Responsible Artificial Intelligence Awareness Course

This material is based upon work funded and supported by the Department of Defense under Contract No. FA8702-15-D-0002 with Carnegie Mellon University for the operation of the Software Engineering Institute, a federally funded research and development center.

DM23-0912

1. Course Description

This course introduces the concept of responsible artificial intelligence (RAI), the DoD Ethical Principles for AI, and provides an overview of best practices that are required to successfully implement RAI systems. Learners will have an understanding how AI enabled systems can accelerate and amplify mission success when designed and implemented correctly. This course enables learners to identify inherent hazards and risks in AI enabled systems and to understand mitigation best practices.

Learning objectives

By the end of the course, learners will be able to:

- Understand how AI enabled systems are different from traditional software systems.
- Understand why AI enabled systems create more risk than traditional software.
- Explain how responsible AI systems can be a force multiplier.
- Be familiar with policies and guidance impacting AI enabled systems.
- Recognize hazards and risks of AI systems and explain mitigation strategies.
- Understand the importance of designing AI to support human capabilities.
- Discuss considerations when operationalizing AI enabled systems.

2. Prerequisites

None

3. Course Delivery Options

This course is expected to be delivered in a remote asynchronous format, in approximately 90 minutes with pre-recorded lectures, slides, and videos.

4. Course Overview

This course consists of *5* units. Assessments should be conducted at the end of each unit and at the end of the course.

Identifier	Topic	Learning Objectives
Unit 1	**RAI Foundations**	
Module 1.1	AI and its applications	Identify potential use cases for AI/ML solutions to support missions. Recognize how AI/ML solutions are developed and architected across their lifecycle.
Module 1.2	Benefits and Risks of AI	Explain Responsible AI and DoD AI Ethical Principles.
Unit 2	**RAI Principles**	
Module 2.1	Understanding Ethics in AI Systems	Explain Responsible AI and DoD AI Ethical Principles.
Module 2.2	DoD Ethical Principles for AI	Explain the meaning and applicability of the DoD's and related Ethical AI principles.
Module 2.3	AI Regulation and Governance	Discuss laws, regulations, and policies related to AI/ML, data security, data privacy, and use of publicly procured data for government.
Unit 3	**RAI Considerations for Developing AI Systems**	
Module 3.1	Data for RAI	Discuss best practices in gathering requirements for AI and data applications. Discuss data risks. Discuss best practices in data labeling and data exploration. Discuss how relationships between data affect outcomes.
Module 3.2	Context for the End User	
Module 3.3	Selecting an AI System	Explain Responsible AI and DoD AI Ethical Principles. Explain the relationship between acquisitions and artificial intelligence (AI) solutions.

Unit 4	**AI in Operation**	
Module 4.1	Human-Machine Teaming	Discuss how to assess user interactions with AI solutions under different scenarios in order to identify design improvements.
Module 4.2	Interactions with an AI System	
Module 4.3	Mitigation of Harms	Describe AI and machine learning risk assessments.
Unit 5	**Final thoughts and course recap**	
	Course Summary	

5. Course Details

Unit 1: RAI Foundations

Unit 1 RAI Foundations places significant emphasis on Responsible AI (RAI) and its core terminology. Learners will gain a solid understanding of ethical considerations and responsible practices in AI development and deployment. Key concepts such as fairness, transparency, accountability, and privacy will be explored, highlighting their importance within the context of AI. Additionally, the module will delve into the unique benefits and risks associated with AI systems when compared to traditional software systems.

Module 1.1 – AI and its Applications

In Unit 1, learners will be introduced to AI, its applications, and to the concepts that make up Responsible AI (RAI). The focus of this module will be on the importance of RAI and its core terminology. Learners will explore definitions and examine the key terms associated with RAI, providing a solid foundation for understanding ethical considerations and responsible practices when developing and deploying AI systems. Learners will explore concepts such as fairness, transparency, accountability, and privacy, among others, and their significance in the context of AI.

Learners will examine real-world examples and delve into the reasons why the practice of responsible AI is crucial in today's rapidly evolving technological landscape. Learners will be asked to consider the potential implications of unethical AI practices and the importance of incorporating responsible approaches to mitigate risks.

Learning Objectives and Indicators

Identify potential use cases for AI/ML solutions to support missions.

- Understand what AI is (and is not) and how AI enabled systems are different from traditional software systems.
- Identify the general benefits and limitations of an AI capability.
- Explain how responsible AI systems can be a force multiplier.

Recognize how AI/ML solutions are developed and architected across their lifecycle.

- Understand and explain Responsible AI (RAI) core terminology.
- Define and recognize the importance of RAI.

Knowledge, Skills, Abilities, and Tasks

- (7048) Knowledge of the benefits and limitations of AI capabilities.
- (7021) Knowledge of emerging trends and future use cases of AI.
- (7065) Skill in explaining AI concepts and terminology.

Topics

Introduction to AI and potential benefits of incorporating AI into existing systems.

- What makes AI systems different from traditional software systems.
- Unique characteristics of AI systems (machine learning algorithms, data-driven decision-making, adaptability).

Introduction to Responsible AI (RAI) and its significance, key terms, and core concepts including:

- Fairness
- Transparency
- Interpretability
- Robustness

Importance of RAI in today's technological landscape.

- Highlight the need for ethical and accountable AI development and deployment.
- Consider AI systems potential for societal impact, the implications of unethical AI practices and ethical considerations of AI systems.

Examples [1]

Importance of Responsible AI Practices
AI-controlled drone swarms represent a promising military application, demonstrating the capability of these cooperative drone groups to communicate and operate synergistically, surpassing the effectiveness of individual drones. These swarms find utility in simulations, training operations, and actual missions, offering collective intelligence that empowers drones to autonomously navigate toward shared objectives. This underscores the crucial influence of Responsible AI (RAI) in refining military endeavors. Through the integration of RAI practices, the ethical foundation of these swarms is fortified. RAI-driven algorithms govern swarm behavior, fortified by human supervision to uphold strategic alignment, heighten situational awareness, and secure mission accomplishment. By emulating nature's swarm intelligence, drones emulate the transfer of critical data, much like insects communicate. In this landscape, the convergence of RAI and the emulation of natural processes accentuate AI's central role in enhancing military operations.

Consequences of unethical AI development and deployment
The outcome of unethical AI development and deployment could result in inadvertent harm and escalation. For instance, misinterpretation by AI-controlled drone swarms might cause erroneous decisions and actions, potentially leading to civilian casualties, infrastructure damage, or unintentional assaults on non-hostile targets.

Module 1.2 – Benefits and Risks of AI

In this module, learners will delve into the benefits and risks of AI systems and explore the characteristics that set AI systems apart from traditional software systems. AI systems have great potential and are already making great contributions to society, but as with any complex system, there are ways that information and the system itself could be used to harm individuals and organizations.

To identify benefits and risks for an AI system, the context that an organization will create and use the system in – the people, organizations, ecosystems, and overall society, that will interact with the system - need to be considered. Risk can come from multiple sources and need to be prevented or at least mitigated. The harms that individuals, groups, and organizations may encounter need to be identified during the early phases of a project.

Organizations making or acquiring AI enabled systems need to consider the full range of impacts, both positive and negative, that the system can have to mission, to warfighters and other end-users, the organization, and society. For example, due to the dynamic nature of AI systems, it may introduce additional risk to financial and business planning, to communication across the organization, to employee's well-being and employment, to customer relationships, to product development and many other potential impacts. The broader impacts of systems across larger groups also need to be considered such as social (for example, misinformation can affect voting and policy) and environmental (electrical consumption) considerations.

