Functional Hazard Analysis (FHA) and Subsystem Hazard Analysis (SSHA) of Artificial Intelligence/Machine Learning (AI/ML) Functions within a Sandbox Program

Presented for the Acquisition Research Symposium, May 11-13, 2021
Panel: Ship Maintenance and Acquisition

Bruce Nagy
NAVAIR

Gunendran Sivapragasam
NAVSEA

Loren Edwards
NAVAIR

Motivation: Naval Ordnance Safety and Security Activity (NOSSA) Investigating Policy and Guidelines Specific to AI/ML Functions

How can we improve our weapon systems using AI and still deploy safe systems?

It's "likely" we are as safe as using non-Ai technology!

Only "likely" as safe?

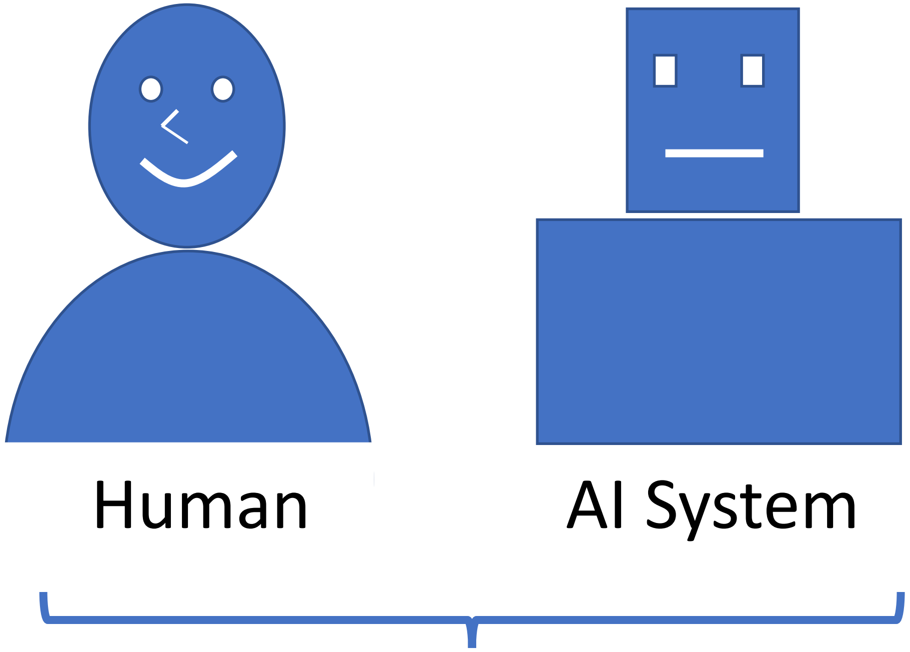
I suggest we ask NOSSA for guidance!

The strength of AI is also its weakness!

AI/ML is a New Development Paradigm

When expected to successfully perform critical tasks, the **Human** needs the "***right/correct***" ***training*** and ***incentives*** to consistently meet expectations.

Expectations need to define a ***likelihood*** that he/she will be successful ***most of the time***.



When expected to successfully perform critical tasks, the **AI System** needs the "***right/correct***" ***training*** and ***algorithm*** to consistently meet expectations.

