Transcript

Welcome to the Certified Information Security Training and certification program for the certified NIST artificial Intelligence Risk Management Framework, 1.0 architect. This is the final and 4th part of four of the introduction to the training and certification program. I'm Alan keel. Thanks for joining me. You may remember from my general introduction to AI that I provided in a previous presentation that one of our overall issues at hand with artificial intelligence is trying to achieve trustworthiness, trustworthiness, that the system is going to be accurate, trustworthiness. That it's not going to harm trustworthiness, that it's going to facilitate positive benefit instead of negative consequence. So let's go ahead and take a look at AI trustworthiness, challenges and risk that the AI risk management framework can help us resolve and effectively manage. So what is AI trustworthiness? AI trustworthiness, aims to create systems that are reliable, ethical, and accountable, fostering public trust in artificial intelligence. The NIST trustworthy and responsible AI Resource Center explains that characteristics of trustworthy AI include that it is safe, secure and resilient, explainable and interpretable privacy, enhanced that it is fair, with harmful bias being managed. That it is also has all of these characteristics that are then tested and validated to be reliable, so that way we can have and achieve overall accountability and transparency, so valid and reliable is a necessary condition of trustworthiness, as shown by the other trustworthiness. Characteristics. Accountable and transparent is shown as a vertical box because it relates to the other characteristics as well. So we learned that valid and reliable is the foundation for all other desired AI trustworthiness characteristics. What is valid and reliable validation is the confirmation through the provision of objective evidence, that the requirements for a specific intended use or application. Has actually been fulfilled. Deployment of AI systems which are inaccurate, unreliable or poorly generalized to data and settings beyond their training creates and increases negative AI risk and reduces trustworthiness. Reliability is defined as the ability of an item to perform as required without failure for a given interval of time under given conditions. So reliability is a goal for overall correctness of AI system operation under the conditions of expected use. And over a given period of time, including the entire lifetime of the system. So our first characteristic that is on that base of validation and reliability is going to be safe. AI system should not under defined conditions lead to a state in which human life, health, property or the environment is endangered. Safe operation of AI systems is improved through responsible design, development and deployment practices. Clear information to developers on responsible use of the system. Responsible decision making by Deployers and end users and explanation and documentation of risk based on empirical evidence of incidents. The next AI trustworthiness characteristic is secure and resilient AI systems as well as the ecosystems in which they're deployed may be said to be resilient if they can withstand unexpected adverse events or unexpected changes in their environment, or use or. If they can maintain their functions and structure in the face of internal and external change. And that they can degrade safely and gracefully when this is necessary. Common security concerns relate to adversarial examples, data poisoning, and the exfiltration of models, training data or other intellectual property through AI system endpoints. Continuing with secure and resilient, we need to ensure that AI systems can maintain confidentiality, integrity, and availability through protection mechanisms that prevent unauthorized access and use, so that way they can be said to be secure. Guidelines in the NIST cybersecurity framework to not 0. Are among those which are very applicable here. Our next desired trustworthiness characteristic is accountable and transparent. Accountability assumes transparency. Transparency reflects the extent to which information about an AI system and its outputs is available to individuals interacting with that system. Regardless of whether they are even aware that they are doing so. With accountable and transparent, we also ensure that meaningful transparency provides access to appropriate levels of information based on the stage of the AI lifecycle and tailored to the role or knowledge of