Learners will be introduced to key principles and considerations that guide the practice of Responsible AI. By understanding the value of RAI, learners will be able to identify AI systems that align with societal needs and uphold ethical standards.

Learning Objective and Indicators

Explain Responsible AI and DoD AI Ethical Principles.

- Give examples of the possible impacts of machine learning blind spots and edge cases.
- Understand potential benefits and risks of incorporating AI into existing systems.
- Understand why AI enabled systems create more risk than traditional software.
- Consider the full range of impacts AI systems can have on mission, warfighters at the edge, on end users, the organization, and society.

Knowledge, Skills, Abilities, and Tasks

(7051) Knowledge of the possible impacts of machine learning blind spots and edge cases.

(7041) Knowledge of remedies against unintended bias in AI solutions.

(7003) Knowledge of AI security risks, threats, and vulnerabilities and potential risk mitigation solutions.

Examples [1]

Potential Hazard of Neglecting RAI Practices
When implementing AI-driven combat simulations, neglecting Responsible AI practices can lead to unrealistic training scenarios and a failure to adequately prepare soldiers. AI-powered simulations might lack proper contextual understanding and could generate scenarios that deviate from real-world complexities, potentially leading to poorly trained soldiers who are ill-prepared for actual combat situations. Moreover, without ethical considerations and human input, AI simulations might inadvertently encourage behavior or tactics that could have negative consequences in real-world operations. Learners will recognize the challenges and complexities associated with working with AI systems.

Distinguishing Features and Implications of RAI Systems
Responsible AI systems for combat simulation would prioritize accuracy and contextual realism, reflecting the complexities of actual military operations. Ethical considerations and human expertise would guide the creation of scenarios that encourage appropriate decision-making and adherence to rules of engagement. Human oversight would ensure that the training focuses on necessary skills while avoiding harmful or unethical practices. By incorporating RAI practices, combat simulations can effectively prepare soldiers for a variety of scenarios.

Summary & Assessment

Throughout both modules, learners will develop a thorough comprehension of AI, its capacity for integration, and the significance of RAI. Practical examples will be used to demonstrate the importance of RAI and emphasize the unique characteristics of AI systems.

By the module's conclusion, learners will have the ability to differentiate AI systems from traditional software systems, empowering them to champion ethical and responsible AI practices while appreciating the nuances involved in working with AI technologies.

Outcomes

- Ability to differentiate AI systems from traditional software systems
- Comprehensive understanding of Responsible AI and its importance
- Understand potential benefits and risks of incorporating AI into existing systems.
- Understand why AI enabled systems create more risk than traditional software.
- Consider the full range of impacts AI systems can have on mission, warfighters at the edge, on end users, the organization, and society.

Unit 2: RAI Principles

In Unit 2 learners will gain an understanding of ethics as they relate to AI-enabled systems, the Department of Defense (DoD) Ethical Principles for AI, and the role of government and industry in AI regulation and governance.

Module 2.1 – Understanding Ethics in AI Systems

This module builds on the previous Unit to extend to the impacts of these dynamic systems. Learners will be exposed to the importance of integrating ethics when AI systems are being developed so that potential harms are identified and prevented and/or mitigation plans are created.

Learning Objective and Indicators

Explain Responsible AI and DoD AI Ethical Principles.

- Discuss best practices for implementing Responsible AI and DoD AI Ethical Principles while designing and developing AI solutions.
- Describe why it's important to have ethical AI systems.

Knowledge, Skills, Abilities, and Tasks

(5896) Maintain current knowledge of advancements in DoD AI Ethical Principles and Responsible AI.

Topics

- Team adoption of technology ethics can improve team communication and collaboration as well as consistency in implementation.
- Technology ethics supports harmonization of cultural variations across diverse teams.
- Technology ethics create an expectation for the team to consider and question the breadth of implications and provides explicit permission to do so.

Module 2.2 – DoD Ethical Principles for AI

The U.S. Department of Defense officially adopted a series of ethical principles for the use of Artificial Intelligence in February 2020, following recommendations provided to Secretary of Defense by the Defense Innovation Board [2].

Learning Objective and Indicators

Explain the meaning and applicability of the DoD's and related Ethical AI principles.

- Explain the meaning and relevant context of the DoD's Ethical AI principles to developers, end users, and senior leaders.

Knowledge, Skills, Abilities, and Tasks

(7020) Knowledge of DoD AI Ethical Principles (e.g., responsible, equitable, traceable, reliable, and governable).

(5896) Maintain current knowledge of advancements in DoD AI Ethical Principles and Responsible AI.

Responsible

Exercise appropriate levels of judgment and care, while remaining responsible for the development, deployment, and use of AI capabilities.

- End-to-end accountability of personnel
- "AI systems are tools, and they have no legal or moral agency that accords them rights, duties, privileges, claims, powers, immunities or status." – DIB

Example of Responsible [1]

When military forces are preparing for a critical operation, exercising prudent judgment and care is paramount for effective target recognition. Artificial Intelligence plays a pivotal role in bolstering the accuracy of pinpointing positions and potential threats within combat environments. By assimilating a wealth of information from various sources—ranging from reports to news articles—AI systems swiftly provide comprehensive insights into the operational theater. Leveraging its predictive abilities, AI can anticipate enemy behavior, evaluate vulnerabilities, and potentially consider environmental conditions. This responsible integration of AI acknowledges that ethical considerations are indispensable for guiding these forecasts. Ensuring unbiased outcomes necessitates meticulous handling of biases within the AI's training data. To maintain alignment with ethical standards and international regulations, human oversight remains a vital component. This harmonious fusion of responsible AI and target recognition not only bolsters operational efficacy but also underscores the technology's dedication to ethical principles and the safeguarding of societal welfare.

Equitable

Take deliberate steps to minimize unintended bias in AI capabilities.

- All systems have some form of bias. Complete objectivity is misleading.
- Bias can have purpose and can be helpful.
- The goal is to reduce unintended and/or harmful bias.

Example of Equitable [1]

Within the realm of military training, the need for thorough preparation without compromising safety has led to the utilization of combat simulation—an innovation powered by advanced AI capabilities. Within this digital domain, soldiers undergo rigorous training, maneuvering lifelike weapon replicas and confronting high-stakes decisions alongside their comrades.

Working behind the scenes, intentional measures are implemented to ensure fairness, precision, and the mitigation of unintended biases. Acknowledging the inherent presence of bias in all systems, AI's involvement extends beyond guiding evaluations and personalizing training; it also can act as a vigilant sentinel against these biases. Achieving this involves curating data to minimize predispositions, thereby fostering inclusivity. This fusion of AI augmentation and human feedback optimizes efficiency, cutting training time and resource costs.

Traceable

AI capabilities will be developed and deployed such that relevant personnel possess an appropriate understanding of the technology, development processes, and operational methods applicable to AI capabilities, including with transparent and auditable methodologies, data sources, and design procedure and documentation.

