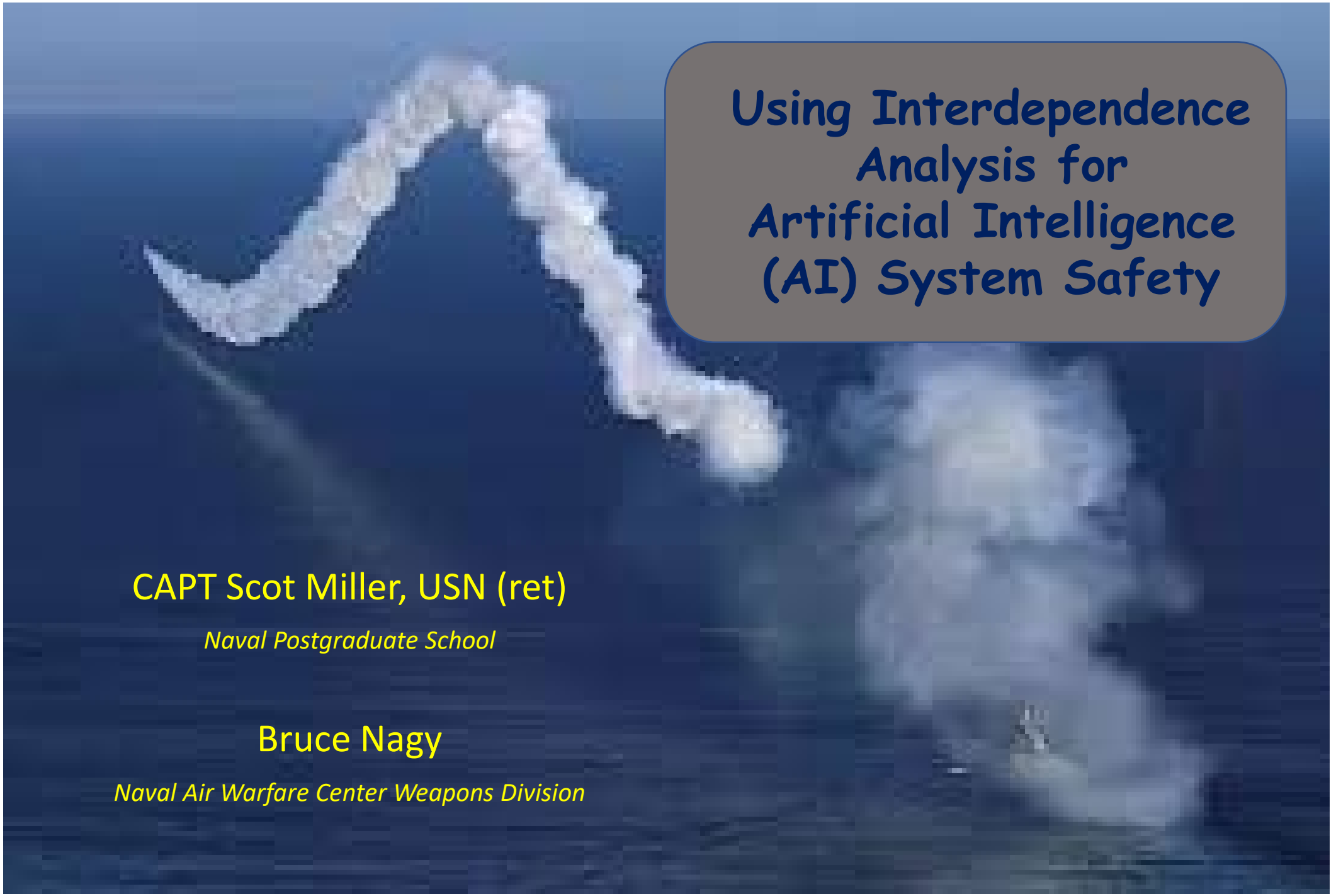


Using Interdependence Analysis for Artificial Intelligence (AI) System Safety

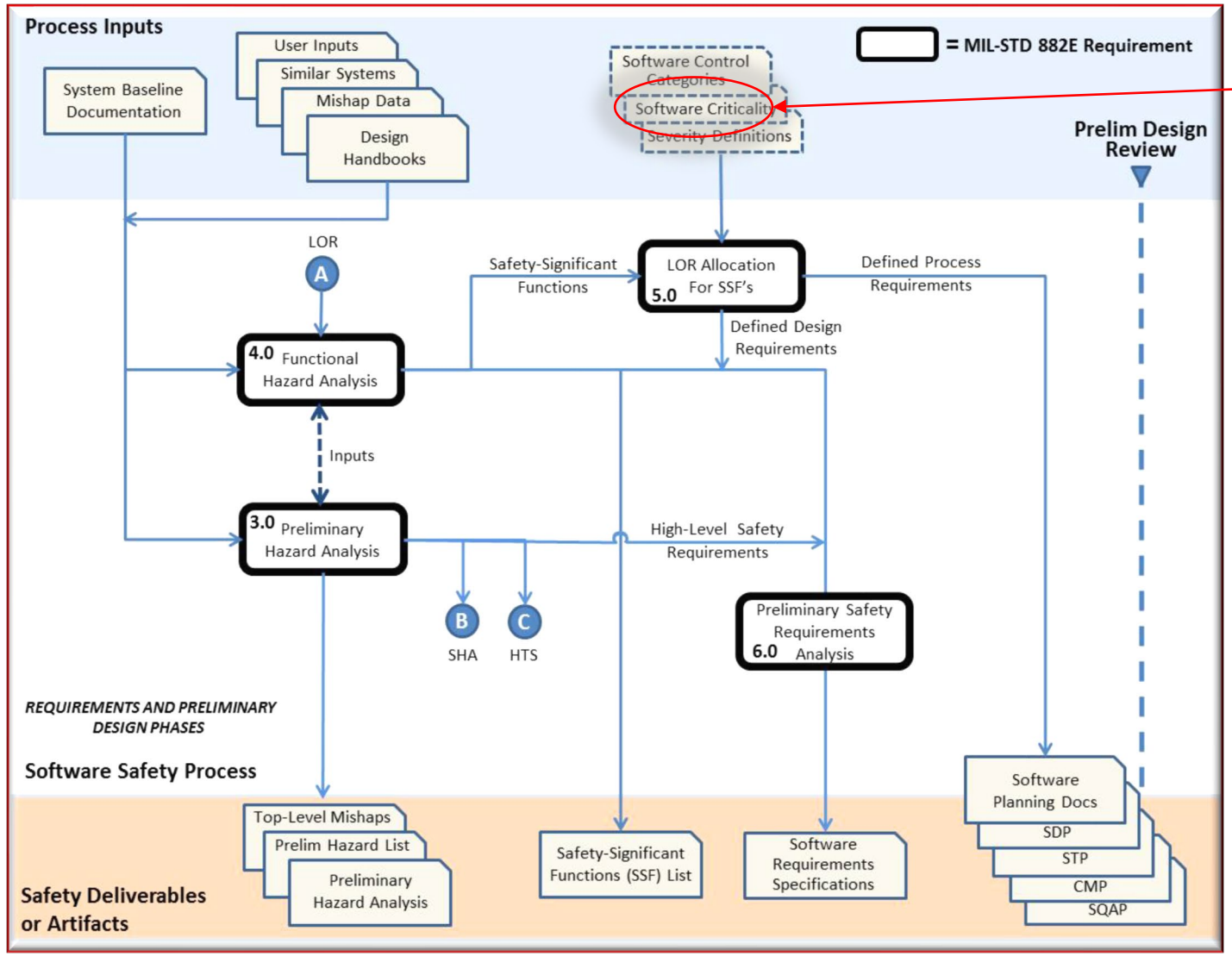
CAPT Scot Miller, USN (ret)

