation due to the amount of research being covered within the assignment. The public does not allow much time for explanation due to education being limited as well as time consuming. As we create a model for DEVCOM this assignment helps our government cut time on explanation and increases time for developments. Environmentally, as we mitigate failures through defining them its implications on the real world is that less damage leads to less catastrophe as these Artificial Intelligent tools are applied at industrial scales.

4. Welfare:

Is Artificial Intelligence good for our society? Was a question we asked in the previous assignment. As this technology progresses, we must determine how effective it is to our problems and if it has a positive impact or not to the welfare of society. The less failures we encounter through Artificial Intelligence the better our economy should be, especially as its influence grows within our current protocols. It is the duty of the people to deem how much of a good this is for us. Globally, we advocate for sustainable innovations through AI that expedite solutions to our current challenges.

Conclusion:

In conclusion, because of the reasons stated above this assignment embodies the ABET outcome number 2. Designing and maintaining something that encompasses all characteristics of design, prioritizing public health, safety, and welfare is not easy, however through brainstorming and innovation our group has managed to stay on track and systemically hit all the marks.

8.2 Curriculum Outcomes

In this final section, we cover the engineering standards that are associated with the project as well as the constraints we faced with the project.

8.2.1 Applied Engineering Standards

This project did not utilize any standards, but the documentation provided in document 1 of 7.1 states "In addition, ensure compliance with relevant regulations and industry standards (e.g., ISO/IEC 27001 [ISO/IEC, 2022b], and DOD Directive 3000.09 [DOD, 2023])." . The standards listed include the military policy and guidelines for Autonomy in Weapons Systems and the International Standards Organization Security Control Framework.

8.2.2 Project Constraints

This project was required to be done with a minimum amount of knowledge on the applications of artificial intelligence and machine learning by US Army DEVCOM. This constraint made understanding the impact and use of the research we were finding difficult, but with consistent communication with our capstone sponsors we were able to validate an area of interest and find several supporting documents.