AI systems should clearly state capabilities and provide:

- Rationale and confidence in decisions and recommendations
- Evidence and resources (data provenance)
- Traceability and auditability

Example of Traceable [1]

Consider a combat scenario where artificial intelligence (AI) comes into play, significantly refining target recognition accuracy and amplifying systems' precision in pinpointing positions. This advancement equips defense forces with a profound operational understanding. AI rapidly scrutinizes reports, documents, and news, delivering insights at a pace surpassing human capacities. It foresees enemy behavior, identifies vulnerabilities, assesses mission strategies, and proposes effective mitigation plans, resulting in resource and time savings while conferring a tactical edge to soldiers. Responsible AI deployment is pivotal, necessitating transparent communication of capabilities, evidence, resources, as well as traceability and auditability encompassing data provenance. This commitment guarantees AI's alignment with ethical standards and operational requisites, cultivating a robust interplay between capability and responsibility.

Reliable

AI capabilities will have explicit, well-defined uses, and the safety, security, and effectiveness of such capabilities will be subject to testing and assurance within those defined uses across their entire life cycles.

AI systems should be able to provide warning signs for abuse/misuse and there should be policies for managing unintentional consequences.

Example of Reliable [1]

When AI solutions are used, the importance of having explicit, well-defined uses becomes evident in threat monitoring and situation awareness, which encompass operational approaches that acquire and analyze information to support diverse military activities. Unmanned systems, whether remotely operated or following predetermined routes, may rely on AI to assist defense personnel in vigilantly monitoring threats and enhancing their situational awareness. Drones integrated with AI further contribute to this capability. These drones effectively surveil border regions, identify potential threats, and promptly alert response teams. Moreover, they play a pivotal role in fortifying military bases' security and augmenting soldiers' safety during combat situations. It's imperative that AI systems incorporate mechanisms to detect warning signs of misuse/abuse, alongside well-defined policies for managing unintended consequences. Ensuring the reliability of

AI capabilities through explicit and well-defined purposes is paramount in these contexts.

Governable

Design and engineer AI capabilities to fulfill intended functions while possessing the ability to detect and avoid unintended consequences, and the ability to disengage or deactivate deployed systems that demonstrate unintended behavior.

For mitigation and communication plans it should be clear who can report and how; who is notified; how to disengage the system; what the consequences of system unavailability are; and backup and reverting processes.

Example of Governable [1]:

In highly secure military systems, the susceptibility to cyber-attacks remains a substantial concern, highlighting the vital role of AI. These attacks could jeopardize classified information and system integrity, endangering both personnel and missions. AI can act as a barrier, shielding programs, data, networks, and computers against unauthorized access. Its analytical adeptness detects and counters cyber-attack patterns through protective strategies. Ensuring AI's governable nature, effective development can equip it to identify and preempt unintended consequences and disengage or deactivate deployed systems that exhibit unintended behavior. Mitigation and communication plans are crucial components of governing AI systems, ensuring responsible and effective deployment to address evolving challenges while upholding safety and ethics.

Module 2.3 – AI Regulation and Governance

In this module learners will gain awareness of the role of the U.S. government in establishing policies and guidelines that impact AI systems. Learners will also gain an awareness of industry and professional standards, and country regulations that can have an influence AI development.

Learning Objective and Indicators

Discuss laws, regulations, and policies related to AI/ML, data security, data privacy, and use of publicly procured data for government.

- Describe strategies for mitigating AI security risks, threats, and vulnerabilities.
- Gain an awareness of the role of government and industry in establishing policies and guidelines that impact AI systems.
- Be familiar with policies and guidance impacting AI enabled systems.

Knowledge, Skills, Abilities, and Tasks

(7036) Knowledge of laws, regulations, and policies related to AI, data security/privacy, and use of publicly procured data for government.

Topics

The NIST AI Risk Management Framework (AI RMF) [3] was created in collaboration with the private and public sectors and is a framework to better manage risks to individuals, organizations, and society associated with artificial intelligence (AI). The AI RMF is intended for voluntary use and to improve the ability to incorporate trustworthiness considerations into the design, development, use, and evaluation of AI products, services, and systems [4]. Additionally, NIST developed the AI Risk Management Framework Playbook to provide suggested actions for achieving the outcomes laid out in the AI RMF.

The DoD RAI Strategy and Implementation Pathway was developed from the DoD RAI Memo and includes lines of effort by the six tenets to guide the implementation of RAI across the Department [5]: RAI Governance, Warfighter Trust, AI Product and Acquisition Lifecycle, Requirements Validation, Responsible Al Ecosystem, and Al Workforce.

The Defense Innovation Unit (DIU) Responsible AI Guidelines - aim to provide a clear, efficient process of inquiry for personnel involved in AI system development (e.g.: program managers, commercial vendors, or government partners) to achieve the following goals [6]:

- ensure that the DoD's Ethical Principles for AI are integrated into the planning, development, and deployment phases of the technical lifecycle.
- effectively examine, test, and validate that all programs and prototypes align with DoD's Ethical Principles for AI; and,
- leverage a process that is reliable, replicable, and scalable across a variety of programs.

Success Stories of Responsible AI

The DIU RAI Guidelines help companies think about and put the DoD Ethical Principles for AI and RAI considerations into practice.

Quantifind [7] leverages analytics derived from commercially and publicly available information to identify, track, and counter international criminals and their networks and were an exemplary case study in the DIU RAI Guidelines Report. Quantified commented that "the guidelines show promise for accelerating technology adoption as it helps identify and get ahead of potentially show-stopping issues."

Through the use of the DIU RAI Guidelines, the company identified [8], [9]

- Trade-offs that existed between performance gains and potential biases and performance irregularities
- The importance of continually measuring performance at both the individual model and end-to-end system levels
- The need to provide a mechanism for stakeholders to communicate about progress, standards, and known problem areas.

Fivecast's goal is to create AI and ML solutions to help organizations explore data to uncover actionable insights that are critical to intelligence missions and protecting global communities [10]. Fivecast has stated that their solutions implement and continuously improve upon ethical techniques and takes many considerations into account including how the solutions [11]:

- Will protect people and communities
- Enable analysts with the tools they need to work with AI tools
- Will be designed so they are reliable, safe and secure
- Will implement explainable and interpretable models
- Ensure accountability and contestability

Eightfold.ai [12] uses their AI platform to highlight talent that employers need to hire new employees, and allows candidates to match to jobs and to see why they are a good match [13]. Eightfold credits their responsible and ethical development and use of AI as foundational to their success [13].

Eightfold prioritizes four AI principles when designing and deploying their AI solutions:

- Fairness
- Transparency
- Safety and Reliability
- Active Monitoring and Response

Unit 3: RAI Considerations for Developing AI Systems

In this unit, learners will be introduced to topics that relate to best practices for developing responsible AI systems. These topics will provide learners with insight into the considerations that are required when designing and developing AI systems that are built for, and will be used by, humans.

Module 3.1 – Data for RAI

In this module, learners will discover the importance that data has on AI systems. This module will touch upon the impacts data can have, as well as provide insight into how to select, prepare, and analyze data responsibly.

Another important component when trying to achieve a successful AI system is the data that is used to train and test the models that are incorporated into the system.

Learning Objectives and Indicators

Discuss best practices in gathering requirements for AI and data applications.

- Explain the basic uses and challenges of test data in developing an AI solution.
- Understand how real-world environments and actions influence AI systems (data considerations)

Discuss data risks.

- Explain how to incorporate risk mitigation into a data plan.
- Outline the different data risks and vulnerabilities that should be mitigated, and how to recover from them if necessary.

Discuss best practices in data labeling and data exploration.

Discuss how relationships between data affect outcomes.

- Describe basic data relationship concepts (i.e. correlation, outliers).
- Discuss how transforming data can affect outputs.