Naval Postgraduate School

Bruce Nagy

Naval Air Warfare Center Weapons Division

System Safety Processes



Problem

- Critical risk factor needs scrutiny
- Scrutiny involves level of rigor
- Level of rigor requires code review
- Reviewing AI/ML enabled functions via code review is not practicable

- AI enabled autonomous functions = critical functions with no human monitoring or intervention

❖ Therefore, avoiding code review requires lessening the autonomy of the AI enabled function

•TABLE I. Examples of AI System Failure Modes (Faria, 2017)

Examples of AI/ML Issues

Failure Category	Failure Mode Examples
System Produces Faulty/Poor Decision Recommendation	Biased outcomes/predictions
	Skewed outcomes/predictions
	Uncertain outcomes/predictions
Human Machine Operation Issues	Operators have lack of trust in the system
	Operators are overly trusting (overreliant) in the system
	Operators ignore the system
	Operators misunderstand the system recommendations/predictions
	Operators introduce errors into the system
System Under Attack (Cyber attack)	System is overtaken by adversary/adversary is controlling system
	System and its outcomes are corrupted by adversary
	Adversary jams or shuts down system
	Adversary gains access to system; decision information/knowledge is compromised

Faria, J. (2017, October 23-26). *Non-determinism and failure modes in machine learning.* Proceedings of IEEE 28th International Symposium on Software Reliability Engineering Workshops, 310-316.

Interdependence Analysis (IA)

Out of left field

IA is used to determine the interdependence relations between a robot and a human. Based on observability, predictability, and directability

Hypothesis: IA will determine potential AI enable critical functions, and offer resolution ideas

Interdependence Analysis Table

Tasks	Hierarchical Sub-tasks	Required Capacities	Team Member Role Alternatives							
			Alternative 1				Alternative 2			
			Performer	Supporting Team Members			Performer	Supporting Team Members		
			A	B	C	D	B	C	D	A
task	subtask	capacity								
task	subtask	capacity								
		capacity								
		capacity								
	subtask	capacity								
task	subtask	capacity								
		capacity								
	subtask	capacity								
		capacity								

