Determining Safety Risks for Artificial Intelligence/ Machine Learning (AI/ML) Enabled Systems

Christopher Green

NDIA 26th Annual Systems & Mission Engineering Conference
October 16 -19, 2023
Norfolk, Virginia

E-mail: cgreen@alumni.vcu.edu

Outline

- Disclaimer
- Bottom Line Up Front (BLUF)
- Introduction
- Background
  - Artificial Intelligence (AI)
  - AI/ML Safety
- Suggestions
- Conclusions

Disclaimer

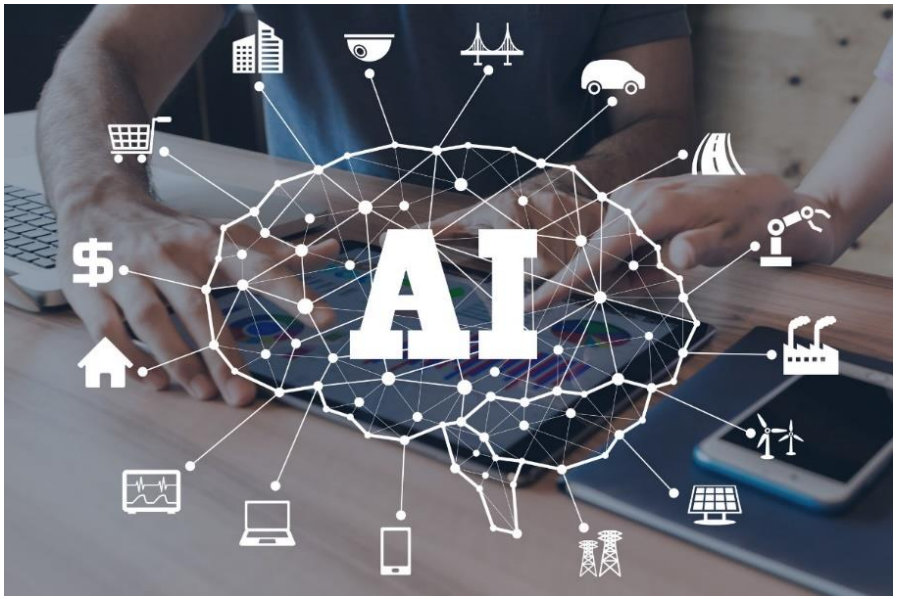
The views presented are those of the speaker and do not necessarily represent the views of the U.S. Department of Defense or its components.

BLUF

- The Department of Defense (DOD) is integrating Artificial Intelligence /Machine Learning (AI/ML) into systems at all stages of the Lifecycle Process
- AI Safety is an emerging area within the field of computer science with the development of self-driving cars and unmanned platforms
- Research in this area that is relevant to the DOD will be presented
- MIL-STD-882E is the Department of Defense Standard Practice for System Safety and currently does not contain guidance on AI/ML but is currently under revision
- The DOD is developing resources to aid system safety professionals in developing and executing safety programs on AI/ML-enabled systems

Introduction

- Advances in computational thinking and data science have led to a new era of artificial intelligence systems being engineered to adapt to complex situations and develop actionable knowledge.
- The increasing volume, velocity, variety, veracity, value, and variability of data are creating challenges in terms of development and implementation.
- For systems supporting critical decisions with higher consequences, **safety** has become an important concern.
- Methods are needed to avoid failure modes and ensure that only desired behavior is permitted.



Johnson, B.. Safety in AI-Enabled Warfare Decision Aids. NAML 2022. March 2022.