Knowledge, Skills, Abilities, and Tasks

(7029) Knowledge of how to collect, store, and monitor data.

(7032) Knowledge of how to use data to tell a story.

(7036) Knowledge of laws, regulations, and policies related to AI, data security/privacy, and use of publicly procured data for government.

(1034C) Knowledge of Personal Health Information (PHI) data security standards.

(1034A) Knowledge of Personally Identifiable Information (PII) data security standards.

Using Data

Data are the basis for all AI systems, and as such care must be taken to identify data that will support the goal of the system and the end-users. Larger curated data sets are typically more robust and enable better AI systems performance.

Things to consider when trying to collect good data:

- Quality
- Machine readability
- Enough variation in the data for applicable use cases
- Appropriate data preparation

Risks with Data

There are many risks that can occur with an AI system if data are not appropriately prepared.

For example, datasets are inherently biased due to many different factors including, but not limited to:

- Decisions made when collecting data
- Selecting data (i.e., what is included in the dataset vs what isn't)
- How data are organized and curated
- When data is collected
- Decisions that are made by individuals that are captured and possibly reflected in the dataset

Even if the risks from the dataset itself are addressed, risks associated with data can still arise throughout the development of the system. It's possible that as the AI system learns, it will identify patterns found in the data that could lead to biased or unintended outcomes. Therefore, it is important to continually monitor how the data is influencing the AI system.

Preparing Data Responsibly

When preparing data for use in AI systems, it is important to prepare it responsibly. This means:

- having accountability during data collection
- thoroughly performing exploratory data analysis to understand what the data contains and what needs to be investigated further
- using the insights from the exploratory data analysis to perform comprehensive data cleaning to remove erroneous data
- labelling data properly
- splitting data into training, validation, and test sets in order to train, validate, and test the model for its future purpose in an AI system

Privacy

Privacy needs to be considered constantly as data is collected, stored, and monitored. There are laws, regulations, and policies that relate to data security and privacy and the use of publicly procured data for government, as well as standards

relating to personal health information (PHI) and personally identifiable information (PII) that learners should be aware of.

Federated Learning enables systems to collaboratively learn a shared prediction model while keeping all the training data on separate systems or devices, meaning there is no need to store training data in the cloud [14]. Federated Learning allows for smarter models, lower latency, and less power consumption, all while ensuring privacy [14] via incorporating the principle of data minimization. For a visual explanation of Federated Learning, please refer to the Federated Learning comic from Google AI (https://federated.withgoogle.com)

Google has been developing federated learning applications since 2017 and in 2022 Google announced a Spanish language next-word-prediction ML model using federated learning with a rigorous differential privacy guarantee [15].

Module 3.2 – Context for the End User

In this module, learners will explore the importance of context when interacting with AI systems. Learners will discover that context encapsulates different elements, from the purpose and goal(s) of the AI system to the people (warfighters) using the system.

Learning Indicators

- Determine what it means for an AI system to operate as intended and is meeting the needs of the mission.
- Understand how real-world environments and actions influence AI systems
- Recognize hazards and risks of AI systems and explain mitigation strategies

An AI system is made successful through several factors. These factors include understanding the environment the system will operate in, the purpose for which the system is intended, and the people who will be using the system

It is important to understand that well developed AI systems have been designed for specific contexts, purposes, and users. Therefore, AI systems should not be used interchangeably.

Context

As an example of the importance of context, think of a cell phone from 10 years ago that did not have the do not disturb or driving detection features that many cell phones have today. These older phones were "context-blind", and could exhibit socially inappropriate or even dangerous behavior, if they rang in an important business meeting or when driving on a busy highway respectively. However, in many of the cell phones available today, there is an option to turn on the do not disturb feature before entering a situation where you should not be distracted, and similarly some phones have a driving detection feature where calls and notifications are hidden until the driver indicates they are not driving. The addition of these features demonstrates the context-awareness capabilities of technology [16].

Purpose

It is important to understand the purpose of the AI system and if the purpose changes, to conduct research to confirm that the system is appropriate for a different purpose.

For example, review the metric of fairness for two different scenarios:

- A cancer detection system used by the VA may define fairness as parity of performance with hopes that the system is equally **accurate** across different groups (i.e., male, female, and those who do not wish to disclose).
- A hiring algorithm used by the Army may define fairness as parity of outcomes, since it is desirable to have a similar **rate** of employment between different groups (i.e., male, female, and those who do not wish to disclose).

Fairness is the metric considered in both scenarios. However, the context in which fairness is used, changes. In the VA scenario it wouldn't make sense to define fairness as a parity of outcomes because the cancer detection system needs to be as accurate as possible across the groups. The rate of detection between groups may be interesting, but it is not the priority for their definition of fairness. For the Army scenario, parity of performance might be interesting, but the fairness metric is about parity of outcomes.

People

Developing an AI systems involves many different types of people, and it is important to consider the different roles as they will impact the system throughout the lifecycle. The different people include:

- Capability developers – people making the AI-enabled system.
- End users - people who are directly interacting with the system, or the outputs of the system.
- Acquisitions – people who are deciding what system to buy, who will manage it and maintain it.
- Maintainers – people who may have developed the system and who are now responsible for monitoring and maintaining it.
- Others – people who are affected by the system and/or whose situations are being changed due to the system (for example their employment status, their financial well-being, etc.), as well as other stakeholders such as family, colleagues, and people in the community.

End Users

End users play a very important role as the people who will be interacting with the AI system. If end users' concerns are not taken into consideration or their needs are not met, the likelihood of those people choosing to use the AI system and finding success with it will be low.

For example, consider the case where US soldiers felt ill and more prone to harm after testing a headset that used mixed reality to aid them in certain missions [17]. During testing it was discovered that soldiers found the headsets heavy and that the headsets limited their head movements and decreased their visibility. Overall,

soldiers reported the headsets would fail to contribute to their ability to complete their mission. These findings mean that a very expensive initiative needs significant changes so that the system works with soldiers and to ensure that soldiers see the value in using the system on their missions.

Consideration of Context to Avoid or Mitigate Harm

It is also important to consider context so that certain harms can be avoided or mitigated against. Provide examples of risks and harms such as the following.

Categories of Harm from the 3Q-Do No Harm Framework [18]

- Financial: Negative impact on finances, property, or other resources
- Health: Negative impact on mental, emotional, or physical health
- Time: Inefficient or unproductive activities, processes, or systems
- Fairness/Equity: Perpetuating or facilitating prejudice, bias and/or unfairness
- Safety: Physical and/or emotional wellbeing compromised by fear, danger, or uncertainty
- Privacy: Lack of control over personal information
- Misinformation: The creation, spread and/or amplification of false or inaccurate information intended to deceive
- Control: Inability to freely direct information, activities, or systems
- Transparency: Lack of disclosure of information, activities, or systems

Additional potential risks and harms include:

- Abuse (physical, mental, social, economic) of individuals, groups, and the system itself.
- Discrimination against individuals, people within an organization, and those affected by the system.
- Disinformation influencing audiences through deception (like misinformation defined above)
- Democracy
- Human dignity

Module 3.3 – Selecting an AI System

In this module, learners will explore different considerations that go into selecting an AI system to use in operation. These considerations range from questions to ask when acquiring an AI system to questions regarding a systems' trustworthiness and development.

Learning Objectives and Indicators

Explain Responsible AI and DoD AI Ethical Principles.