AI actors or individuals interacting with or using the AI system. And by promoting higher levels of understanding, transparency increases confidence in the AI system itself. Accountable and transparent scope spans from design decisions and training data to model training the structure of the model, its intended use cases, and how and when deployment post deployment or end user decisions were made and by whom. Transparency is often necessary for. Actionable redress related to AI system outputs that are incorrect or otherwise lead to negative impacts. And transparency should also include and consider human AI interaction. For example how a human operator or user is notified when a potential or actual adverse outcome caused by an AI system is detected. A transparent system is not necessarily an accurate privacy enhanced, secure or fair system, but at least it's well understood by all that use it. So this all means that the role of AI actors should be considered when seeking accountability for the outcomes of the AI systems. The relationship between risk and accountability associated with AI and technological systems more broadly differs across cultural, legal, sectoral. And societal context. And when consequences are severe, such as when life and liberty are at stake, AI developers and DEPLOYERS should consider proportionately and proactively adjusting their transparency and accounting practices. Maintaining organizational practices and governing structures for harm reduction like risk management can help lead to more accountable systems. The next desired characteristic is explainable and interpretable. Explain ability refers to a representation of the mechanisms underlying AI systems operation, whereas interpretability refers to the meaning of AI systems output in the context of their designed functional purposes. So together explain ability. And interpretability assist those operating or overseeing an AI. System as well as users of an AI system to gain deeper insights into the functionality and trustworthiness of the system, including its outputs. So to put it all together, transparency explain ability and interpretability are distinct characteristics that support each other. Transparency can answer the question of what happened in the system. Explain ability can answer the question of how a decision was made in the system. Interpretability can answer the question of why a decision was made by the system and its meaning or context to the user. So with explainable and interpretable, the underlying assumption is that perceptions of negative risks stem from a lack of ability to make sense of or contextualize system output appropriately. Explainable and interpretable AI systems offer information that will help and users understand the purposes and potential impact of an AI system. And the next AI trustworthiness characteristic is privacy enhanced. Privacy refers generally to the norms and practices that help to safeguard human autonomy, identity, and dignity. These norms and practices typically address freedom from intrusion, limiting observation, or an individual's agency to consent to disclosure or control facets of their own identities. Privacy values such as anonymity, confidentiality, and control generally should guide choices for AI system design, development, and deployment. So privacy related risk may influence security bias and transparency and come with trade-offs with these other characteristics. Similar to safety and security specific technical features of an AI system may promote or potentially reduce privacy. AI systems can also present new risk to privacy by allowing inference to identify individuals or previously private. Information about individuals. And the next characteristic for achieving AI trustworthiness is perhaps one of the most important and elusive. We want our AI system to be fair and we want to be able to manage any potential harmful bias that may be inherent to the system. So what is fairness, fairness and AI includes concerns for equality and equity. By addressing issues such as harmful bias and potential discrimination. Standards for fairness, however, can be complex and difficult to define because. How fairness is perceived differs from culture to culture. Even systems in which harmful biases are mitigated are still not necessarily fair. For example, systems in which predictions are somewhat balanced across demographic groups still may be inaccessible to individuals with disabilities or affected by the digital divide or may. Exacerbate existing disparities or systemic biases?

