AUTONOMOUS CYBERSECURITY AND AI RISK MANAGEMENT FOR UNCREWED SYSTEMS: CHALLENGES AND OPPORTUNITIES USING THE NIST FRAMEWORKS

Raymond Sheh,[1] Karen Geappen,[2] Donald Harriss[3]

Uncrewed systems are increasingly connected and automated. With these increases come significant changes in the cybersecurity and Artificial Intelligence (AI) risks associated with such systems. This is particularly important in safety-critical applications, such as public safety, security, infrastructure inspection, hazardous materials handling, transportation and delivery, and entertainment.

This paper provides an overview of uncrewed systems and the ways in which increasing connectivity and automation pose new and increased cybersecurity and AI risks. We provide some examples of these risks and suggest examples of controls that may be employed to manage these risks, with reference to the NIST Cybersecurity Risk Management Framework (CSF) 2.0 and the NIST AI Risk Management Framework (AI RMF). We also outline our working group's efforts to produce detailed, actionable guidance and other resources to help stakeholders better manage these risks, particularly for the uncrewed aircraft systems (UAS) for Public Safety application.

INTRODUCTION

Uncrewed systems, such as uncrewed aircraft systems (UAS), are increasingly widely used in critical applications such as public safety, security, infrastructure inspection, hazardous materials handling, transportation and delivery, and entertainment. Previous generations of commonly available uncrewed systems have been mostly human-controlled. Artificial Intelligence (AI) provides basic control of the motors and sensors in response to simple commands, such as to move in a given direction. Such systems also have limited integration into other critical software systems. While cybersecurity risks have always been a reality, they were usually limited in scope with realized risks only impacting the individual system.

Systems like Drone as First Responder (DFR) and autonomous patrol and inspection systems are now frequently deployed. These have increasing levels of autonomy and connectivity with critical software systems, such as emergency dispatch and digital twins as well as interconnection with other uncrewed systems. Certainly, highly connected, intelligent systems are not new. Specialized industries and the military have had a wide variety of such systems for several decades, often pushing the forefront of the technology. However, technology leaps come with a significant

[1] Uncrewed Aircraft Systems Research Lead, Public Safety Communications Research Division, National Institute of Standards and Technology, 100 Bureau Drive, Gaithersburg MD 20877, USA
[2] Independent Researcher, PO Box 174, Bull Creek, Western Australia, Australia, 6149
[3] Uncrewed Aircraft Systems Technical Lead, Public Safety Communications Research Division, National Institute of Standards and Technology, 325 Broadway, Boulder, CO 80305, USA

https://doi.org/10.52202/075106-0003

cost in custom research and development, in addition to the policies, procedures, and other controls to manage risk. The results have historically been difficult, or impossible, to apply widely to mainstream users and other stakeholders in the uncrewed systems ecosystem due to specialization, classification, and other restrictions on dissemination.

There are few resources that help mainstream stakeholders apply existing language, standards, metrics, governance, and risk management frameworks to manage the novel cybersecurity and AI risks posed by these increasingly connected and intelligent uncrewed systems. Contextualizing cybersecurity and AI risks related to the adoption of such technologies can support users in making informed choices for effective risk management. Proper risk management can remove barriers to adoption, improve efficiencies, reduce organizational risk exposure, and increase return on investment.

This paper aims to provide an overview of cybersecurity and AI risk management and outline some of the challenges to managing risks posed by highly connected and intelligent uncrewed systems. We provide some salient examples of how these risks can manifest themselves, and we end with some actionable suggestions that organizations can take to improve their risk management. While we will be mainly using UAS[4] and public safety as the example modalities and applications, the general principles can be applied across other applications, including ground and water vehicles.

BACKGROUND

Definitions

There are various competing definitions for the key terms of cybersecurity and artificial intelligence. While there is no universal definition for these terms, we will use the following definitions for this discussion.

Cybersecurity. For this discussion, we use the definition of cybersecurity used in NIST Special Publication 800-53 Rev. 5[5]:

"Prevention of damage to, protection of, and restoration of computers, electronic communications systems, electronic communications services, wire communication, and electronic communication, including information contained therein, to ensure its availability, integrity, authentication, confidentiality, and nonrepudiation."