- Discuss leading practices for ensuring Responsible AI principles are reflected in an organization's approach to AI acquisition, development, deployment, and monitoring.
- Understand the importance of user experience and how human and machine interaction is critical to mission success.

Explain the relationship between acquisitions and artificial intelligence (AI) solutions.

Knowledge, Skills, Abilities, and Tasks

(7020) Knowledge of DoD AI Ethical Principles (e.g., responsible, equitable, traceable, reliable, and governable).

(7051) Knowledge of the possible impacts of machine learning blind spots and edge cases.

(7041) Knowledge of remedies against unintended bias in AI solutions.

(5896) Maintain current knowledge of advancements in DoD AI Ethical Principles and Responsible AI.

Selection Considerations

Accountability and ownership in contracts – Define who is responsible for the acquired system. Define who, specifically, understands the implications of context and use of the system and will be held responsible for the ethical and responsible use of the AI system.

Data and models – Consider and define who is responsible for data and models used and what measures are in place to address concerns about accuracy, reliability, and safe handling of sensitive data and the models that utilize data.

Defining problems to determine AI appropriateness – Gather and consider information about primary end users, data, risk of application space, overall risk tolerance, and overall situation to determine if AI is an appropriate solution based on needs. Consider alternate solutions and compare potential outcomes.

Example questions to ask vendors about their "AI":

- What data is the system based on?
- Performance (e.g. accuracy, confidence, measure of success, etc.)?
- Benefits and limits of the system?
- Measure compliance of AI Tools with DoD Ethical Principles for AI?
- Updates/Maintenance of AI systems?

Example questions to Explore Trustworthiness of new AI Technologies [19]

- As new and innovative technologies become available, there will be many different experiences to consider. Before using an AI system in operation, various questions should be considered, including:
- What is the intended use of the AI product?
- How was the model trained?
- How do the AI product's characteristics align to the responsible AI dimensions of the use case and context that I am considering?
- What are limitations of the technology's functionality?

Example questions to understand the process to audit and verify the AI product performance:

- What are the product performance metrics?
- How is the product continuously monitored for failure and other risk conditions?
- What implicit biases are embedded in the technology?
- How are aspects of trustworthiness assessed? How frequently?
- Is there a way that I can have an expert retrain this tool to implement fairness policies?

The Defense Innovation Unit RAI Guidelines [6]

Additionally, when considering selecting a AI system to use in operation, the Defense Innovation Unit's (DIU) RAI Guidelines can be used as a reference to understand if the AI system was designed responsibly, as the guidelines pose questions that seek to understand the design processes behind the planning, development and deployment phases of an AI system [6].

For example:

- Questions posed in the planning phase provide insight into the work behind understanding the ownership of the data/models that are used in the system as well as the identification of potential harms and errors.
- Questions in the development phase consider manipulation of data models, audit mechanisms, and system performance monitoring.
- Questions in the deployment phase focus on three sets of continuous evaluation procedures that should be scoped and performed on an ongoing basis throughout the AI System's life cycle.

Unit 4: AI In Operation

In this unit, learners will begin to understand what RAI looks like in practice. The purpose of this unit is to expose learners to the considerations, decisions, and applications of AI when deployed in real-world contexts.

Module 4.1 – Human-Machine Teaming

In this module, learners will dive deeper into other considerations that are required when developing successful responsible AI systems. The purpose of this module is to introduce and set expectations for human-machine teaming, as well as discuss user experience and usability topics. These components are important because they impact what users think of, how they interact with, and their opinions of the usefulness and usability of an AI system that they may team with.

Learning Objective and Indicators

Discuss how to assess user interactions with AI solutions under different scenarios in order to identify design improvements.

- Describe basic user experience (UX) and user interface (UI) design principles.
- Understand the importance of designing AI to support human capabilities.

- Understand how the effectiveness, security, and robustness of an AI system affect the end user and mission and the related risks.
- Discuss considerations when operationalizing AI enabled systems.

Knowledge, Skills, Abilities, and Tasks

(7053) Knowledge of the user experience (e.g., decision-making, user design, and human-computer interaction) as it relates to AI systems.

(7047) Knowledge of the basics of customer experience, customer design, psychology of customer decision-making, and human-computer interaction.

(7027) Knowledge of how humans interact with and/or are impacted by AI solutions within the DoD context.

Human-machine teaming

Human-machine teaming is achieved when systems made with and for people, meet their needs and augment the person's efforts by partnering in effective support [20][21].

What does it mean to be a part of a human-machine team?

- "Human-machine teaming is a relationship—one made up of at least three equally important elements: the human, the machine, and the interactions and interdependencies between them [22]."
- When designing AI systems intended for human-machine teams, the systems should be designed through speculation of risks and benefits, value elements of transparency and usability, encourage human accountability, promote respect for human operators, and uphold security standards [23]

Trustworthiness

To trust an AI system and to find an AI system trustworthy are two different things. Trust is *a user's psychological state based on expectations of the system's behavior,* whereas finding an AI system trustworthy depends on how the *system demonstrates that it will fulfill its promise by providing evidence that it is dependable in the context of use* [19].

When a user finds a system trustworthy, it means that the person has awareness of the system's capabilities during its use, and they choose to use it. AI enabled systems must be trustworthy for human-machine teaming to be successful. Humans build trust between themselves and adjust their perception of another person's trustworthiness based on the situation. For example, one member of a human sports team may find a second team member more trustworthy to receive a long pass, than a third member.

AI systems cannot experience trust, but successful human-machine teams require trustworthy AI systems, that provide information to their human counterparts so that the human can maintain calibrated trust of the system. Calibrated trust is a psychological state of adjusted confidence that is aligned to end users' real-time perceptions of trustworthiness [19]. Users can develop a level of calibrated trust with a system by understanding the system's capabilities through its performance

in a specific context [20]. A user's calibrated trust toward an AI system that they team with can be impacted by a variety of different factors, including **User Experience** and **Usability.**

User Experience and Usability

As mentioned above, human-machine teaming is a relationship made up of three elements: the human user, the machine, and the interactions between them. These interactions are important because they play a vital role in establishing the user's calibrated trust of the machine. Making trustworthy systems requires an understanding of the primary users' needs and then designing interactions between the user and machine that are helpful and assist users in their workflows. This work is often referred to as user experience.

User Experience (UX) refers to research that provides a deep understanding of users, what they need, what they value, their abilities, and their limitations [24]. It also considers the business goals and objectives of the group managing the product [24] and includes the work to design interactions for users.

UX centers around creating and improving the quality of a user's interaction with a product, system, and/or physical experience. Successful UX efforts play a large role in ensuring users are comfortable with an AI system, that it meets their needs and expectations, and they are aware of its abilities in what in can and can't do.

These considerations are of great importance when developing AI systems for human-machine teams as they will have an impact on what the user thinks about the system, if they would want to use the system, and their level of calibrated trust with the system.

Usability is a quality attribute that assesses how usable and useful the interfaces are. It also refers to methods for improving ease-of-use during the design process [25]. Usability looks at the effectiveness, efficiency, and the overall satisfaction of the user. Therefore, usability also plays an important role with AI systems. If a user finds the system frustrating or is not able to understand what the system is doing, or why it is suggesting something, that user is unlikely to want to use the system and will not readily adopt it into their workflow. The same is true for AI systems that are used in human-machine teaming.