Implicit bias is a stereotype. It's an unconscious, involuntary opinion that informs our actions all day, every day. These implicit biases are now embedded in our algorithms.

If we capture the world completely accurately. We might capture inappropriate phenomena that we wish to exclude from the. Systems that we are building.

If the data is unjust and have biases in them, the model will learn and propagate them.

I'm very worried about a world where AI is used to increase disparities between groups where AI is biased, but there's a lot we can do to prevent that world from happening. There's a lot we can look at in terms of existing laws and regulatory frameworks, their best practices.

There are a.

Whole host of things we can do to make sure that the AI. Adopting is not biased.

The time to act is. Now we're far enough along that we know that these harms are not theoretical, but we're not too far, that we can't stop them from recreating decades and centuries of discrimination in our society.

At the end of the day, computers are not programmed by themselves, they're actually programmed by human beings and human beings bring their values, assumptions and norms about the world that factor into these models that we're actually creating. So we have to care about bias and AI. If we as humans don't care about bias in AI, then we honestly don't care about the flow of democratic participation and how. We actually democratize systems so that everybody gets to be able to participate.

So bias is broader than demographic balance and data representativeness. NIS has identified three major categories of AI bias to be considered and managed. The first, computational and statistical that's known as. Technical bias. We have human cognitive bias and systemic bias. We'll be looking at each of these in turn. So each of these categories of AI bias can occur in the absence of prejudice, partiality or discriminatory intent. So again, computational and statistical biases can be present in the AI data sets themselves as well as the algorithmic processes, and they often stem from systematic and errors due to non representative. Examples. We also have human cognitive biases that relate to how an individual or group perceives AI system information to make a decision or to fill in missing information, or how humans think about purposes and function of an AI system itself. So human cognitive biases. Are omnipresent in decision making processes across the AI life cycle and system use, including the design, implementation, operation and maintenance of AI. And finally, systemic bias can be present in AI data sets. The organizational norms, practices, and processes across the area out life cycle, as well as the broader society that uses AI systems. So as we summarize the challenges to achieving trustworthy AI systems, we know that creating trustworthy AI really requires balancing each of these characteristics based on the AI systems context of use. And while all characteristics are sociotechnical system attributes, accountability and transparency also relate to the processes and activities internal to an AI system and its external settings. So neglecting these characteristics can increase AI risk. And these AI trustworthy characteristics are interrelated, so addressing AI trustworthiness characteristics individually will not necessarily ensure AI system trustworthiness. trade-offs are usually involved and rarely do all characteristics apply in every setting and some will be. More or less important in any given situation. Ultimately, trustworthiness is a social concept that ranges across a spectrum and is only as strong as its weakest characteristics. And as I explained, when managing AI risk, organizations can face difficult decisions in balancing these characteristics. trade-offs may emerge between optimizing for interpretability. While also still achieving privacy in other cases, organizations might face a trade off between predictive accuracy. And interpretability. Under certain conditions, such as data sparsity, privacy enhancing techniques can result in a loss of accuracy, which would negatively affect resulting decision. So the NIST AI risk Management Framework 1.0 again is enhanced with an AI risk management playbook that is also published by this. And this is an ever evolving resource. And what we're going to do now is take a look at some of the. Initial use cases that NIST's released to give us a better feel for how this framework is used in business. So what are AI? RF, 1.0 profiles? They are use case profiles that are implementations of the AI risk management functions, categories and sub categories for a specific setting or application based on the requirements, risk tolerance and resources. Of the framework user now. Currently NIST doesn't make so many of these use cases available, but they do have a few links are provided here for your convenience. So again, what are use cases good for? Well, they may illustrate and offer insights on how risk can be managed at various stages of the AI life cycle or in a specific sector. Technology or end use application. Armv use case, profiles assist organizations in deciding how they might best manage AI risk that is well suited with their goals and that considers legal, regulatory requirements and best practices, as well as reflecting risk management priorities. Now NIST publishes several different flavors of AI RMF profiles, the first being temporal. These are current state or desired state target state profiles. They are descriptions of either the current state or desired target state. Of specific AI risk management activities within a given sector, industry organization or application context. And AI RMF current profile indicates how AI is currently being managed and the related risk in terms of current outcomes. A target profile indicates the outcomes needed to achieve the desired or target AI risk management goals. Typically, we will have both a current state and a target state profile because what is in between will be our road map to improvement. So comparing current and target profiles gives us that road map by revealing gaps to be addressed to meet AI risk management objectives. We know that action plans could then be developed to address these gaps, fulfilling outcomes and a given category or subcategory of. Objective and prioritization of gap mitigation is driven by the user's needs and risk management processes and the risk based approach also enables framework users to compare their approaches with other approaches to gauge the resources needed to achieve AI risk management goals. In a an effective and comprehensive and prioritized manner. NIST also publishes cross sectoral AI profiles. These are similar to what NIST has done with the cybersecurity framework, with its community profiles. It's essentially covering risk of models or applications that can be used across use cases or sectors cross sectoral profiles. Can also cover how to govern map, measure and manage risk for activities or business processes coming across sectors such as the use of large language models, cloud based services or acquisition. So this introduction has been quite a journey just getting.