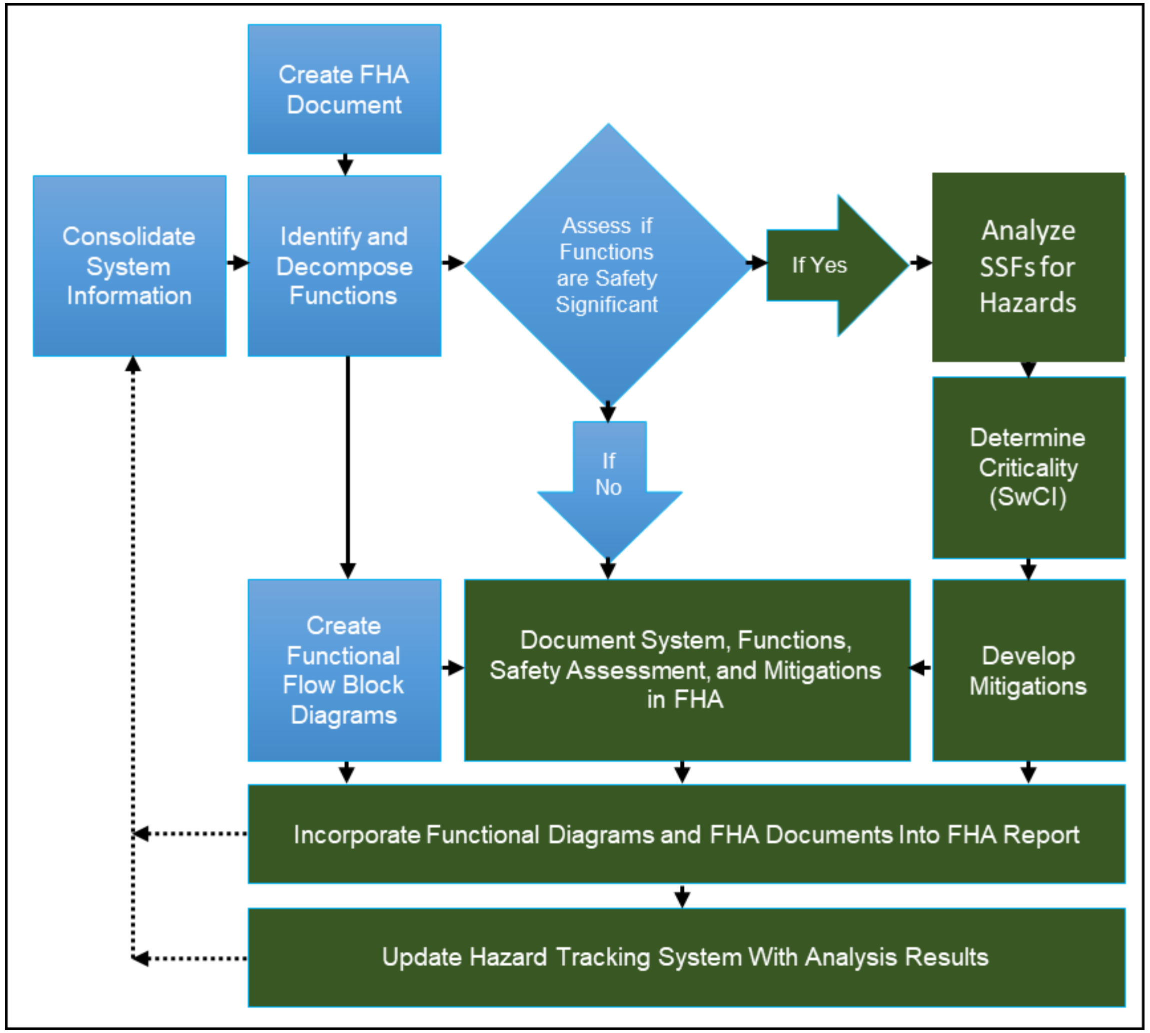
Expectations need to define a ***likelihood*** that the machine will be successful ***most of the time***.

This comparison "right/correct" training analogy applies to AI developed code but not traditional code.

Basic System Safety definitions:

- Software Control Category (SCC) -- A numeric number resulting from applying a standard method to categorize safety significant software based on its level of autonomy.
- Software Criticality Index (SwCI) -- A numeric number resulting from a combination of SCC and severity to determine the LOR tasks required for safety significant software.
- Level of Rigor (LOR) -- per MIL-STD-882 “A specification of the depth and breadth of software analysis and verification activities necessary to provide a sufficient level of confidence that a safety-critical or safety-related software function will perform as required.“ A specific set of tasks to be completed before that safety significant software is considered “safe” or representing a certain level of acceptable risk for the system.
- Functional Hazard Analysis (FHA) – The primary analysis used to determine SCC and SwCI determinations for safety significant software. Each function is evaluated for level of autonomy and safety criticality.
- Subsystem Functional Hazard Analysis (SSHA) – A detailed subsystem analysis used to determine LOR for safety significant software.