We will delve into three critical factors that users must consider to understand the definition and scope of cybersecurity in general and for uncrewed systems in particular. The first is identifying the assets and interests to be protected. This may be physical, but it may also be something more abstract, such as privacy, integrity, trust (in the system, data, or operation), reputation, continuity of operation, or financial, regulatory, or legal exposure. The second factor is identifying the entities from which the asset or interest must be protected. The third involves assessing what the potential

[4] In this paper, we will make the distinction between the Uncrewed Aerial Vehicle (UAV), which is the equipment that flies, and the overall Uncrewed Aircraft System (UAS), which includes the UAV and all of the systems that directly support and integrate with its operation, such as the Operator Control Unit (OCU), flight planning software, maintenance and logistics software, and interfaces to collaboration and Common Operating Picture (COP) systems.

[5] Joint Task Force, "NIST Special Publication 800-53 Revision 5 Security and Privacy Controls for Information Systems and Organizations," National Institute of Standards and Technology (NIST), 2020-12-10, https://doi.org/10.6028/NIST.SP.800-53r5

https://doi.org/10.52202/075106-0003

damage might be. Identifying these three critical factors informs priorities for applying cybersecurity.

In the past, industry attention was mostly focused on attackers looking to take over control of systems and/or gain access to protected information. Such attackers still exist; however, the modern threat landscape, coupled with a higher degree of interconnection between systems, requires considering a wider variety of potential attackers and their targets.

For example, as uncrewed systems become more tightly interconnected with other critical systems, such as dispatch, collaboration, and geographical information systems (GIS), the uncrewed system can become an entry point for pivoting into the overall network of the organization or partner organizations. An attacker might have no interest in the uncrewed system except that it might be the least-protected device with access to the network.

As technology becomes cheaper and more widely available, it is easier for a more diverse range of adversaries to attack critical systems. Also, as critical systems become more interconnected, it becomes easier for adversaries to find paths through the network to those systems, paths that might not even be known to the owners of those systems. For example, the threat of ransomware exists against anyone who attackers think will pay to regain access to their data or systems. At the same time, "hacktivists" might only care that their actions will result in disruption and publicity regardless of the actual value of their target. There are also individuals and groups who simply want to cause disruption regardless of the target.

As we will discuss later, this changing threat landscape, in terms of both advancements in uncrewed systems and changes in threat actors, dramatically changes the risk calculus for many users of uncrewed systems who might have previously considered themselves an unlikely target.

Artificial Intelligence. Artificial Intelligence, or AI, is a term that has very different meanings in different contexts. ISO/IEC 22989:2022 defines AI as "a technical and scientific field devoted to the engineered system that generates outputs such as content, forecasts, recommendations or decisions for a given set of human-defined objectives.". This definition reflects the broad scope of AI, although it also encompasses a lot of traditional software systems that one might not immediately consider to be AI, particularly in the popular media. For our discussion, we will focus on dividing up the space of AI in two ways.

First, AI systems can be divided by how much visibility humans have into their inner workings. They can tend towards being more transparent or more opaque in terms of the ability of humans to understand, at an actionable level, within a reasonable timeframe, and with reasonable resources, how it works and why it made a given decision. This distinction is critical for discussing risk management as much of traditional risk management relies on predicting what a system might do and performing meaningful root cause analysis.

Second, AI systems can be divided into the ways in which they make their decisions and the sources of knowledge used to make those decisions. Much of the recent discussion in popular media has focused on deep learning (DL). It is important to understand that there are many other forms of AI with different capabilities, benefits, and, more importantly, risks and risk management techniques.

For example, DL is a type of statistical machine learning (statistical ML). DL depends on fitting highly complex, randomly initialized sets of linear equations, with some non-linear elements, to the observed input data. Other forms of Statistical ML fit equations to the observed input data that are less complex and less random. Such systems might not be able to represent the same complexity as DL, but they might also be easier for humans to understand and the system may behave more consistently.

Statistical ML is a type of machine learning (ML) that bases its behavior on the statistics of the data it sees. While statistical ML is the dominant form of ML, there are forms of ML that also take in other forms of information, such as mathematical models of the system, and use the observed data to augment these models. Such systems might require more knowledge of how the system works, such as a knowledge of physics, but can also behave more predictably and require less training data.