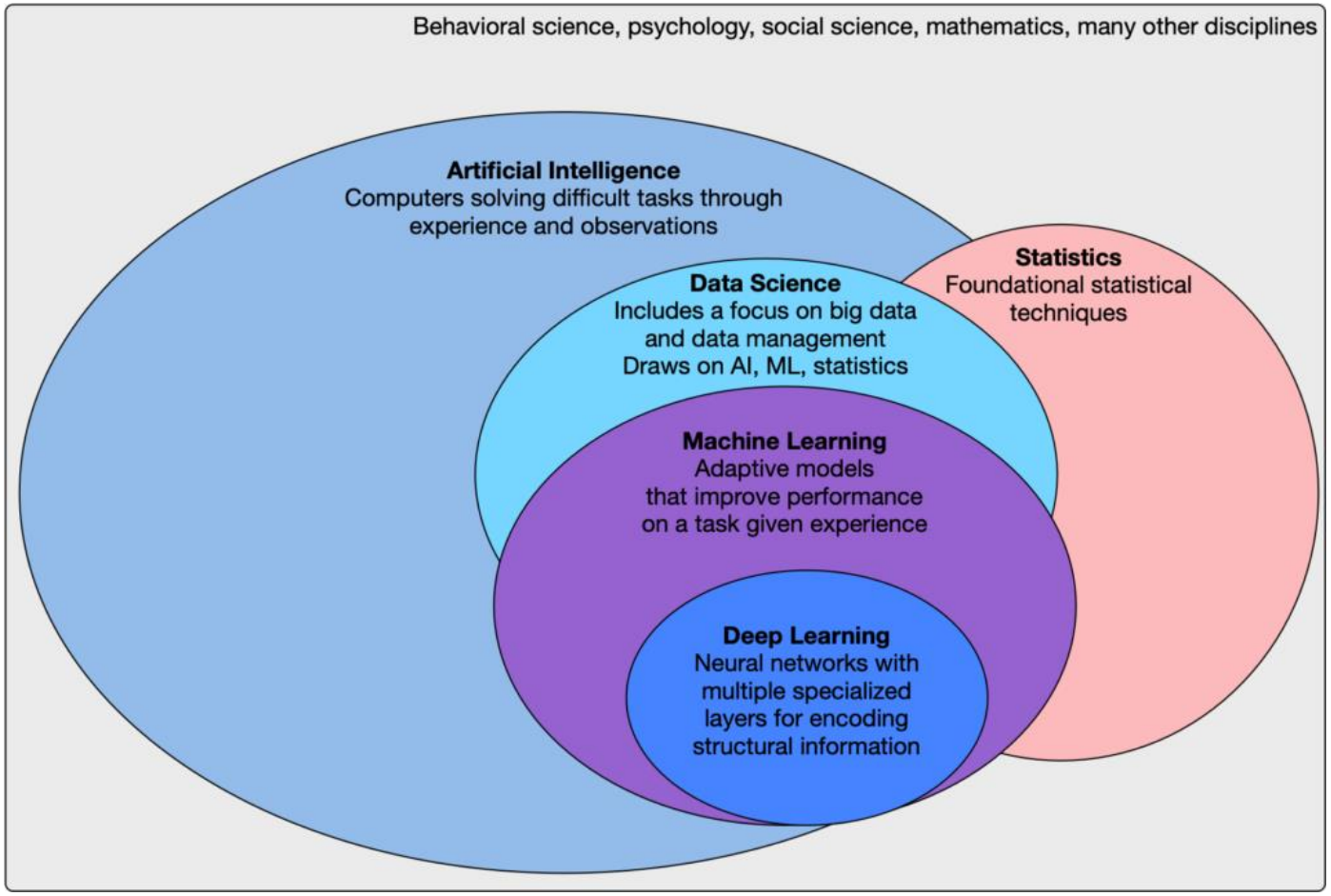
Background

Artificial Intelligence (AI)

Artificial Intelligence

The ability of machines to perform tasks that normally require human intelligence, such as recognizing patterns, learning from experience, drawing conclusions, and making predictions.

-JAIC AI Primer



ARTIFICIAL INTELLIGENCE FALL 2021. Amy McGovern. University of Oklahoma – Norman. https://ai-fall2021.ai2es.org/

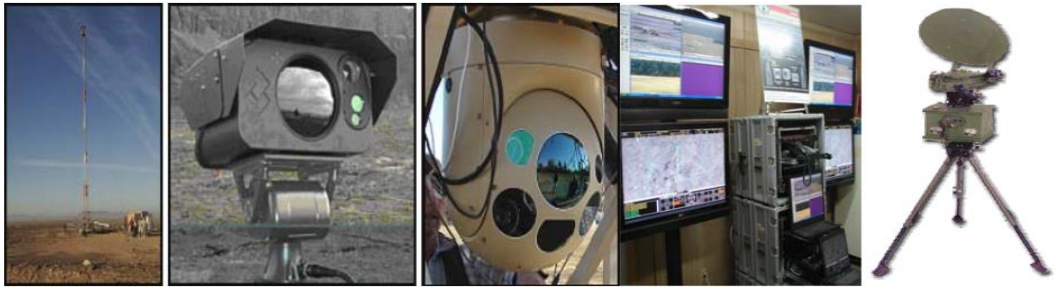
AI/ML Tasks

Inputs	Question	AI Tasks	Example Outputs
Text Data Image Data Video Data Audio Data	Is "it" present or absent?	Detection	Drone Detection
	What type of thing is "it"?	Classification	Type of Drone
	To what extent is it present? Where is "it"?	Segmentation	Drone Quantity and Size
	What is the likely outcome?	Prediction	Survivability Prediction
	What will likely satisfy the objective?	Recommendation	Engage or not to engage

Application Areas: Computer Vision; Human Language Technology; Robotics and Autonomous Systems

AI Enabled Military Systems

Example : G-BOSS



- Ground Based Operational Surveillance System (G-BOSS): Light, Medium, and Heavy Variant
- An expeditionary ground-based, integrated surveillance system, which employs the following: a multi-spectral Electro-Optic/Infrared sensor suite with multiple detection and assessment technologies in a self-contained, mobile platform.
- Used to observe, collect, detect, identify, classify, track, and report on contacts, objects of interest, and assessed threats 24-hours a day utilizing a video and sensor data display.
- Capable of video capture, storage, and transmission. The G-BOSS can also integrate signals from Unmanned Aerial Vehicles (UAVs) .
- Provides both local and regional commanders increased battle space awareness with the use of near-real time surveillance throughout their Area of Operation.
- Software utilizes Full Motion Video and Machine Learning algorithms to create models to:
  - Detect Personnel and Vehicles
  - Track Detected Objects Frame to Frame
  - Create Clusters of Detected Objects

Some AI Challenges

- Data
  - Hard to obtain
  - Determining how much data is needed
  - Classification
  - Security (Handling needs to be tracked)
  - Needs to be validated
- Systems Engineering
  - AI/ML systems learn and change during operation
  - Lack of Expertise
  - Major changes to SE are needed to “engineer” these types of systems.
    - Requirements
    - Acquisitions
    - Test and Evaluation
    - Logistics