FHA Workflow Conducted by System Safety Practitioners

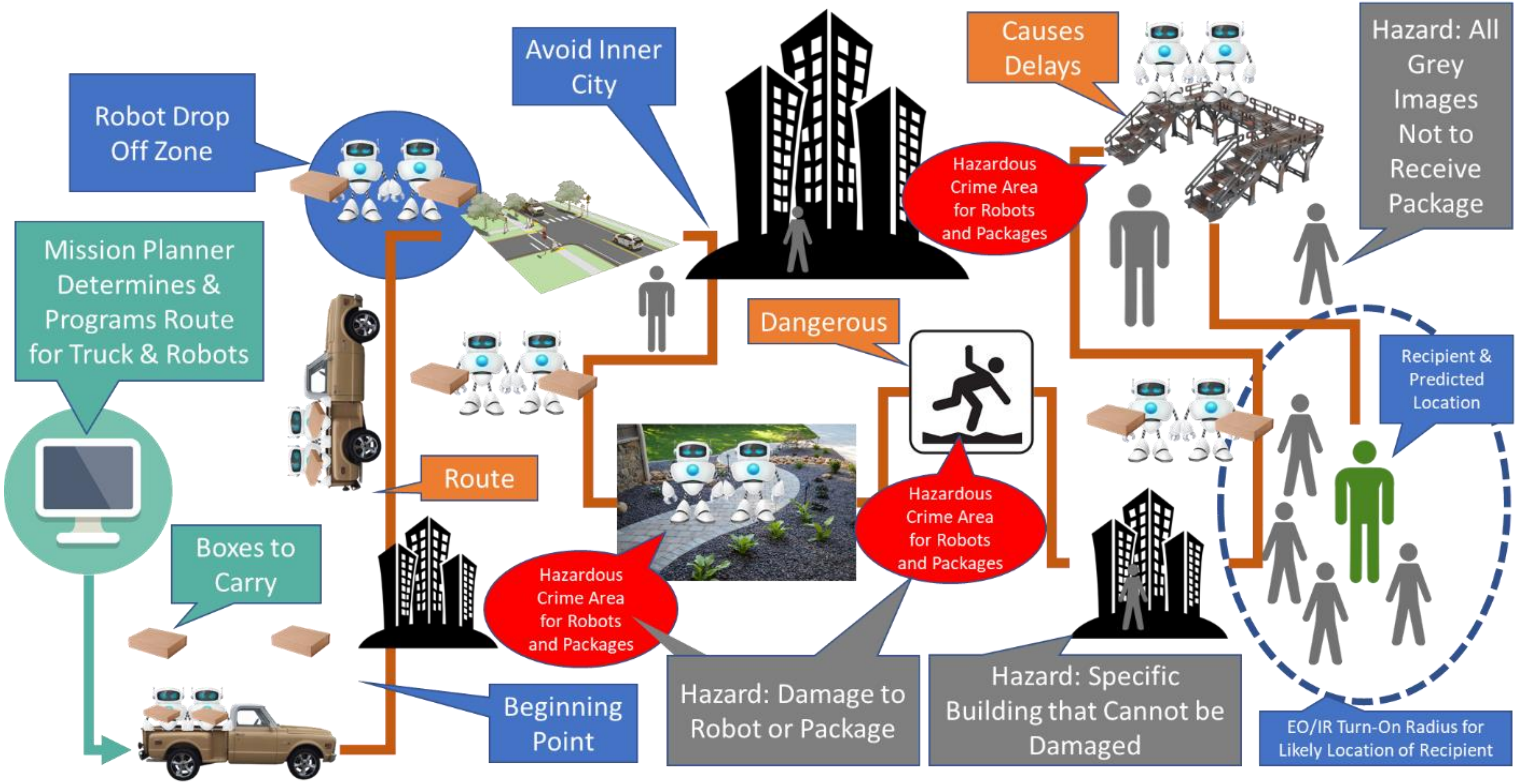


Stakeholder's Analysis Table (Subset of List)