ML is a type of AI that takes in, processes, and stores some form of information from the task or environment and uses this stored information to influence the system's future behavior. Its behavior is not only dependent on its programming but also on historical information with which it was presented. Other forms of AI, such as classical planning, reasoning, logic programming, and expert systems, generally take input from outside the task to inform its behavior.

For example, we might choose two different ways to teach a system how to catch a ball, based on observing its trajectory at the moment that it is thrown. An AI system based on DL would observe the ball being thrown at it many times, learn to predict where the ball is going to land based on what it sees of the initial trajectory, and then learn how to move to that location. This is similar to a human learning a skill by practicing it repeatedly.

Instead, an AI system based on reasoning, rather than ML, might have background knowledge that includes the equations that represent physical properties such as ballistics, gravity, and inertia. It might infer from this knowledge how the ball will move given its observed initial trajectory and calculate where it should move to catch the ball. It represents the task in an analytical manner rather than modeling it based on observation.

Both approaches have benefits and risks. For example, the DL system will naturally account for more complex aspects of the environment, such as air resistance and variations in the wind, that are incorporated into its observations of the flight of the ball. In contrast, the reasoning system might not have background knowledge that includes these factors. Instead, the reasoning system might only account for these as random errors.

In contrast, the DL system is opaque. For example, it does not represent the concept of "gravity" in a way that a human can adjust. If the system were taken to the Moon or Mars with a different gravitational field, it would have to relearn how to catch the ball. Worse yet, as there is no way for a human to tell the system that gravity has changed, it will be impossible to tell if there are any remnants of its behavior that still rely on Earth's gravity. In contrast, the gravity term in the reasoning system could be adjusted explicitly, and the system is likely to perform in an appropriate manner.

Taking risk management designed for a system with one type of AI and applying it to a system using a very different type of AI, even if it performs a similar task, can easily lead to a situation where benefits are foregone and risks are ignored. Blindly adopting risk management plans becomes even more complicated in systems that are hybrids of these techniques. For example, some systems might be analytical in nature but also include a learning component to account for complexities in the world about which it does not otherwise know. Hybrid AI systems can lead to the best of both worlds, but they also risk becoming the worst of both worlds if the risks are not appropriately managed.

Autonomy in Uncrewed Systems

Uncrewed systems vary considerably in how they make use of autonomy and automation. For this discussion, we refer to autonomy as systems that make decisions independently based on their own sensing and computation. Automation, on the other hand, refers to systems that execute

https://doi.org/10.52202/075106-0003

predefined instructions and are only capable of simple deviations from the instructions in response to unexpected events. An example would be a robot that stops when it encounters an obstacle.

Many uncrewed systems are remotely operated with a single human operator directing the vehicle's movements through an Operator Control Unit (OCU). For example, many UAVs are still controlled by humans using basic direction commands like forward, backward, left, right, climb, descend, and yaw.

This commonality can be somewhat deceiving; while a wide variety of UAVs might take the same basic commands, they can vary in the amount of autonomy they apply to follow these commands. Increasing levels of autonomy usually make it easier for the operator to control the UAV. They move the UAV's behavior closer to what the operator expects from an ideal system. This autonomy handles low-level control of the UAV and leaves the operator to handle the movements required to achieve the task at hand. For example, even modest UAVs now incorporate Global Positioning System (GPS) and cameras to determine their position and help them to automatically counteract wind. This allows the operator to fly the UAV as if it were operating in still air, transparently delegating the task of dealing with wind to the autonomy of the UAV.

From a risk management perspective, this example of autonomously dealing with changing wind conditions presents both benefits and risks as compared to systems without it. Clearly, such a system reduces the risk associated with an operator, particularly one who is less experienced, losing control of the UAV if they are insufficiently skilled to respond to changing wind conditions. This is particularly important near obstacles, both because of the increased risk of collision and the more complex air currents around obstacles that even experienced pilots can have trouble managing.

However, it also presents a significant risk, especially if the pilot is unaware of what the autonomy is doing, the limit of its capabilities, and from where the autonomy gets its information. For example, UAVs are mechanically limited in how fast they can fly relative to the air around them based on factors such as their aerodynamics and the power of their motors. Complex maneuvers close to these mechanical limits can pose risks to the stability of the UAV. An operator might command the UAV to fly forward slowly relative to the ground, but if there is a strong headwind, the UAV might actually be flying fast and close to its mechanical limits relative to the air. As the UAV is autonomously compensating for the headwind, the pilot may not even be aware that the UAV is close to its mechanical limits. If the operator then commands a complex maneuver, thinking that the UAV is still moving slowly and far from its mechanical limits, the UAV may be at greater risk of loss of control than the operator would expect.