Traditional hierarchical task analysis

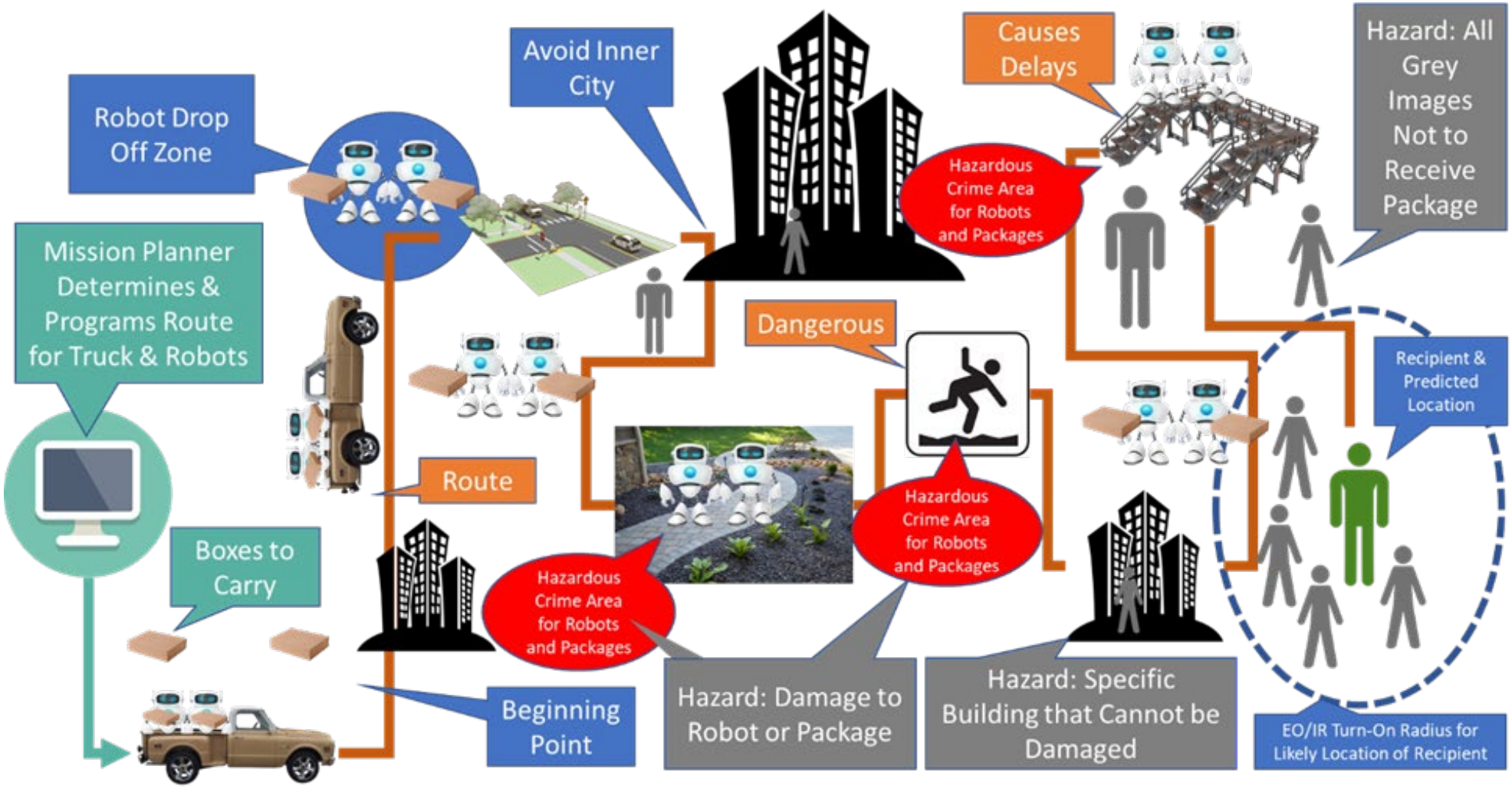
Identification of required capacties including situation awareness information, knowledge, skills, and abilities

Enumeration of viable team role alternatives

Assessment of capacity to perform and capacity to support, as well as identification of potential interdependence relationships in the joint activity

Observability, Predictability, and Directability (OPD) requirements derive from the role alternatives the designer chooses to support, their associated interdependence relationships, and the required capacities.

Scenario



Nagy, B.N. (2021, March 25). *Using event-verb-event (EVE) constructs to train algorithms to recommend a complex mix of tactical actions that can be statistically analyzed.* [On line conference presentation] Fifth Annual Naval Application of Machine Learning, San Diego, CA, United States.

Scenario Graphical User Interface

Nagy, B.N. (2021, March 25). *Using event-verb-event (EVE) constructs to train algorithms to recommend a complex mix of tactical actions that can be statistically analyzed.* [On line conference presentation] Fifth Annual Naval Application of Machine Learning, San Diego, CA, United States.