AI Safety

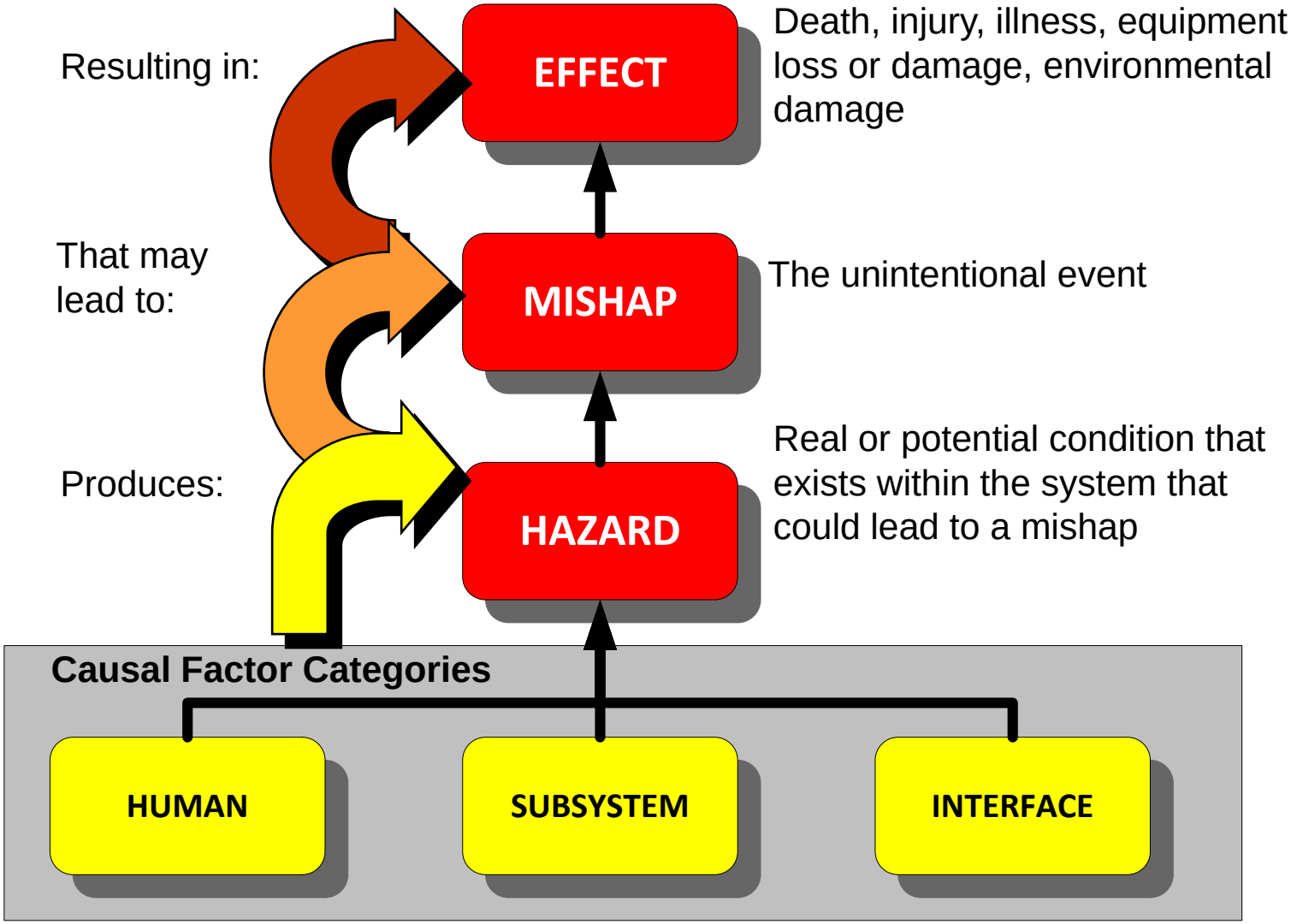
System Safety Definitions

- **System**: The organization of hardware, software, material, facilities, personnel, data, and services needed to perform a designated function within a stated environment with specified results
- **Safety**: Freedom from conditions that can cause death, injury, occupational illness, damage to or loss of equipment or property, or damage to the environment
- **System Safety Definition 1**: The application of engineering and management principles, criteria, and techniques to achieve acceptable risk within the constraints of operational effectiveness and suitability, time, and cost throughout all phases of the system life-cycle (MIL-STD-882E)
- **System Safety Definition 2**: An engineering discipline that employs specialized knowledge and skills in applying scientific and engineering principles, criteria, and techniques to identify hazards and then to eliminate the hazards or reduce the associated [mishap] risks when the hazards cannot be eliminated
- System Safety addresses hazards to personnel, equipment or environment during all lifecycle phases
- **Domains**: aviation, energy, medicine, military, etc.
- **Safety Critical Systems**: Systems where safety is the top priority.