UAVs, and uncrewed systems in general, have a wide variety of forms of autonomy beyond simply accounting for variations in the environment. Uncrewed systems can vary in complexity in their abilities to perform tasks such as driving or flying to a given point and have different capabilities for autonomously addressing unexpected events on their way to their commanded destination. For example, they may:

- Take a straight-line path.
- Drive or fly around unexpected obstacles.
- Plan a variety of paths and choose the one that they deem to be the best based on a wide variety of factors such as power consumption, the riskiness of the different paths (by some measure), and so-on.

Even more complex levels of autonomy include mission-level decision making and behavior, such as finding and tracking an object of interest or coordinating with other entities.

Autonomy at task and mission levels presents particular challenges to risk management because of the wide variety of ways in which they can be implemented and the unique risks posed. Proper risk management requires information from vendors and developers that often does not flow to the end users outside of highly integrated sectors such as defense. Short of acquiring the requisite information, end users can at least be informed of the potential risks for which they should be watching and managing.

The aforementioned levels of autonomy are split across levels of abstraction loosely based on the NIST Autonomy Levels for Unmanned Systems (ALFUS)[6]. In the uncrewed systems community, levels of autonomy are also divided based on how much human supervision is required. There are loose correlations between the two different ways in which levels of autonomy are defined, but they do represent very different concepts.

One popular definition is from the Society of Automotive Engineers (SAE) Taxonomy and Definitions of Terms Relating to Driving Automation Systems for On-Road Motor Vehicles J3016_202104 (J3016)[7], often simply referred to as the "SAE Levels of Autonomy." This six level system ranges from level 0, where the operator must devote focused attention to the operation of the system, through to level 5, where the operator need not be present or on call.

It is important to note that the SAE levels of autonomy are designed specifically for roadgoing vehicles on publicly accessible roads and associated locations like parking lots. Specifying a level of autonomy for a system is only meaningful when associated with a definition of its nature, task, and environment. For example, a level 5 autonomous car is not expected to perform autonomously on an off-road trail or to back up a large trailer. Similarly, it is generally meaningless to state that a UAV has or requires "Level 5 Autonomy" if the environment and task are not specified.

Good Risk Governance for Uncrewed Systems

Governance is the collection of systems of processes, mechanisms, and rules by which an organization is controlled and operates. It ensures that the people within the organization are held accountable when moving towards organizational objectives. Appropriate governance bridges both horizontally and vertically across the organization. It is a traceable collection of artifacts linking organizational objectives with processes, procedures, and technology. It also needs to flow in both directions, with approvals informing teams and metrics reported to determine effectiveness. The focus on organizational objectives amalgamates what is important to the organization, such as social considerations, regulatory requirements, technology, and organizational skills and capabilities.

Governance is a key pillar when it comes to risk management. The aim of governance in risk management is to identify and manage relevant risks to the organization achieving its objectives. When done well, it sets the boundaries for acceptable and unacceptable behavior and risk-taking, provides definitions to remove ambiguity, and enables an organization to have clarity and coherence across all its subject matter expert teams.

[6] Ad Hoc ALFUS Working Group Participants, "NIST Special Publication 1011-II-1.0, Autonomy Levels for Unmanned Systems (ALFUS) Framework Volume II: Framework Models Version 1.0," National Institute of Standards and Technology (NIST), 2007-12-28, https://doi.org/10.6028/NIST.sp.1011-II-1.0

[7] Society of Automobile Engineers (SAE) International, "Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles J3016_202104," 2021-04-30, https://www.sae.org/standards/content/j3016_202104/

https://doi.org/10.52202/075106-0003

Good governance becomes even more critical when an organization uses AI or uncrewed systems. By definition, AI implies that elements of the decision are taken out of humans' hands (and minds). Similarly, by definition, uncrewed systems implies that human oversight of the system is reduced. Furthermore, this reduction of human oversight is often also accompanied by greater connectivity and access to other systems, as the AI and uncrewed systems directly access mission data that humans would have normally handled. Good governance is required to manage the risks of activities that humans do not directly oversee.