To improve the UX and usability of systems, research must be conducted to understand the user's needs and behaviors regarding their current work. Through analysis of that research, more informed design decisions can be made regarding how the system (including the AI portions) will be designed. UX and usability methods include interviews, observations, prototyping, usability studies to ascertain the systems' ease of use at various points during planning and development, analysis of system use and analytics, and many other methods.

Scenarios

Content developers should provide examples that tie back to the DoD Ethical Principles for AI (responsible, equitable, traceable, reliable, and governable).

Module 4.2 – Interactions with an AI System

In this module, learners will investigate some of the implications that may arise when interacting with an AI system, and what they should keep in mind when interacting with an AI system.

Learning Indicators

- Understand the importance of user experience and how human and machine interaction is critical to mission success.
- Understand how real-world environments and actions influence AI systems.
- Understand how AI risk is managed.

Human-in-the-Loop

Humans should remain "in-the-loop" for AI systems in operation. Having a human-in-the-loop when an AI system is in operation means that there is an interaction between human and machine before a decision is made.

The concept of human-in-the-loop emphasizes the active engagement of human experts who provide oversight, guidance, and intervention through the decision-making process to try and achieve accurate and responsible outcomes [26]

Human-in-the-loop systems also bring several benefits, including:

- Increased quality and accuracy [26]
- Lowering the number of errors [26]
- Increasing transparency and trustworthiness of the system [27]
- Improves safety and precision [28]

Setting Expectations for AI System Interaction

Introducing an AI system to an existing process does not mean changing the end users' approach to making decisions. The responsibility in decision-making lies with humans, not with an AI system.

While technology enhanced by AI can factor in many inputs and correlate different variables to identify patterns or make predictions, at the end of the day, if something goes awry based on a decision that was made with the system, questions will be asked of the operator. Therefore, when working in a human-machine team, it is vital that the people making the system understand what the system is doing, how it is producing the predictions or outputs, and what should be inferred from the information the system presents to make decisions.

AI systems may produce unexpected results, particularly if the same system is used for different tasks. There is no "one size fits all" AI system solution because while they are not programmed as typical software systems are, they are based on data that is specific to their purpose. Applying that information to a new situation may not provide the desired results.

Due to the dynamic nature of AI systems and the difference in data there are not plug-and-play solutions that can be implemented every time a new AI system is

needed. To be designed responsibly, AI systems are designed specifically for the context, users, and use case that the system is intended for.

Example of Environmental Differences

The conditions data are collected in directly correspond to the conditions the resulting AI system will be effective in. For example, depending on the time of year and weather, the sun's strength, and the color of light change. People wear different clothing in different weather, as well as depending on local culture and other influences. Photographs taken on a hot, sunny day will contain different information about that environment than those taken on a cool, rainy day – even in the same location.

For example, consider a computer vision system that will be used to identify military tanks. The data used to train this model includes images of military tanks (unobstructed) in desert-like environments (minimal grass, some bushes, and trees) with the sun shining and blue skies. Similar style media was used to test the model before releasing it for use.

Then upon release, the system is used on a day when the skies are dark grey, it is raining, and the wind is blowing. It is plausible to assume that the computer vision system trained on clearly visible military tank media, will not be successful identifying a tank in this situation. The system has no point of reference for obstructed military tanks (due to the rain and environmental movement), and will therefore likely output incorrect results.

Examples of Different Contexts

Just as was discussed in Unit 4, context is very important when using AI systems. If an AI has been developed for a specific context, then using that system in a different context may lead to incorrect or troublesome results. In more serious cases, if an AI system is used in a different context that it was intended for, it can lead to serious and harmful impacts for the people involved.

Data and Context Mismatch
An AI healthcare device that was trained using a military health data source that is then used in a community hospital. It is likely that the AI model will provide less accurate results about the female population it is asked to diagnose, since it was trained on a military health data source, and historically, there are more male service members than female [29].

Large Language Models
If a nine-year-old asks a Large Language Model (LLM) (like ChatGPT) to explain what a black hole is, ChatGPT will respond with a very educated and scientific answer, because it does not have the context that a nine-year old is asking the question and likely requires a more simplified response. In trying the question again, but this time providing the context that the question is being asked by a nine-year old, ChatGPT provides a much different response, using analogies that a child is more likely to relate to.

Activity (learners on their own):

- Ask an LLM like ChatGPT or BARD to explain what a black hole is and read the response.
- Next, change the questions by asking the LLM to explain what a black hole is to a nine-year old and read the response.
- What differences do you notice?

Fabricated Data
For example, consider the case when a lawyer asked ChatGPT to find existing cases that would strengthen their prosecution to move the case forward for their client [30]

However, it came as a surprise to the lawyer when they received correspondence from the judge indicating at upon further review, the cases that the lawyer had brought forward as precedent for the case, we're inaccessible, due to the fact they were fabricated [30].

In using ChatGPT in this way, the lawyer not only hurt their own creditability, but also the credibility and trustworthiness of their firm and the trustworthiness they had with their clients. Additionally, this also hurt the client the lawyer was representing as it caused their case to be called into question [30].

New Data Integration
Models that are used in AI systems go through training and testing to ensure the data that is being used produces the intended outputs, if unexpected data is input into the system, the system may respond incorrectly. In RAI systems, this concern should be mitigated throughout the design and development process, however in some cases, if these considerations are not considered, the system may behave unexpectedly with new input data.

For example, consider an autonomous driving system that is trained to stop when it detects a stop sign. Imagine the system was trained and tested in the western hemisphere (ex. the United States), using various pictures of stop signs (ex. octagonal shaped, red, big white lettering) in different weather conditions, lighting, angles, etc.

However, consider after the model is trained and tested it is used in the Eastern Hemisphere (ex. Japan) [31]. When the car detects a stop sign in Japan, it does not stop and continues driving. Why? Because even though the system was trained to detect stop signs, it likely learned to look for the shape, color, and lettering. However, in some countries, like Japan, a stop sign is a red, upside-down triangle.

Therefore, when the system received a new input, it did not respond in the way it should, due to contextual differences that it was not trained to detect.

Module 4.3 – Mitigation of Harms

In this module, learners will be provided with an overview of preventing harm as well as discover steps, techniques, and tools can be used to potentially mitigate harms that may arise due to introducing or using AI in a new context or environment.

Learning Objective and Indicators

Describe AI and machine learning risk assessments.

- Understand how real-world environments and actions influence AI systems.
- Understand how AI risk is managed.

Knowledge, Skills, Abilities, and Tasks

(7003) Knowledge of AI security risks, threats, and vulnerabilities and potential risk mitigation solutions.

(7040) Knowledge of Personal Health Information (PHI), Personally Identifiable Information (PII), and other data privacy and data reusability considerations for AI solutions.

Measuring Harm

Measuring the likelihood of harm is something that should be done regularly throughout the design and development of an AI system. Harm can be measured in a variety of different ways and should be thoroughly explored so that as many possibilities of harm can be identified and prevented or mitigated against while a system is being developed, rather than when users are interacting with the system. Harm can be looked at in different lenses, such as:

- Qualitative
- Quantitative
- Probability estimation based on indicator monitoring
- Harm risk analysis – (AML topic: Can we enter a harm state by providing plausible input?

Measuring the impact of a harm is also important to understand – those that affect people and impacts on society, organizations or reputation. Therefore, when thinking about harm, it is important to ideate not only on harm to users, but also how wide-reaching harms can be. Harm can manifest in different areas [32], such as:

- Societal
- Organizational
- Mission
- Safety
- Financial
- Reputational
- Can we predict harm?