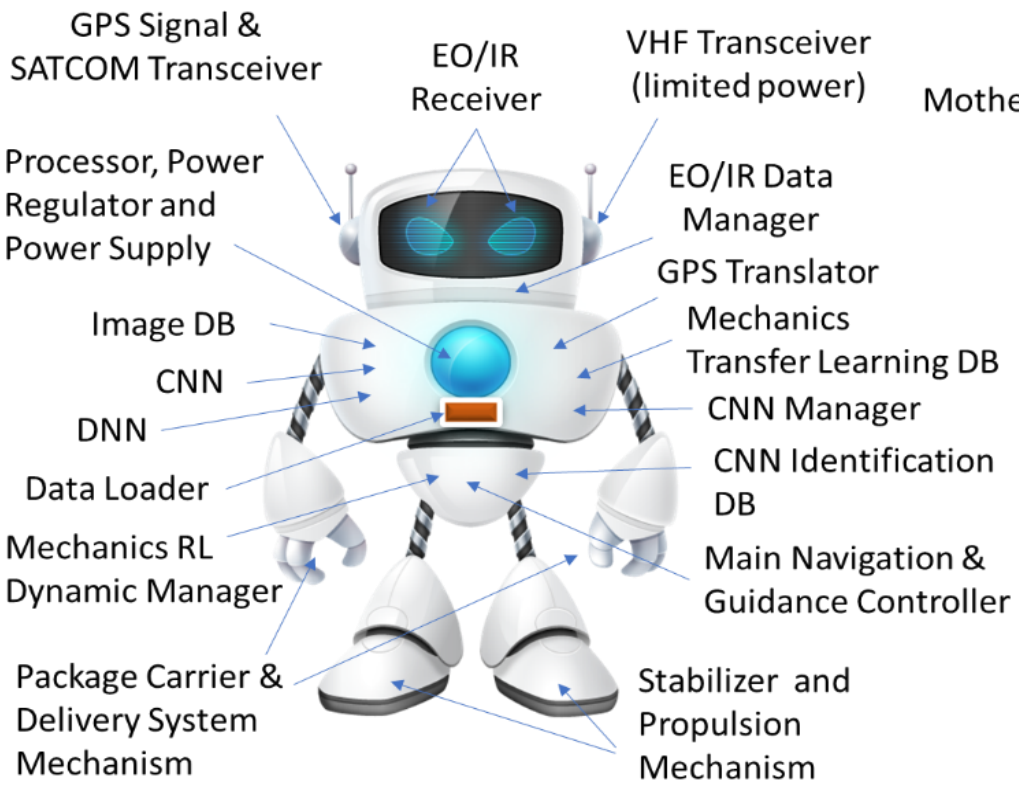
	Name/Organization	Type	Want/Need	Concern/Loss
9	NOSSA	Sponsor	What tools, guidance and documentation would need to be created to support the processes and policy per each group's needs? Groups: Developers need from system safety, System safety practitioners from system safety and Oversight folks from system safety.	NOSSA: Unsafe deployed system
10	NOSSA	Sponsor	Along with the processes, what analytics need investigation for each user group?	NOSSA: Unsafe deployed system
11	NOSSA	Sponsor	How would various AI/ML software designs affect the analytical approach?	NOSSA: Unsafe deployed system
12	NOSSA	Sponsor	What kind of OQE is required per a given AI/ML technique and implementation structure to support a program moving forward?	NOSSA: Unsafe deployed system
13	NOSSA	Sponsor	Will data and analytics be considered as separate pieces to inspect?	NOSSA: Unsafe deployed system
14	NOSSA	Sponsor	During a WSESRB or Technical Review Panel review that involves AI/ML, how would systems, data and numbers be presented to allow for proper investigation and analysis to ensure contextual accuracy based on group technical background?	NOSSA: Unsafe deployed system
15	NOSSA	Sponsor	What are the factors and limitations associated with confidence of numbers presented regarding AI/ML performance?	NOSSA: Unsafe deployed system
16	NOSSA	Sponsor	AI/ML performance is always associated within the context of the training data?	NOSSA: Unsafe deployed system
17	NOSSA	Sponsor	What does it mean to perform architecture, design, or code analysis (see MIL-STD-882E Table V) with an AI/ML system, especially when, for example, even the developer has limited understanding on how the neural network works?	NOSSA: Unsafe deployed system

Note: NOSSA is investigating software safety processes to appropriately address ML/AI.

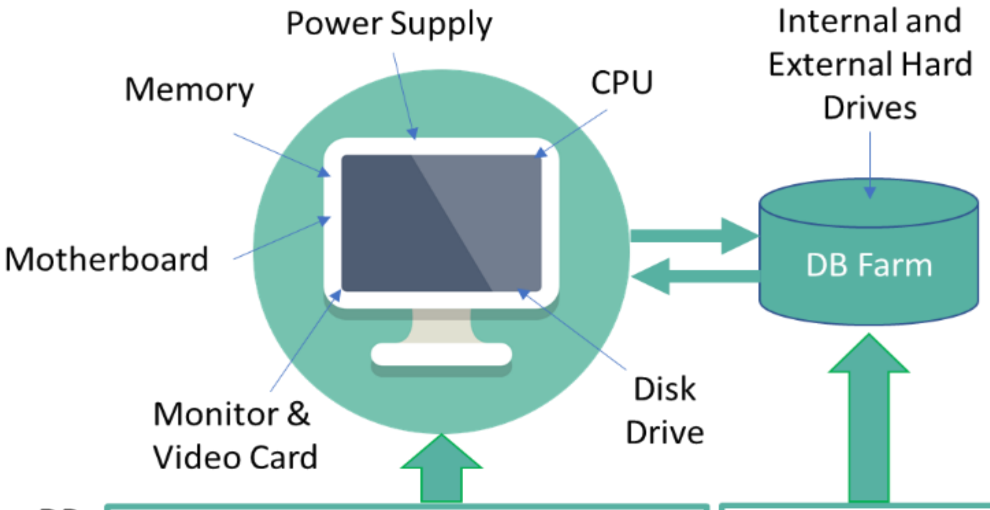
Operational Use Case of Two Robots Delivering Packages



Robot and Mission Planner Subsystems



Note: Robot includes other passive sensors not listed but used to support navigation and guidance.