Interdependence Analysis

A. Tasks	B. Sub Tasks	C. Capacities	D. Performing Robot Component	E. Supporting Human	F. Observability, Predictability, and Directability Assessment WRT NOSSA Evaluations, plus Specific GUI functions from above
P-thru GUI	Map obstacles	1. Use leg route & obstacle DB	Data Loader Manager	User	OPD-thru GUI and MDP (1, 7)
		2. Use wx DB	Data Loader Manager	User	OPD-thru GUI
		3. Use police intel DB	Data Loader Manager	User	OPD-thru GUI
	Characterize legs	4. Use naive Bayes (nB) to determine best input attributes	nB, DB Farm, DB Manager	Evaluator	OPD-thru GUI Leverage statistical output part of GUI to verify inputs for attributes make sense. (2, 3, 6)
		5. User Random Forest (RF) to estimate probability and missing attributes	RF, DB Farm, DB Manager	Evaluator	OPD-thru GUI While RF is a black box to evaluators, in this case techniques exist to prove that the results are useful. Evaluators need to understand this proof and how to apply. (2, 3, 6)
	Select robot/route pairs	6. Apply temporal greedy search (TGS) to create robot /route candidates	TGS, Business Rule Manager, DB Farm, DB Manager	Evaluator	OPD-thru GUI While TGS is an algorithm, it is not ML, no special attention required (1, 2, 3, 6, 7)
		7. Use non-linear optimization (NLO) to determine combos that provide highest likelihood of mission success			OPD-thru GUI While NLO is an algorithm, it is not ML, no special attention required (1, 2, 3, 6, 7)
Unload robot in delivery Zone	Remove from truck	8. Activate robots	Processor, Power Regulator, and Power Supply	Truck Driver	OPD-thru GUI (1)
Robot navigation	Determine lead	9. Select robot as lead	Main Navigation and Guidance Controller	User	OPD-thru GUI (1)
	Navigate	10, Access planned waypoint DB	Main Navigation and Guidance Controller	User	OPD-thru GUI and MDP (1, 2, 7)
		Update status	Main Navigation and Guidance Controller	User	OPD-thru GUI and MDP (1, 2, 3, 5, 6, 7)
Delivery	Enter delivery zone	11. Compare up date to plan	Main Navigation and Guidance Controller	User	OPD-thru GUI and MDP (1, 2, 3, 5, 6, 7)
		12. Adjust location as necessary	Main Navigation and Guidance Controller	User	OPD-thru GUI and MDP (1, 3, 7)
	Identify customer	13. Use computer vision (CV) to identify customer	Image DB and CV	User, Recipient	OPD-thru GUI and MDP; CV is ML, so a human on the loop checking the identity as seen by the robot reduces "Autonomy". In other systems, this may not be feasible. May have to assume risk here. (1, 4)
		14. Check time so delivery can be synchronous	GPS Signal & SATCOM Transceiver, GPS Translator	User	OPD-thru GUI and MDP ((2,3,5)
		15. Deliver package	Robot arms	Recipient	(1, 4, 7)

Conclusions

- IA adds detail to those ML functional areas that need to be evaluated.
  - Not all designers appreciate the interdependence that should exist between user and the algorithm and therefore build no OPD connections. This makes reducing "autonomy" infinitely harder. Conducting IA rapidly speeds that discovery
- Adding the three main fault areas of ML into the IA raises very specific evaluation details and questions are raised.
  - While it may not solve the emerging ML evaluation conundrum, it does add considerable detail to the kinds of discussions that system developers and NOSSA ought to consider, especially when evaluators use the root cause details to inform their questions.
- Authors recommend adding a seventh column to IA table, suspect root causes and why, to the IA table
- Recommend NOSSA evaluators frequently review deployed system performance.
- This scenario benefits from a very capable GUI already informed by a knowledge of IA; NOSSA evaluators should not expect all systems will be as well developed.

Conclusions, continued

- Consider making an IA a requirement for submission of a system for NOSSA certification
- Withhold updating training data sets to the deployed edge at this point, since processes are not well understood
- Each updates to training data sets ought to be reexamined by NOSSA. Not a long term solution, though. Needs more research, tie to OVERMATCH
- Updating training data sets is similar to Navy's development security operations (DEVSECOPS) efforts to update patches to the Fleet in hours, not weeks. NOSSA should learn from those lessons learned; recognize training data set size makes over the air updates challenging and unreliable.
- Introducing ML techniques into systems may suggest changes to standard SE practices,
- New SE 'Vee' may change to be continuous for the entire lifecycle of a system. This means, in theory, that NOSSA has a continuous responsible to monitor system safety. That is a significant change, and worth thinking about. It may be that IA provides at least a way to wrap one's head around this potentially new responsibility. IA could be used to identify those functions that do require continuous evaluation.