Effectively, the technology becomes a risk multiplier by opening up different attack vectors in an organization and increasing the number of exposed assets and interests, including the aforementioned abstract assets, such as reputation and legal exposure. This makes risk management control that can cross-cut organizational teams and hierarchy even more important. However, like any risk management control, incorrect application of a control can cause more damage than no control.

Risks that should be covered by governance are those that would hinder the organization from achieving its goals. This generally revolves around protecting an organization's assets and interests, both tangible and intangible, including those of connected systems. Adherence to regulations and legal requirements can be seen as either an intangible reputational risk or a tangible risk related to finances, exposure to legal action, or if applicable, a license to operate.

Many frameworks can be used for the various levels and domains that need risk management governance. The NIST AI Risk Management Framework and the Cybersecurity Framework 2.0, discussed later in this document, may support an organization in establishing operating behaviors and processes. In the context of new technologies and novel applications, the process layer of governance is where we find gaps with published guidance. These gaps include relevant and materially useful metrics that need to be relayed up the governance hierarchy to establish overall organizational risk.

Existing Relevant Frameworks, Standards, and Guidance

As awareness of cybersecurity and AI risks becomes more widespread, various frameworks and guidance are now available to help organizations better manage these risks. Most current guidance is general and not written with uncrewed systems in mind. However, it can provide a good starting point in developing appropriate risk management. In this section, we discuss some salient examples of recent frameworks, standards, and guidance relevant to uncrewed systems.

Cybersecurity Frameworks and Guidance. In early 2024, NIST released the updated Cybersecurity Framework (CSF) version 2.0, providing a set of outcomes across six cybersecurity functions for reducing cybersecurity risks. The functions are Identification, Protection, Detection, Response and Recovery with a cross-cutting function of Governance. The outcomes are industry and organization-agnostic, and therefore, the CSF does not define the implementation of the outcomes. Instead, the framework suggests relevant references, such as NIST Privacy and other frameworks, for guidance on implementation within functions. Examples of resources that the CSF can assist in contextualizing include but are not limited to:

- The NIST Privacy and Risk Management Frameworks.
- NIST SP 800-53 Rev. 5, Security and Privacy Controls for Information Systems and Organizations.
- NIST SP 800-171 Rev. 2, Protecting Controlled Unclassified Information in Nonfederal Systems and Organizations.
- The Criminal Justice Information Services (CJIS) Security Policy.

https://doi.org/10.52202/075106-0003

- ISO/IEC 27001:2022 Information security, cybersecurity and privacy protection.

This contextualization allows organizations to more easily and comprehensively assess their current cybersecurity posture relative to the posture that they wish to adopt and identifies steps to move forward.

The CSF is outcomes-based and technology-agnostic, making it relevant to all technologies. However, without domain-specific guidance for uncrewed systems, best practice implementation to achieve these outcomes is left to the organization, particularly for more unusual aspects of these highly connected, intelligent systems. For many organizations, adapting this general guidance to the specific challenges of uncrewed systems can be difficult.

Specific guidance regarding the management of cybersecurity risk for uncrewed systems is still in its relative infancy. Organizations that have issued guidance that relates to uncrewed systems include:

- The Cybersecurity & Infrastructure Security Agency (CISA), part of the U.S. Department of Homeland Security (DHS), has issued guidance that focuses on the procurement and use of Uncrewed Aircraft Systems[8] rather than the development, manufacture, or broader integration into wider organizational systems and processes.
- Both the Defense Innovation Unit (DIU) of the U.S. Department of Defense (DoD) and the Association for Uncrewed Vehicle Systems International (AUVSI) have issued guidance on the procurement of UAS based on the National Defense Authorization Act (NDAA)[9]. DIU maintains a "Blue UAS" list[10] while AUVSI maintains a "Green UAS" list[11], which list the UAS that they have evaluated according to the NDAA criteria. This guidance is primarily focused on geopolitical, cybersecurity, organizational, and supply chain considerations of the UAS manufacturer rather than its integration into the organization of the end user's broader systems and risk management.
- The NIST Information Technology Laboratory (ITL)[12] produces guidance on a wide variety of cybersecurity topics, including the resources mentioned earlier in this section. While it does not have an uncrewed systems-specific focus, it does focus on applications such as Automated Vehicles[13]. Much of the guidance within the cyber-physical systems and internet of things (IoT) topics is also highly relevant to uncrewed systems.