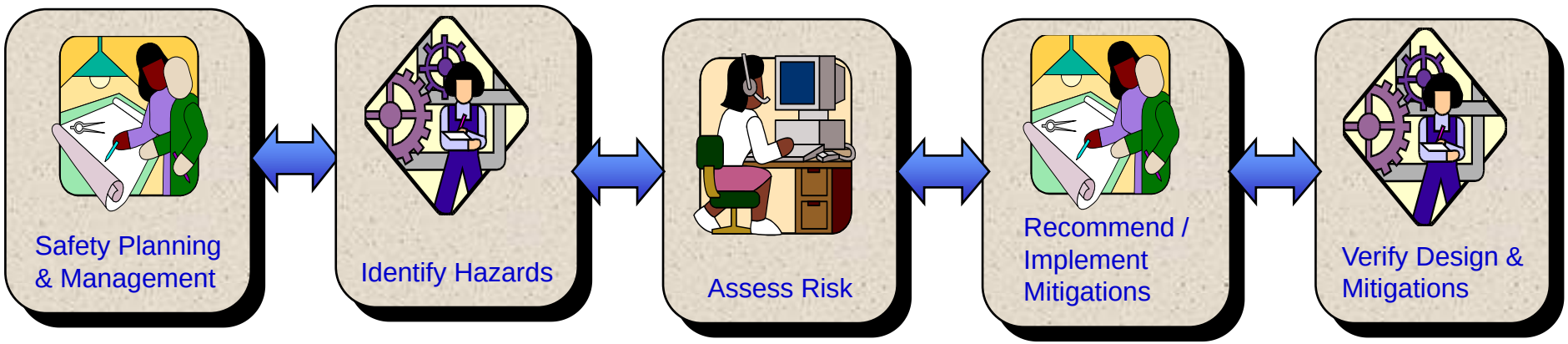
Standards

- MIL-STD-882E “Standard Practice for System Safety” – 11 May 2012
- Air Force System Safety Handbook – July 2000
- NASA System Safety Handbook
- ARP4761: Guidelines And Methods For Conducting The Safety Assessment Process On Civil Airborne Systems And Equipment
- ARP4754:Guidelines for Development of Civil Aircraft and Systems
- MIL-HDBK-516: Airworthiness Certification Criteria
- DoD Manual 5000.69: DoD Joint Services Weapon System Safety Review Process.
- Joint Software System Safety Engineering Handbook: 2010.
- AOP-52 (EDITION 1): NATO Guidance on Software Safety Design and Assessment of Munition-Related Computing Systems

Components of System Safety



System Safety Process: How it all comes together



- Plan: Plan to get system safety involved in a program as soon as possible
- Identify: Testing; Data; safety situations, scenarios, failures and conditions that may uncover, define, characterize or validate hazards
- Assess: Assess risk; Various standards available (MIL-STD-882E)
- Recommend/ Implement Mitigations: Get buy in from stakeholders
- Verify Design and Mitigations: Use standards such as MIL-STD-1472 and Test results

RISK ASSESSMENT MATRIX

SEVERITY / PROBABILITY	Catastrophic (1)	Critical (2)	Marginal (3)	Negligible (4)
Frequent (A)	High	High	Serious	Medium
Probable (B)	High	High	Serious	Medium
Occasional (C)	High	Serious	Medium	Low
Remote (D)	Serious	Medium	Medium	Low
Improbable (E)	Medium	Medium	Medium	Low
Eliminated (F)	Eliminated			

SEVERITY CATEGORIES

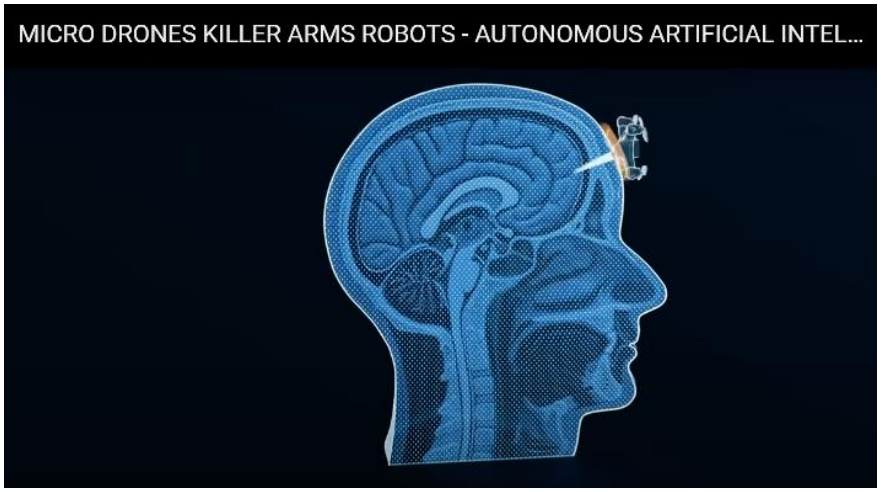
Description	Severity Category	Mishap Result Criteria
Catastrophic	1	Could result in one or more of the following: death, permanent total disability, irreversible significant environmental impact, or monetary loss equal to or exceeding $10M.
Critical	2	Could result in one or more of the following: permanent partial disability,injuries or occupational illness that may result in hospitalization of at least three personnel, reversible significant environmental impact, or monetary loss equal to or exceeding $1M but less than $10M.
Marginal	3	Could result in one or more of the following: injury or occupational illness resulting in one or more lost work day(s), reversible moderate environmental impact, or monetary loss equal to or exceeding $100K but less than $1M.
Negligible	4	Could result in one or more of the following: injury or occupational illness not resulting in a lost work day, minimal environmental impact, or monetary loss less than $100K.