Loaded Software Containing the Following Code:
- Route Segments Algorithm
- Mapping Segments Algorithm
- Legs ID Algorithm
- Mapping Characteristics Algorithm
- Candidate Selection Algorithm
- Naïve Bayes for Classification Algorithm
- Random Forest Algorithm
- Naïve Bayes for Statistics Determination Algorithm
- Minimax Algorithm
- Group Legs Algorithm
- Temporal Greedy Search Algorithm
- GUI Objects
- GUI Object Manager
- Data Base Manager
- Statistical Graphics Package
- Data Loader Manager
- Data Collection & Processing Algorithm

Loaded DBs/Data Farm:
- Robot & Truck Availability & Performance Characteristics DB
- Metamodels (includes input characteristics to output probabilities) by Generic Segment and Leg DB
- Mission Segments, Constraints and Coordinates Template DB
- Upload Package DB
- Weather & Police Intel per Geo Region DB
- Geographic Route, Legs and related Obstacles DB

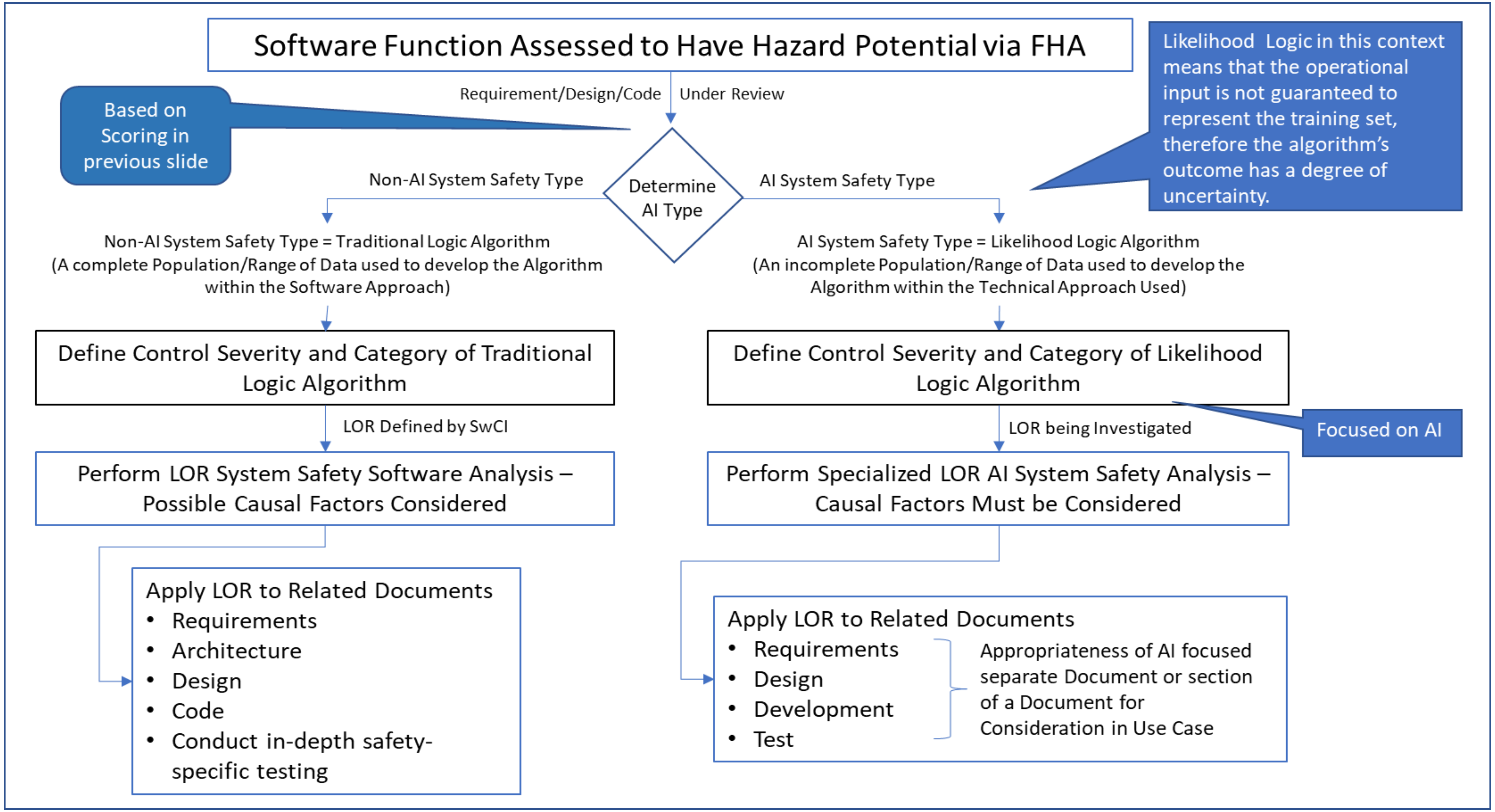
AI Type Definition

AI Type (Working Definition): For system safety concerns, an AI Type of function means that an algorithm will be developed:

(1) from using data approximations to build its algorithm, e.g. from simulations and synthetic data vs an equation that accurately represents real world physics, and/or
(2) when data samples used to build its algorithm is a subset of the actual population size, e.g., training data samples from population to support machine learning, training data samples requiring clutter backgrounds.

AI Type Examples of Specific Algorithms	Algorithm built based on using data approximations	Algorithm built based on using data samples from larger population	Final Score
CNN	x (if synthetic data used for CNN)	x (training data samples)	0 to 2
DNN + SL	x (if synthetic data used for DNN)	x (training data samples)	0 to 2
DNN + RL	x (if synthetic data used for DNN)	x (training data samples)	0 to 2
RNN (LSTM)	x (if synthetic data used for RNN)	x (training data samples)	0 to 2
RNN (Simple)	x (if synthetic data used for RNN)	x (training data samples)	0 to 2
Naïve Bayes	x (if modeling and sim data used to produce statistics for Naïve Bayes)	x (if RL used during opponent interaction to train algorithm)	0 to 2

Flow to Assess AI Type using Special FHA and SSHA Rigor



FHA Example for Mission Planner from Sandbox

Each Column represents a recommendation

Recommendations 1, 2 and 3

Haz ID #	Phase	Activity	State/ Mode	Function	Functional Failure	Hazard Title	Hazard Description	AI Type Scoring	AI Type Justification	AI Type Autonomous Justification	System Item(s)	Causal Factor Description
	Test & Deployment	Metamodel Selection	Setup	Mission Planner - Route/Robot Selection (Naïve Bayes Metamodel Selection)	Function Unavailable	No hazard since no metamodel will be selected		2				
					Function Malfunctions (degraded, partial, or unexpected results of the function, signal too small/strong/intermittent)	Wrong Metamodel selected	If a wrong metamodel is selected, then the wrong robots could be selected resulting in packages not being delivered or being too early or late and so delivered to wrong receipient, or the package could be lost. (Assumption: Package not delivered means package will be lost - Catastrophic hazard)		Input to function has infinite number of combinations - no traditional approximation is sufficient	Semi-Autonomous -- Man in Loop for approval of route	Mission Planner	Training Data is incorrect or incomplete, corner case occurence
					Function out of sequence/In another combination (Same thread/separate thread)	No hazard - if metamodel is not selected at the right sequence then route selection will not progress.						
					Function at incorrect time (too early, too late, outside defined window, function never ends)	No hazard since function is only relevant at a certain sequence, not time based.						

Scoring Grade of 2

Continued

Mishap	Effects	Existing Mitigations	Software Control Category (SCC)	Criticality Index	Software Criticality Index (SwCI)	Recommended Mitigations	Component(s)	Follow-On Actions	Comments
Delivering the package to the wrong receipient could be catastrophic since the material is hazardous and/or very valuable.	Personnel Injury / Equipment Loss	1. Operator review of this function output is required before system operations can progress 2. Test Data is used to ensure ML is comprehensive	2- Semi-Autonomous	1	SwCI 1	1. Implement threshold on propotionality used for selection e.g. if propotionality of selected item is not significantly different from next choice (10x), then declare fault or select default metamodel. 2. Implement redundant, independent non-ML functions that review route/robot selection for compatibility	Business Rule Manager Data Manager Database Farm Graphics User Interface (GUI)	Implement recommended mitigations if possible	User review of this function's output is required before the system can progress operation. The user must approve of the route/robot selection, including compatability between robot and route. With a trained user, detailed procedures, and appropriate time windows for review and approval, the user increases the SCC from 1 (Autonomous) to 2 (Semi-Autonomous), though the lack of other redundant interlocks does not further increase the SCC. Implementing of recommended mitigations would increase the SCC.