Often, it is difficult to predict how an organization will use novel technologies and how they will create and implement the necessary cybersecurity policies and other controls to protect organizational assets and interests. The CSF 2.0, focusing on risk management outcomes, provides

[8] Cybersecurity & Infrastructure Security Agency (CISA), "Be Air Aware," accessed 2024-04-03, https://www.cisa.gov/topics/physical-security/unmanned-aircraft-systems

[9] House Armed Services Committee, "NDAA - National Defense Authorization Act," accessed 2024-04-03, https://armedservices.house.gov/ndaa

[10] Defense Innovation Unit, "About Blue UAS", accessed 2024-04-03, https://www.diu.mil/blue-uas

[11] Association for Uncrewed Vehicle Systems International (AUVSI), "Green UAS," accessed 2024-04-03, https://www.auvsi.org/green-uas

[12] National Institute of Standards and Technology (NIST), "Information Technology," accessed 2024-04-03, https://www.nist.gov/information-technology

[13] National Institute of Standards and Technology (NIST), "NIST Automated Vehicles Program," accessed 2024-04-03, https://www.nist.gov/programs-projects/nist-automated-vehicles-program

https://doi.org/10.52202/075106-0003

a useful reference even when domain-specific guidance has not yet caught up. However, leveraging these resources requires a high level of understanding of these novel technologies.

Applying existing frameworks to new applications requires an ongoing, iterative effort by all stakeholders with a keen understanding of established acceptable risk tolerances and how to meaningfully apply them in situations outside of the original use case. Organizations need to be flexible when dealing with cutting-edge, connected, and autonomous systems. Standard security practices and past experiences might not always apply perfectly. This is especially true in such a fast-moving field.

AI Frameworks and Guidance. Society has been aware for a long time that AI poses risks. Indeed, American writer and professor of biochemistry, Isaac Asimov, coined the "Three Laws of Robotics." Many of his stories, such as those in his collection "I, Robot"[14], revolve around how simplistic characterizations of robots with AI result in unintended consequences. When developing guidance on the use of AI, it is easy to fall into the trap of generalizing based on an understanding of how humans, or fictional systems, make decisions. Managing the risks of real AI systems requires considerable nuance and understanding. Unfortunately, this understanding can be very hard to obtain, especially with a supply chain of software and hardware that does not expose the necessary information about included AI systems to the end user.

Several recent efforts have generated frameworks, guidance, policies, and regulations to facilitate the management of AI risk. These are vital in helping organizations to generate more thorough, nuanced, technically actionable, and ethically meaningful policies and procedures. Perhaps most prominent of these are the NIST AI Risk Management Framework (AI RMF)[15], the U.S. Department of Energy (DoE) AI Risk Management Playbook (AI RMP)[16], and the European Union Artificial Intelligence Act. In this discussion, we will focus on the AI RMF as a tool to help organizations better characterize their current AI risk, determine their acceptable level of AI risk, and provide guidance to manage these risks better.

Like the various cybersecurity frameworks, the AI RMF is general and not focused on uncrewed systems. It divides the problem of analyzing an organization's exposure to AI risk through four functions: Govern, Map, Measure, and Manage. Much like the CSF 2.0, it is not a checklist. Rather, it provides a high-level overview that helps organizations better understand their risks and apply other guidance, resources, and policies to manage risks.

The guidance landscape for AI risk management is much less mature than it is for cybersecurity. The existing guidance focuses more on risks associated with the use of AI for decision-making and content generation. This includes such issues as privacy, fairness, ethics, and intellectual property. Government and standards organizations need to provide more guidance that specifically assists organizations with managing the AI risks unique to uncrewed systems.