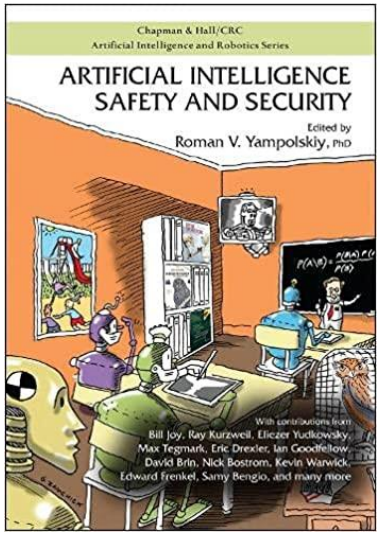
PROBABILITY LEVELS

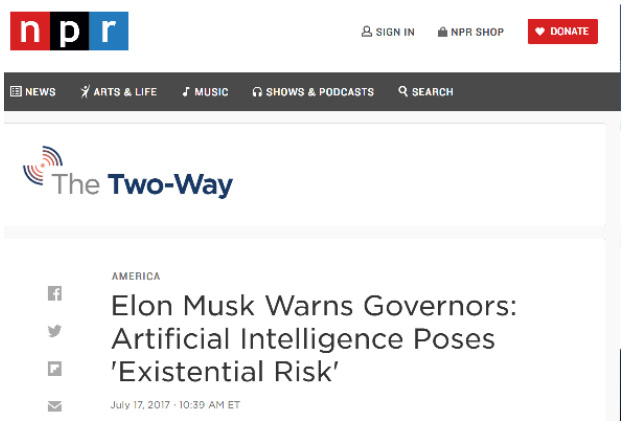
Description	Level	Specific Individual Item	Fleet or Inventory
Frequent	A	Likely to occur often in the life of an item.	Continuously experienced.
Probable	B	Will occur several times in the life of an item.	Will occur frequently.
Occasional	C	Likely to occur sometime in the life of an item.	Will occur several times.
Remote	D	Unlikely, but possible to occur in the life of an item.	Unlikely, but can reasonably be expected to occur.
Improbable	E	So unlikely, it can be assumed occurrence may not be experienced in the life of an item.	Unlikely to occur, but possible.
Eliminated	F	Incapable of occurence. This level is used when potential hazards are identified and later eliminated.	Incapable of occurence. This level is used when potential hazards are identified and later eliminated.

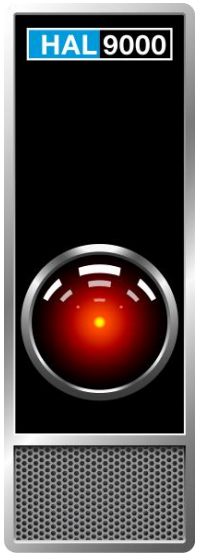
AI Safety

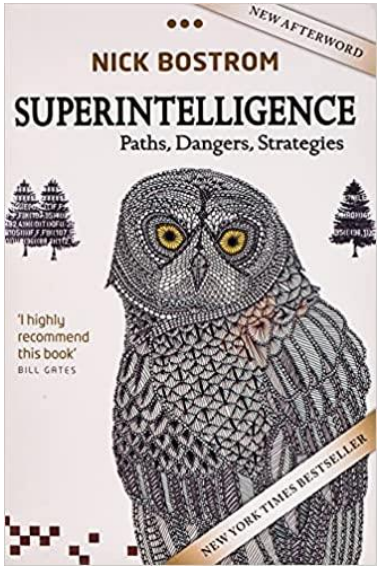
- "Artificial Intelligence Safety Engineering" (AI Safety) first coined in 2010
- "Artificial Intelligence Safety and Security" by Roman Yampolskiy
- Emerged in computer science with research on autonomous vehicles
- Young and underfunded outside of industry











Relevant Research

Roman V. Yampolskiy. Computer Engineering and Computer Science. University of Louisville. roman.yampolskiy@louisville.edu

- Artificial Intelligence Safety and Cybersecurity: a Timeline of AI Failures. 2015.
- Unexplainability and Incomprehensibility of Artificial Intelligence. 2019.
- Understanding and Avoiding AI Failures: A Practical Guide. 2021
- Unpredictability of AI. 2019.

Dario Amodei, Google Brain. Chris Olah, Google Brain. Jacob Steinhardt, Stanford University. Paul Christiano, UC Berkeley. John Schulman, OpenAI. Dan Mane. Google Brain.

- Concrete Problems in AI Safety. arXiv:1606.06565v2 [cs.AI] 25 Jul 2016

Pedro A. Ortega, Vishal Maini, and the DeepMind safety team

- "Building Safe Artificial Intelligence: Specification, Robustness, and Assurance," Medium, September 27, 2018, https://medium.com/@deepmindsafetyresearch/buildingsafe-artificial-intelligence-52f5f75058f1.

Dr. Bonnie Johnson. Naval Postgraduate School. Systems Engineering bwjohnson@nps.edu

- Safety in AI-Enabled Warfare Decision Aids. NAML 2022. Naval Applications of Machine Learning. March 2022
- Artificial Intelligence Systems: Unique Challenges for Defense Applications. 2021 Acquisition Research Symposium Pre-symposium Webinar: Developing Artificial Intelligence in Defense Programs. 3 March 2021

Relevant Research (cont)

Nancy Leveson. Massachusetts Institute of Technology. Professor of Aeronautics and Astronautics and also Professor of Engineering Systems.

- System Safety and Artificial Intelligence* Roel I.J. Dobbe1. arXiv:2202.09292v1 [eess.SY] 18 Feb 2022. leveson@mit.edu

Bruce Nagy. NAWCWD, China Lake, Systems Engineering Department, Systems Safety. bruce.m.nagy.civ@us.navy.mil

- Level of Rigor Tasks for AI Development. NAWCWD 03/08/2022. Sponsor: Naval Ordnance Safety and Security Activity (NOSSA)
- Using Event-Verb-Event (EVE) Constructs to Train Algorithms to Recommend a Complex Mix of Tactical Actions that can be Statistically Analyzed. NAML 2021.
- Applying Generative Adversarial Network constructs to Mission-based Simulations to produce “Realistic” Synthetic Training Data for Machine Learning Algorithms. NAML 2021

Christopher Green. NSWCDD. Dahlgren, VA. christopher.w.green.civ@us.navy.mil or cwgreen@alumni.vcu.edu

- Designing Safety into AI Enabled Systems. Sixth Annual Workshop on Naval Applications of Machine Learning (NAML 2022). Mar 23, 2022

Causal Factors

Pre-Deployment: Design, Development, Testing

- Bias in the training data sets
- Incompleteness---data sets don't represent all scenarios
- Rare examples – data sets don't include unusual scenarios
- Corruption in the training data sets
- Mis-labeled data
- Mis-associated data
- Poor validation methods
- Poor data collection methods
- Underfitting in the model – model cannot capture the structure of the data
- Cost function algorithm errors – model is optimized to the wrong cost function
- Wrong algorithm – training data is fit to the wrong algorithmic approach (regression neural network, etc.)