SSHA for Meta-Model Selection Algorithm within Mission Planner from Sandbox

Haz ID#	Phase	State/ Mode	System	Subsystem	Component	Element	Hazard Title	Hazard Description	Causal Factor Description	Mishap	Effects	Existing Mitigations
Identifier used to reference specific hazard.	*The life-cycle phase for which the risk and risk assessment apply. Multiple phases can be specified if the risk & mitigations are equivalent.*	*The State and/or Mode of the system for the hazard of concern.*	*The composite at any level of complexity of interworking parts (personnel, procedures, equipment, hardware, software, et al) used together to perform a task or accomplish a mission.*	*A functional or physical portion of a system designed, used or integrated to accomplish one aspect of the system task or mission.*	*A functional or physical portion of a subsystem designed, used or integrated to accomplish one aspect of the subsystem task or objective.*	*A functional or physical portion of a component designed, used or integrated to accomplish one aspect of the component.*	*Short title of the hazard*	*The detailed description of the conditions under which hazardous energy may be released in an uncontrolled or inadvertent way.*	*The detailed description of the failures, conditions, or events that contribute either directly or indirectly to the existence of the hazard.*	*The event or series of events where hazardous energy release could negatively effect equipment, personnel or environment; accident*	*The results of the mishap to include injury or death, damage to equipment and property, or damage to the environment.*	*Controls that are already planne existing to mitigate the risk.*
SSHA-001	Test & Deployment		Mission Planner				AI Function for metamodel selection (Naïve Bayes) failure	If a wrong metamodel is selected, then the wrong robots could be selected resulting in packages not being delivered or being too early or late and so delivered to wrong receipient, or the package could be lost. (Assumption: Package not delivered means package will be lost - Catastrophic hazard)	Inadequate quality or quantitiy of training data.	Delivering the package to the wrong receipient could be catastrophic since the material is hazardous and/or very valuable.	Personnel Injury / Equipment Loss	
									Incorrect algorithm selection			
									Improper curation of data			Multiple sources (primary, secondary and tertiary sources) accommodate sources that fail/missing.
									Too much or too little data (Underfitting and Overfitting of model)			

NAWC Weapons Division NAVAL AIR WARFARE CENTER

NAVAIR

SSHA LOR Table Example Based on Data Flow Analysis of Meta-Model Selection within the Mission Planner

Level or Rigor (LOR) Activity	Phase	Focus	LOR Description	Primary Responsibility	Support Responsibility	Baseline	Software Criticality Index (SwCI) 4	3	2	1	Representative Artifacts Produced
			Would the corruption of API/MSG/SQL/Other affect data variations requiring additional training of the Target Algorithm?	AI/ML Algorithm Developer	Data Analytics Engineer						Data Analytics Report
			If so, will quality (composition/complexity/structure) of Training Data significantly increase? Explain specific to the API/MSG/SQL/Other.	AI/ML Algorithm Developer	Data Analytics Engineer						Data Analytics Report
			Will these variations be part of the analysis for selecting the "best" algorithm? Explain.	AI/ML Algorithm Developer	Data Analytics Engineer						Data Analytics Report
		API/MSG/SQL Interface	Because of this issue, will quantity (more instances) of Training Data significantly increase? Explain specific to the API/MSG/SQL/Other.	AI/ML Algorithm Developer	Data Analytics Engineer			R	R	R	Data Analytics Report
			Will creating/finding enough training data replicating the corruption be an issue? Explain.	AI/ML Algorithm Developer	Data Analytics Engineer						Data Analytics Report
ALG6: Data Flow Analysis for the Mission Planner	Algorithm Design, Algorithm Code and Test and Evaluation		Are you confident that any additional data created/found will adequately represent the effects associated with replicating the corruption? Explain.	AI/ML Algorithm Developer	Data Analytics Engineer						Data Analytics Report
			Based on Modality Table: Describe Data Source Precedent for Improving Success Rate (ranking of primary, secondary tertiary… n attributes) -- by addressing related question in the table.	AI/ML Algorithm Developer	Data Analytics Engineer						Training Data Curation Report
		ML Modality: During Deployment & Curation Congruency	Based on Modality Table: Describe how missing and sparse data issues are modeled -- by addressing related question in the table.	AI/ML Algorithm Developer	Data Analytics Engineer			R	R	R	Training Data Curation Report
			Based on Modality Table: Describe how the quality of Training Data Characterized -- by addressing related question in the table.	AI/ML Algorithm Developer	Data Analytics Engineer						Training Data Curation Report
			Based on Modality Table: Describe how the quantity of Training Data Characterized -- by addressing related question in the table.	AI/ML Algorithm Developer	Data Analytics Engineer						Training Data Curation Report

- PR: Prerequisite Requirement – Required regardless of LOR or required in order to assess and determine LOR
- AD: As directed by Customer/Contract
- R: Required for assigned LOR
- IV&V: Independent Verification and Validation
- N/A: Not Applicable for this program or LOR