Uncrewed System specific guidance and resources. The NIST Public Safety Communications Research (PSCR) Division has established a working group to produce guidance and other resources to assist organizations with managing cybersecurity and AI risk associated with UAS,

[14] Asimov, I., "I, Robot," *Gnome Press*, 1950-12-02

[15] National Institute of Standards and Technology (NIST), "NIST Trustworthy & Responsible AI Resource Center," accessed 2024-04-03, https://airc.nist.gov/

[16] U.S. Department of Energy (DOE), "DOE AI Risk Management Playbook (AIRMP)," accessed 2024-04-03, https://www.energy.gov/ai/doe-ai-risk-management-playbook-airmp

focusing on public safety applications[17]. The goal of this effort is to work with a wide variety of stakeholders to bridge the gap between general cybersecurity and AI risk management guidance, as well as the specific risks and challenges of uncrewed systems in general and UAS in particular. Although still in its early stages, PSCR researchers developed some preliminary resources, which we will discuss further in this document.

Factors in Managing Cybersecurity and AI Risks for Uncrewed Systems

A comprehensive treatment of the management of all risks is beyond the scope of this paper. This section will highlight a few lesser-known factors to consider when managing cybersecurity and AI risks for uncrewed systems.

Hidden Risks of a Common User Experience

One of the most significant sources of risk for uncrewed systems and intelligent, highly connected systems is their interaction with humans. The risks of an overly complex, unique user interface are well known in terms of cognitive load, errors, and user acceptance.

Having a new system behave in a way that resembles the system it replaces is often seen as reducing risk, as users are already familiar with the previous system and are less likely to make mistakes. However, if the new system has very different underlying risks, the users and the broader organization may not be aware of or understand the need for a change in their risk management. They may, therefore, continue to follow risk management procedures and policies that are inappropriate to the new system. This wastes resources on unnecessary risk management and leaves new risks unmanaged.

A common user experience can hide a change that increases an existing risk, even if the change ostensibly reduces risk. For example, a new UAV that autonomously compensates for wind is often considered to reduce risk. Even in windy conditions, to the operator the new UAV behaves as the previous one did in calm air. However, the new UAV only does so until the point that the wind speed becomes too great. If the user experience of the new UAV is the same as the previous UAV, this poses a risk. The operator was not compensating for the wind and may have therefore been unaware of it. When the autonomy begins to fail, the operator has much less time to react. A common user experience can also hide new risks. For example, failure in the autonomy system of the new UAV due to sensor or GPS errors may require the operator to intervene suddenly and in situations that would not be expected given their experience with the previous UAV.

Nevertheless, there are solutions. For the example of high wind, some UAVs have displays that show an estimate of the surrounding wind speed and warnings if mechanical limits are being approached. However, these are only useful if the operator is aware of these warnings and pays attention to them. This can be challenging if the operator's cognitive load is already high. In such a situation, policies, procedures, and training need to accompany the rollout of the new system, even if the user interface stays the same.

Attack Surface Challenges for Uncrewed Systems

Highly connected, intelligent uncrewed systems present additional attack surfaces compared to more traditional IT systems. A system's attack surface refers to the sum of all points, known as attack vectors, in the system where an unauthorized agent can perform malicious actions. Using the

[17] National Institute of Standards and Technology (NIST), "PSCR UAS Working Group, Cybersecurity and AI Risk Management for Uncrewed Aircraft Systems (UAS) in Public Safety," updated 2024-02-27, https://www.nist.gov/ctl/pscr/pscr-uas-working-group

https://doi.org/10.52202/075106-0003

analogy of a secured room in a building, a physical attack surface might include the doors, windows, walls, floor, and ceiling. Analyzing a system's attack surface ensures we consider how an unauthorized agent may enter. To continue the analogy of the secured room, while placing good locks on the doors and windows might be an obvious measure, good attack surface analysis might also highlight the need to check that an unauthorized person cannot enter the room through the ceiling or ductwork.

Like many IoT systems, uncrewed systems are mobile computing devices operated in public venues, often beyond an organization's logical or physical boundaries. Operating outside of the organization's usual boundaries exposes these devices to external and environmental threats that require greater situational awareness and cyber threat analysis. Furthermore, embedded computer equipment within uncrewed systems can be non-standard due to the need for specialized hardware and software to control the vehicle. The embedded equipment is often impossible to replace or upgrade and might run software that is impossible for central IT services to manage, making them blind spots in traditional cybersecurity monitoring. This void increases cybersecurity risk as uncrewed systems become more connected to critical organizational systems. Uncrewed systems can also threaten other organizational IT systems to which they are connected. This is particularly the case if their integration is ad-hoc and requires formal or informal deviations from IT security policy. As we discuss later, these new, less visible attack surfaces may require careful enumeration and active re-evaluation of security and privacy controls across the rest of the organization to determine how connected they may be to the Uncrewed System.