Post-Deployment: Operations & Sustainment

- Uncertainty/error in operational datasets Corruption in operational datasets
- Inaccuracy in the algorithm model (prediction error)
- Operational complexity that overwhelms the AI system
- Overfitting – tracks the data to closely thus failing to generalize
- Lack of explainability
- Trust issues
- Operator-induced error
- Adversarial attacks – hacking, deception, inserting false data, controlling automated systems

Johnson, B.. Safety in AI-Enabled Warfare Decision Aids. NAML 2022. March 2022.

AI System Modes

- Operation
  - Manual
  - Semi-autonomous
  - Fully Autonomous
- Failure
  - Bad decision made in fully autonomous mode
  - Compromised by an adversary (Cybersecurity)
  - Wrong Predictions
  - Uncertain predictions
  - Biased Outcomes
  - Skewed Outcomes
  - Operators lose trust
  - Operators overly trust
  - Operators ignore
- Mishaps
  - Crash
  - Malfunction
  - Explosion
- Effects
  - Damage to System
  - Damage to other systems/ infrastructure
  - Damage to the Environment
  - Injury/ Death to Personnel

Safety Use Case

Use Case: a fictional robot whose job is to clean up messes in an office using common cleaning tools.

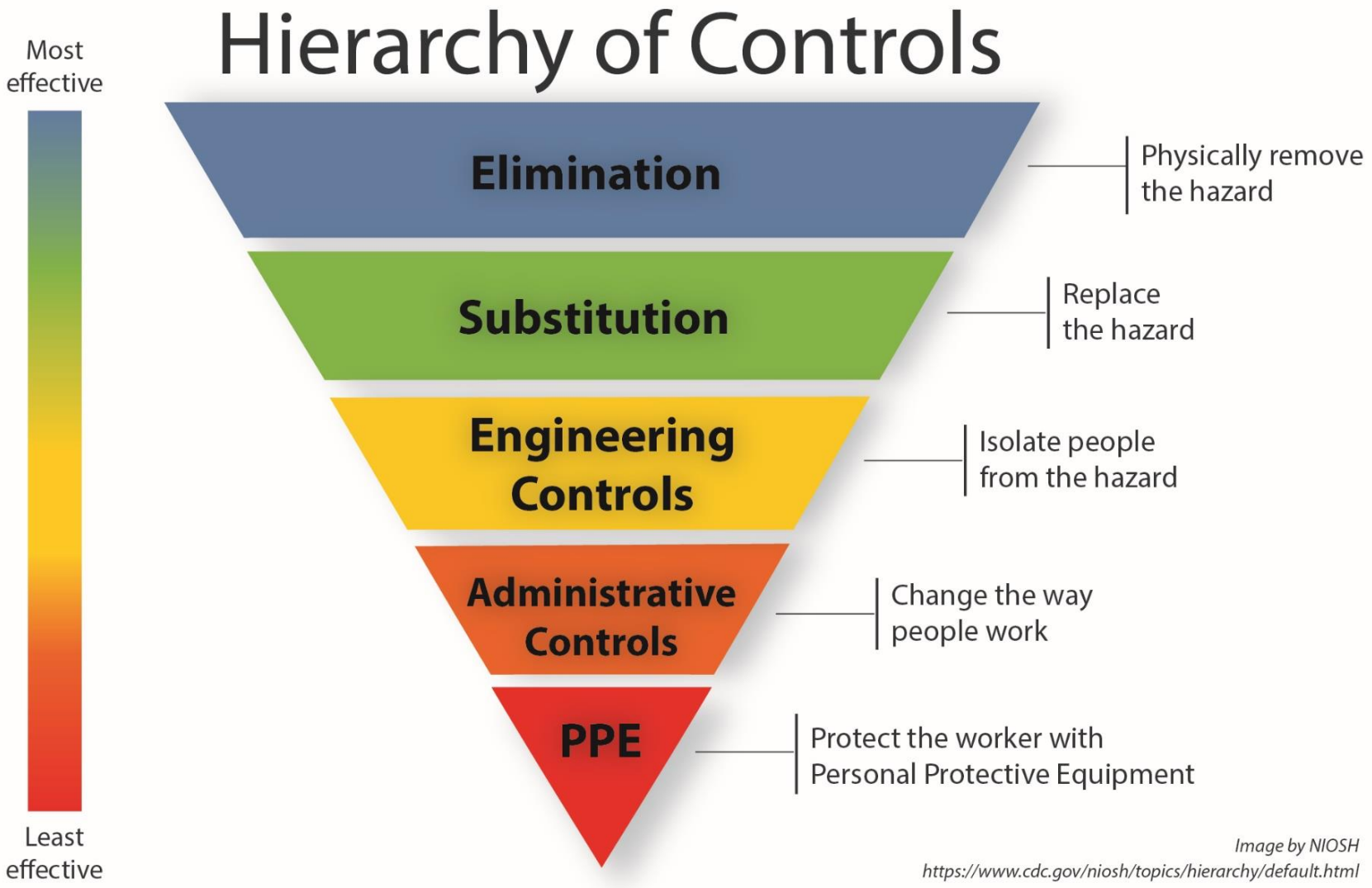
Design Mitigations for Possible Failure Modes

- Negative Side Effects: How can we ensure that our cleaning robot will not disturb the environment in negative ways while pursuing its goals
- Reward Hacking: How can we ensure that the cleaning robot won't game its reward function?
- Scalable Oversight: How can we efficiently ensure that the cleaning robot respects aspects of the objective that are too expensive to be frequently evaluated during training?
- Safe Exploration: How do we ensure that the cleaning robot doesn't make exploratory moves with very bad repercussions?
- Robustness to Distributional Shift: How do we ensure that the cleaning robot recognizes and behaves robustly when in an environment different from its training environment?