Steps and Activities to Predict and Prevent Harm

Harm can be wide reaching and hard to envision, or understand the impact, in many cases. There are activities that people can use to ideate on potential harms as well as ways to prevent or mitigate the harms. In an effort to predict and or prevent harm, different activities can be used to get people thinking about possible harms that could emerge and various ways to prevent those harms:

- Pre-mortem (i.e., imaginative techniques to inspire scenario-planning) [33], [34]
- Abusability Testing [35]
- Designing Ethical AI Experiences Checklist [36]
- What Could Go Wrong? Card Game [37]
- Data and Model Cards [38], [39]
- Implementing Mitigation Frameworks [40]

Integrating Mitigation Steps into Practice

- Complete regular checks with the system to ensure harm is not occurring
- Ensure diversity in design
- Work with cross-functional teams to develop different perspectives

Encourage technical and non-technical people work together to better understand the inconsistencies with the system and to ideate on creative and innovative solutions to address any errors or irregularities.

Have or increase the amount of end user involvement throughout the life cycle.

It is not just the technical needs that should be taken into account throughout the life cycle, but many human elements should also be considered. This helps to ensure certain risks pertaining to users do not manifest because they weren't factored in throughout the design process.

Unit 5: Final Thoughts and Course Recap

This course is designed to offer a thorough exploration of Responsible AI (RAI), addressing various aspects across its four distinct modules. Starting with the foundational Unit 1, participants will gain a comprehensive understanding of how AI can be applied in different contexts, exploring both its potential benefits and the inherent risks it poses. Unit 2 delves into the realm of DoD ethical principles that underpin AI development, shedding light on the ethical considerations and dilemmas that emerge in this evolving field. Unit 3 takes a deeper dive into the complex process of developing Responsible AI systems. This unit emphasizes the significance of incorporating ethical awareness and bias mitigation during AI system development. Participants will explore strategies and methodologies to ensure that AI systems are crafted responsibly, minimizing unintended consequences and potential biases. The final module, Unit 4, offers a practical perspective by examining how Responsible AI principles are put into operation. This unit showcases real-world applications of AI across diverse domains, demonstrating how ethical considerations and Responsible AI practices can be seamlessly integrated into various operational scenarios.

By engaging with these four units, students will not only acquire a comprehensive knowledge of Responsible AI but also develop the skills and insights needed to navigate the ever-changing landscape of AI applications with an ethical and responsible mindset.

Unit 1: RAI Foundations- Understanding how AI can be applied and its benefits and risks

Throughout this unit, learners have comprehensively analyzed the unique characteristics and challenges of AI, focusing on machine learning algorithms, data-driven decision-making, and adaptive learning capabilities. By understanding these distinctions, learners are now better equipped to grasp the potential advantages and complexities involved in working with AI systems, enabling them to make informed decisions during development, deployment, and management processes. We have explored the distinct benefits and risks of AI compared to traditional software systems, gaining insights into its implications and ethical considerations. This knowledge forms a solid foundation for responsibly harnessing AI's potential while navigating its complexities.

Unit 2: DoD Ethical Principles for AI

In this course unit, learners have gained an understanding of the U.S. government's role in shaping AI policies and guidelines, as well as industry standards and country-specific regulations influencing AI development. They explored the NIST AI Risk Management Framework (AI RMF), designed to manage risks associated with AI by promoting trustworthiness considerations in AI products and services. Additionally, learners learned about the DoD RAI Strategy and its Implementation Pathway, which outlines key tenets guiding the Department of Defense's responsible AI implementation, covering aspects such as responsible, equitable, traceable, reliable and governable.

Unit 3: RAI Considerations for Developing AI Systems

In this unit, learners were introduced to some of the considerations that are necessary when developing RAI systems. It is important to understand that RAI systems are about more than the technical requirements and strive to consider the context the system will be used in, the data that is a part of the system, and how the system will be used and accepted by humans.

Unit 4: AI in Operation

In this unit, learners explored the potentials for an AI system's use in real world applications. These applications can range from knowing the questions to ask when selecting a new AI system for an application to using the AI system for an application. It is important to understand that AI systems may behave in unknown or unexpected ways when introduced to new environments or when it encounters new data. Furthermore, it is also important to understand these risks, as well as ways to mitigate and protect against them before they happen.

Provide learners with additional resources to keep up with RAI

- People to watch
- Institutions to watch
- Labs/Groups to watch
- Opportunities to improve your understanding and knowledge of RAI
- Tools (or broad types of tools of different contexts) to consider

Bibliography

[1] Future AI, "Traceability," Future AI, 2023. [Online]. Available: https://future-ai.eu/principle/traceability/#:~:text=The%20Traceability%20principle%20states%20that,validation%20to%20deployment%20and%20usage..

[2] U.S. Department of Defense, "DOD Adopts Ethical Principles for Artificial Intelligence," 24 February 2020. [Online]. Available: https://www.defense.gov/News/Releases/Release/Article/2091996/dod-adopts-ethical-principles-for-artificial-intelligence/.

[3] National Institute of Standards and Technology (NIST), "NIST AI RMF Playbook," NIST, 2023. [Online]. Available: https://airc.nist.gov/AI_RMF_Knowledge_Base/Playbook.

[4] https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.pdf, "Artificial Intelligence Risk Management Framework (AI RMF 1.0)," January 2023. [Online]. Available: https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.pdf.

[5] DoD Responsible AI Working Council, "Responsible Artificial Intelligence Strategy and Implementation Pathway," U.S. Department of Defense, 2022.

[16] J. Dunnmon, B. Goodman, P. Kirechu, C. J. Smith and A. Van Deusen, "Responsible AI Guidelines in Practice," Defense Innovation Unit, 2021.

[17] Quantifind, "Quantifind," Quantifind, [Online]. Available: https://www.quantifind.com.

[18] Quantifind, "Putting Responsible AI into Practice," Quantifind, [Online]. Available: https://www.quantifind.com/resources/putting-responsible-ai-into-practice/.

[19] Quantifind, "Quantifind Project with the Department of Defense and Defense Innovation Unit Serves as Case Study in New "Responsible AI Guidelines in Practice" Report," Quantifind, 8 December 2021. [Online]. Available: https://www.quantifind.com/press/responsible-ai-guidelines-in-practice/.

[20] Fivecast, "Our Mission," Fivecast, [Online]. Available: https://www.fivecast.com/company/mission-and-heritage/.

[21] D. S. James, "Responsible And Trustworthy AI For Open-Source Intelligence," Fivecast, 24 January 2023. [Online]. Available: https://www.fivecast.com/blog/responsible-and-trustworthy-artificial-intelligence-in-osint/#:~:text=All%20Fivecast%20AI%20technology%20solutions,and%20run%20efficiently%20in%20production..

[22] Eightfold.ai, "eightfold.ai Home Page," [Online]. Available: https://eightfold.ai.

[23] Eightfold, "Responsible AI at Eightfold," [Online]. Available: https://eightfold.ai/wp-content/uploads/Responsible_AI_at_Eightfold.pdf.

[24] B. McMahan and D. Ramage, "Federated Learning: Collaborative Machine Learning without Centralized Training Data," Google, 06 04 2017. [Online]. Available: https://ai.googleblog.com/2017/04/federated-learning-collaborative.html.

[25] B. McMahan and A. Thakurta, "Federated Learning with Formal Differential Privacy Guarantees," 28 02 2022. [Online]. Available: https://ai.googleblog.com/2022/02/federated-learning-with-formal.html.