Investigation Questions Based on Modality

Investigation Topic	(Modality 1) multiple data sources, where each source contains one or more attributes	(Modality 2) single data source containing multiple data attributes, e.g., CNN	(Modality 3) combination of multiple data streams, where each stream contains one or more attributes and from a single data stream containing multiple aggregated data attributes, e.g., Naïve Bayes aggregated with CNN
Describe Data Source Precedent for Improving Success Rate (ranking of primary, secondary tertiary… n attributes)	Which sensor, communication link or human input content elements take precedent over others for improving success rate when training the ML algorithm under normal to stressed operational conditions?	Which attributes within the single data source take precedent over others for improving success rate when training the ML algorithm under normal to stressed operational conditions?	What data source content is more significant with regard to normal to stressed operational conditions? When dealing with separate streams, which sensor, communication link or human input content elements take precedent for improving success rate when training the ML algorithm under normal to stressed operational conditions? When dealing with combined streams, which attributes within the single data source are identified as primary, secondary and tertiary regarding importance for ML algorithm to improve success rate under normal to stressed operational conditions?
Describe how missing and sparse data issues are modeled	How is sensor malfunction, message corruption and human input errors on the higher precedent attributes forcing lower level attribute mixes of training data to ensure algorithm can deal with "real" operational issues?	Corruption in parts of image, especially containing higher precedent attributes forcing secondary and tertiary attribute mixes of training data to ensure algorithm can deal with "real" operational issues.	Combinations on modalities 1 and 2 regarding training of algorithm to deal with "real" operational issues.
Describe how the quality of Training Data Characterized	What is the precedent list (from highest to lowest) of attributes being used for training.	Same as Modality 1 for this row.	Same as Modality 1 for this row.
Describe how the quantity of Training Data Characterized	How much more emphasis Is placed on quantify of training data variations that have higher precedent than lower?	Same as Modality 1 for this row.	Same as Modality 1 for this row.

References

Brose, C. (2020): The Kill Chain, Defending America in the Future of Hi-Tech Warfare, Hachette Books, New York.

National Defense Authorization Act (NDAA) Fiscal Year 2021. Section C Artificial Intelligence and Emerging Technology.

National Security Commission on Artificial Intelligence (NSCAI). (2021). Draft Final Report. Autonomous Weapon Systems and Risks Associated with AI-Enabled Warfare. Chapter 4. https://www.nscai.gov/wp-content/uploads/2021/01/NSCAI-Draft-Final-Report-1.19.21.pdf

Naval Sea Systems Command. (2008) Department of the Navy Weapon Systems Explosives Safety Review Board (NAVSEAINST 8020.6E)

National Institute of Standards and Technology (NIST). (2019). U.S. Leadership in AI: A Plan for Federal Engagement in Developing Technical Standards and Related Tools. https://www.nist.gov/system/files/documents/2019/08/10/ai_standards_fedengagement_plan_9aug2019.pdf

Joint Software Systems Safety Engineering Workgroup. (2017) Software System Safety Implementing Process and Tasks Supporting MIL-STD-882E (JS-SSA-IF Rev. A). Washington, D.C.: Pentagon.

Shodani, S., (2018) Opinion: A summary of Concrete Problems in AI Safety. https://futureoflife.org/2018/06/26/a-summary-of-concrete-problems-in-ai-safety/

Joint Services – Software Safety Authorities (JS-SSA). (2017). Software System Safety Implementation Process and Tasks Supporting MIL-STD-882E (JS-SSA-IG Rev. A).

Defense Standardization Program Office. (2012) System Safety (MIL-STD 882E). Washington, D.C.: Pentagon.

National Security Commission on Artificial Intelligence (NSCAI). (2019). Final Report. Establishing Justified Confidence in AI Systems. Chapter 7.

Radio Technical Commission for Aeronautics (RTCA). (2012). Software Considerations in Airborne Systems and Equipment Certification. (DO-178C). Washington D.C.: Federal Aviation Administration.

DoD Architecture Framework Version 2.02. DoD Deputy Chief Information Officer. https://dodcio.defense.gov/library/dod-architecture-framework/

Dam, S. (2006) DoD Architecture Framework A Guide to Applying System Engineering to Develop Integrated, Executable Architecture, SPEC.

Booch, G., (2017): Unified Modeling Language User Guide 2nd Edition. Addison-Wesley Professional.

Hastie, T., Tibshirani, R., Friedman, J. (2017) The Elements of Statistical Learning Data Mining, Inference and Prediction. Second Edition. Springer.