Yampolskiy, R.. "Artificial Intelligence Safety and Security".Chapman and Hall/CRC; 1st edition (August 23, 2018)

Other Issues Related to Safety

- Privacy: How can we ensure privacy when applying machine learning to sensitive data sources such as medical data?
- Fairness: How can we make sure ML systems don't discriminate?
- Security: What can a malicious adversary do to an ML system?
- Abuse: How do we prevent the misuse of ML systems to attack or harm people?
- Transparency: How can we understand what complicated ML systems are doing?
- Policy: How do we predict and respond to the economic and social consequences of ML?



Controls for the AI/ML Enabled System (B. Johnson)

1. Inherently Safe Design
    - Focus: ensuring robustness against uncertainty in the training data sets
    - Interpretability – ensuring designers understand the complex AI and ML systems that are produced
    - Causality – reducing uncertainty by eliminating non-causal variables from the model
2. Safety Reserves
    - Focus: achieving safety through additive reserves, safety factors, and safety margins
    - Validating training data sets – eliminating uncertainty in the data sets; ensuring data sets are accurate, representative, sufficient, bias-free, etc.
    - Increasing/improving model training process – ensuring adequate time and resources are provided
3. Safe Fail
    - Focus: system remains safe when it fails in its intended operation
    - Human operation intervention – the operation of AI systems should allow for adequate human-machine interaction to allow for system overrides and manual operation
    - Metacognition – the AI system can be designed to recognize uncertainty in predicted outcomes or possible failure modes and then alert operators
4. Procedural Safeguards
    - Focus: measures beyond ones designed into the system; measures that occur during operations
    - Audits, training, posted warnings, ongoing evaluation

Leveson's Lessons for AI/ML System Safety

	Leveson Lesson	AI System Safety Implication	Example System Safety Strategy
1	Component reliability is insufficient for safety	Identify and eliminate hazards at system level	System hazard-informed system design and safety control structure
2	Causal event models cannot capture system complexity	Understand safety through socio-technical constraints	System-theoretic accident models: integrating safety constraints, the process model and the safety control structure
3	Probabilistic methods don't provide safety guarantees	Capture safety conditions and requirements in a system-theoretic way	Process model: AI system goals, actions, observation and model of controlled process and automation
4	Operator error is a product of the environment	Align mental models across design, operation and affected stakeholders	Leveson's design principles for shared human-AI controller de-sign: redundancy, incremental control and error tolerance
5	Reliable software is not necessarily safe	Include (AI) software and its organizational dependencies in hazard analysis	System-theoretic process analysis
6	Systems migrate to states of higher risk	Ensure operational safety	Feedback mechanisms (audits, investigations and reporting systems)
7	Blame is the enemy of safety	Build an organization and culture that is open to understanding and Learning	Just Culture

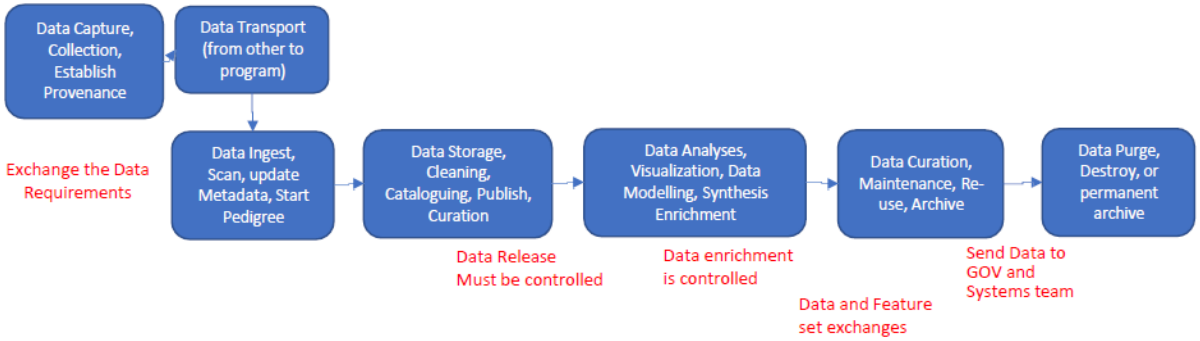
System Safety and Artificial Intelligence* Roel I.J. Dobbe1. arXiv:2202.09292v1 [eess.SY] 18 Feb 2022.

AI/ML System Level of Rigor for AI Development (Nagy)

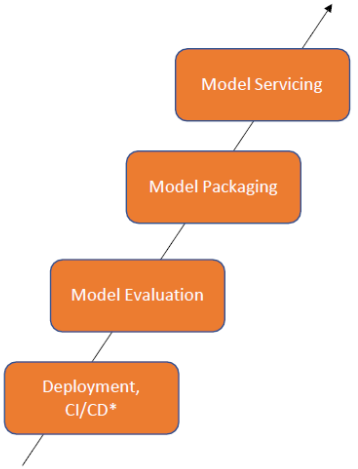
- Use Case
  - AI Systems/Products Under Development: (1) Two autonomous delivery robots (AI system), and (2) intelligent route planners (AI system).
  - Operational Scenario/Products Purpose: Delivering a package.
  - Environmental Requirements: Robots must be able to perform under a pre-determined set of requirements
- Detailed guidelines for the acquisition and development of systems incorporating Artificial Intelligence (AI) functions.
- The guidelines allow the user to create varying degrees of confidence in the behavior of the AI function during the challenges of operational deployment.
- The degree of confidence determines which of the fourteen Level of Rigor (LOR) tasks are being applied across five stages: (1) requirements, (2) architecture, (3) algorithm design, (4) algorithm code, and (5) testing stages.
- Each LOR task provides questions and/or considerations that allow developers to objectively evaluate the safety and reliability of the AI/ML function.