[26] R. DeVaul, J. Gips and M. Sung, "Why is context-awareness important?," MIT Media Lab, [Online]. Available: https://www.media.mit.edu/wearables/mithril/intro/topic3.html#:~:text=Context%2C%20AI%2C%20the%20frame%20problem%20and%20you.&text=Context%2C%20not%20hard%20AI%2C%20is,and%20more%20like%20human%20beings..

[17] S. Harding, "US Army soldiers felt ill while testing Microsoft's HoloLens-based headset," arsTechnica, 13 October 2022. [Online]. Available: https://arstechnica.com/gadgets/2022/10/microsoft-mixed-reality-headsets-nauseate-soldiers-in-us-army-testing/.

[18] L. D. Dance, "3Q-Do No Harm Framework," 2019. [Online]. Available: https://serviceease.net/3q-do-no-harm-framework.

[19] C. Gardner, K.-M. Robinson, C. J. Smith and A. Steiner, "Contextualizing End-User Needs: How to Measure the Trustworthiness of an AI System," Software Engineering Institute, 17 July 2023. [Online]. Available: https://insights.sei.cmu.edu/blog/contextualizing-end-user-needs-how-to-measure-the-trustworthiness-of-an-ai-system/.

[20] C. J. Smith, "Designing Trustworthy AI for Human-Machine Teaming," Sofware Engineering Institute, 9 March 2020. [Online]. Available: https://insights.sei.cmu.edu/blog/designing-trustworthy-ai-for-human-machine-teaming/.

[21] C. J. Smith, "Designing Trustworthy AI: A Human-Machine Teaming Framework to Guide Development," *ArXiv*, 2019.

[22] M. Konaev and H. Chahal, "Building trust in human-machine teams," Brookings, 18 February 2021. [Online]. Available: https://www.brookings.edu/articles/building-trust-in-human-machine-teams/#:~:text=Human%2Dmachine%20teaming%20is%20a,interactions%20and%20interdependencies%20between%20them..

[23] C. J. Smith, "Beyond Interaction: Human-Machine Teaming," 13 July 2021. [Online]. Available: https://apps.dtic.mil/sti/trecms/pdf/AD1145415.pdf.

[24] Usability.gov, Usability.gov, [Online]. Available: https://www.usability.gov/what-and-why/user-experience.html.

[25] J. Nielsen, "Usability 101: Introduction to Usability," Nielsen Norman Group, 3 January 2012. [Online]. Available: https://www.nngroup.com/articles/usability-101-introduction-to-usability/.

[26] I. Sydorenko, "Human-in-the-Loop in Machine Learning: A Handful of Arguments in Favor," Label Your Data, 20 July 2023. [Online]. Available: https://labelyourdata.com/articles/human-in-the-loop-in-machine-learning.

[27] G. Wang, "Humans in the Loop: The Design of Interactive AI Systems," Stanford University Human-Centered Artificial Intelligence, 20 October 2019. [Online]. Available: https://hai.stanford.edu/news/humans-loop-design-interactive-ai-systems.

7
]
[A. Wolfewicz, "Human-in-the-Loop in Machine Learning: What is it and How Does it
2 Work?," Levity, 16 November 2022. [Online]. Available: https://levity.ai/blog/human-in-the-
8 loop.
]
[W. Watson and C. Marsh, "Artificial Intelligence Bias in Healthcare," Booz Allen Hamilton,
2 [Online]. Available: https://www.boozallen.com/c/insight/blog/ai-bias-in-healthcare.html.
9
]
[K. Armstrong, "ChatGPT: US lawyer admits using AI for case research," BBC, 27 May 2023.
3 [Online]. Available: https://www.bbc.com/news/world-us-canada-65735769.
0
]
[W. Japan, "A Complete Guide to Japanese Road Signs: Meanings and Differences," W.Japan,
3 6 September 2019. [Online]. Available: https://www.tsunagujapan.com/wow_02312/.
1
]
[AI Incident Database, "AI Incident Database," 2023. [Online]. Available:
3 https://incidentdatabase.ai.
2
]
[S. Klassen and C. Fiesler, "“Run Wild a Little With Your Imagination”: Ethical Speculation
3 in Computing Education with Black Mirror," in *53rd ACM Technical Symposium on*
3 *Computer Science Education*, Providence, 2022.
]
[C. Fiesler, "Black Mirror, Light Mirror: Teaching Technology Ethics Through Speculation,"
3 15 October 2018. [Online]. Available: https://cfiesler.medium.com/the-black-mirror-writers-
4 room-teaching-technology-ethics-through-speculation-f1a9e2deccf4.
]
[D. Brown, "UX in the Age of Abusability," 18 September 2018. [Online]. Available:
3 https://greenonions.com/ux-in-the-age-of-abusability-797cd01f6b13.
5
]
[C. J. Smith, "Designing Ethical AI Experiences: Checklist and Agreement," Software
3 Enigneering Institute, 12 December 2019. [Online]. Available:
6 https://resources.sei.cmu.edu/asset_files/FactSheet/2019_010_001_636622.pdf.
]
[N. Martelaro and W. Ju, "What Could Go Wrong? Exploring the Downsides of Autonomous
3 Vehicles," in *International Conference on Automotive User Interfaces and Interactive*
7 *Vehicular Applications (AutomotiveUI)*, 2020.
]

[T. Gebru, J. Morgenstern, B. Vecchione, J. Wortman Vaughan, H. Daumé III and K. 3 Crawford, "Datasheets for Datasets," *Communications of the ACM,* vol. 64, no. 12, pp. 86-92, 8 2021.
]
[M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. 3 Raji and T. Gebru, "Model Cards for Model Reporting," *ArXiv,* 2019.
9
]
[J. A. Goldstein, G. Sastry, M. Musser, R. DiResta, M. Gentzel and K. Sedova, "Generative 4 Language Models and Automated Influence Operations: Emerging Threats and Potential 0 Mitigations," 10 January 2023. [Online].
]
[Sentient Digital Inc., "The Most Useful Military Applications of AI In 2023 And Beyond," 4 Sentient Digital Inc., February 2023. [Online]. Available: https://sdi.ai/blog/the-most-useful-1 military-applications-of-ai/.
]
[M. Anagnostou, O. Karvounidou, C. Katritzidaki, C. Kechagia, K. Melidou, I. Konstantinidis, 4 E. Kapantai, C. Berberidis, I. Magnisalis and V. Peristeras, "Characteristics and challenges in 2 the industries towards responsible AI: a systematic literature review," *Springer Link,* 2022.
]
[C. Wai-Loon Ho, "Operationalizing "One Health" as "One Digital Health" Through a Global 4 Framework That Emphasizes Fair and Equitable Sharing of Benefits From the Use of 3 Artificial Intelligence and Related Digital Technologies," *Frontiers,* vol. 10, 2022.
]
[J. Patrick, "How to Check the Reliability of Artificial Intelligence Solutions—Ensuring Client 4 Expectations are Met," *Applied Clinical Informatics,* vol. 10, no. 2, pp. 269-271, 2019.
4
]
[N. Barney, "Artificial Intelligence (AI) Governance," TechTarget, [Online]. Available: 4 https://www.techtarget.com/searchenterpriseai/definition/AI-governance.
5
]
[Quantifind, "Quantifind Home Page," [Online]. Available: https://www.quantifind.com.
4
6
]