OUSD(R&E) DAU AI in SSE Course

Data and Metadata Life Cycle

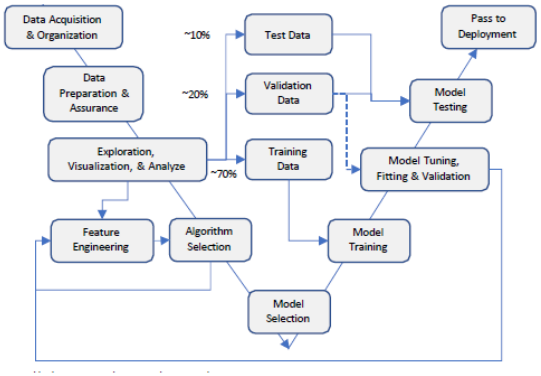


AI Deployment to SE Life Cycle



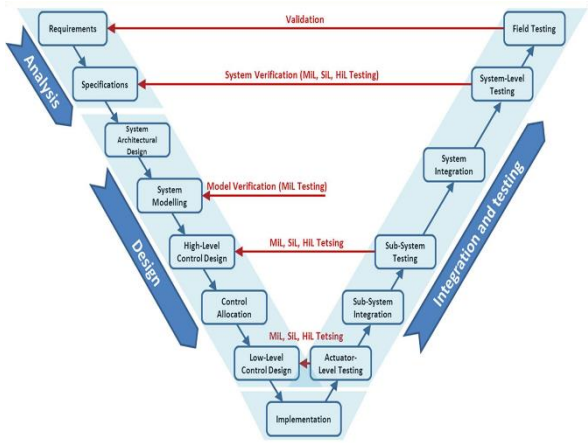
AI Development Life Cycle

Left side of this V is: Data and Model Selection



Right side of this V is: Model Validation & Testing

T&E V diagram with AI integrated



Safety Team Responsibilities

NOT EXHAUSTIVE

- Implements and manages the System Safety Program per MIL-STD-882 (Series)
- Keeps the Program Manager informed of the status of the System Safety Program.
- Coordinates on all system safety contractual matters.
- Identifying hazards and associated risks
- Establishing system safety requirements
- Mitigating or controlling hazards utilizing the safety order of precedence per MIL-STD-882 (Series)
- Manages the hazard-tracking system.
- Chairing the system safety working group or integrated product team meetings
- Serving as a member of the configuration control board.
- Prepare System Safety Documentation
- Incorporating MIL-STD-882 (Series) in the list of contractual compliance documents.
- Developing safety design precepts.
- Executing WSESRB reviews.
- Reviewing and approving risk assessments and hazard closures.
- Evaluating contractors' proposed system safety program.
- Monitoring contractors' system safety program.

The Potential Effects of AI on the SS Team

- 23 August 2023
  - "OUSD(R&E) Artificial Intelligence (AI) in System Safety Engineering Workforce Development" (DeLuca and Vega)
  - "Proposing the Use of Hazard Analysis for Machine Learning Data Sets in Collaboration with the Data Science Team" (Carter, Chan, Vinegar, Rupert)
- Impacts on Hazard Analyses
  - Preliminary Hazard Analysis (PHA): Early interaction with the data set
  - Requirements Hazard Analysis (RHA) and system-level Functional Hazard Analysis (FHA) with AI Level of Rigor (LOR) are necessary to justify data and metadata acquisitions
  - System Hazard Analysis (SHA) and Subsystem Hazard Analysis (SSHA) for integration of test cases and data of models or inference
  - Human-Machine Analysis and Operating & Support Hazard Analysis (O&SHA) address test data, simulators, training, and risk
  - New Analyses based on the need for Data Assurance
    - Data Hazard Assessment; Data FMEA; Data Assurance Assessment; Data V&V
- New or Modified SS Tasks
  - SS Workforce Lack of Understanding of AI/ML

MIL-STD-882

- MIL-STD-882E is the Department of Defense Standard Practice for System Safety and currently does not contain guidance on AI/ML
- MIL-STD-882F
  - Revision was drafted in January 2021 by Headquarters Air Force Material Command Air Force, Wright Patterson AFB
  - Cover letter: “Since MIL-STD-882E was last published in 2012, technologies associated with software have exponentially increased. For years, software safety hazard analyses has been based on the premise of deterministic software. That is no longer the case. Machine learning and Artificial Intelligence (AI) have also been increasingly prevalent in systems, yet MIL-STD-882E guidance is woefully lacking to address these topics.”
  - AI level of Rigor needs development
    - Artificial Intelligence Criticality Matrix Category was still under development
    - Artificial Intelligence Criticality Index (AICI) was still under development
  - AI/ML included in the Software Safety Assurance Process
  - Some contractor AI safety tasks were defined
- (23 Aug 2023) Upcoming updates may include: Enhanced FHA; LOR Task for AI; Safety Data Management Plan; ML SEE Handbook; Modified HA tasks that include AI

Conclusions

- AI has huge potential for many diverse applications (data products, cyber-physical, decision sciences)
- AI Enabled systems are being developed, and new processes have to be created, or existing processes have to be modified to accommodate
- The Department of Defense (DOD) is integrating AI/ML into systems at all stages of the Lifecycle Process
- The non-deterministic nature of AI will present new causal factors, failure modes, consequences, hazards, risks, etc., in the safety of AI-enabled systems
- AI Safety is an emerging area in research, and research relevant to the DOD was presented
- The DOD is developing resources to aid system safety professionals in developing and executing safety programs on AI/ML-enabled systems
- MIL-STD-882E does not contain guidance on AI/ML but is currently under revision

Questions???