As discussed previously, a variety of general cybersecurity controls and guidance exist. They are still relevant for uncrewed systems but need to be adopted with a more complex attack surface in mind. For this discussion, we will focus on NIST SP 800-53. This document is used across the US Government for security and privacy controls for information systems (with the equivalent for non-government systems being NIST SP 800-171). Various other countries have equivalent frameworks and/or standards, but the core principle is a set of implementable control targets for information systems.

We will also use, as an illustrative example, the Drone as First Responder, or DFR, application that is becoming more common in public safety. While DFR can be implemented in a wide variety of ways, a common model is for one or more UAS to be pre-deployed on rooftops around the service area within automated unattended weatherproof pods. When required, the UAS launches autonomously from its pod, performs its mission with minimal human intervention, returns to its pod, and then charges, ready for its next mission. The DFR is often connected to other systems within the public safety organization, including the computer-aided dispatch (CAD) system, GIS, and collaboration tools. These connections to other critical public safety systems introduce risks that need to be addressed.

The controls in the NIST SP 800-53 can be customized to meet the diverse requirements of public safety organizations and stakeholders and are implemented as an organization-wide process. The control catalog consolidates requirements from various sources, including mission and business needs, laws, executive orders, directives, regulations, policies, standards, and guidelines. It also includes security and privacy controls from a functional perspective by breaking down topics into focused control families; see the full list in the table below.

ID	FAMILY	ID	FAMILY

AC	Access Control	**PE**	Physical and Environmental Protection
AT	Awareness and Training	**PL**	Planning
AU	Audit and Accountability	**PM**	Program Management
CA	Assessment, Authorization, and Monitoring	**PS**	Personnel Security
CM	Configuration Management	**PT**	Personally Identifiable Information (PII) Processing and Transparency
CP	Contingency Planning	**RA**	Risk Assessment
IA	Identification and Authentication	**SA**	System and Services Acquisition
IR	Incident Response	**SC**	System and Communications Protection
MA	Maintenance	**SI**	System and Information Integrity
MP	Media Protection	**SR**	Supply Chain Risk Management

Table 1: The families of controls in NIST SP 800-53.

We can first look outwards from the UAS towards the systems with which it communicates. To gain perspective on how these control families fit into a first responder IT system, one must break technology into individual tangible assets, policies, and processes to identify cybersecurity requirements. For example, UAS are no longer simple two-device systems consisting of an OCU and flight vehicle but are now a complex, connected ecosystem. This new model consists of a connected interface from the UAS and/or OCU to an internet-connected device, such as a smartphone, cellular modem, or router on a 4G or 5G LTE network. The cellular endpoint may then connect to multiple remote systems. It may also connect to one or multiple local and/or remote first responder dispatch systems; each dispatch system may then connect to multiple networks that provide other response services.

Many DFR systems are hybrid and include services requiring connections to external vendors and cloud-hosted systems. The number of entry points, hardware, software, network, and human interfaces exponentially increases for each connected system. Software sources and supply chains are also potential attack surfaces that must be considered in first responders UAS. Accordingly, the number of responsible parties involved in the operation and maintenance also increases with each connected system. The complex threat landscape heightens as each provider exports functions to another provider, and so on.

Next, as we continue to develop organizational risk policy, we look inwards at how the UAS operates to see where other parts of the attack surface might lie. For example, the UAV will have receivers for other radio signals, such as GPS and Automatic Dependent Surveillance – Broadcast (ADS-B), which allows the UAV to locate and identify other aircraft in the vicinity. While an

https://doi.org/10.52202/075106-0003

attacker is perhaps unlikely to be able to use these parts of the attack surface to penetrate the rest of the organizational network, this attack surface is still important to consider in terms of the security of the mission and the data gathered. For example, an attacker might spoof the GPS on the UAV[18], causing it to think that the collected data is from somewhere else, resulting in incorrect data ingested into the organization's system.

The UAS also possesses a wide variety of computers, most of which do not run conventional operating systems. For example, the UAV itself would have at the very least a main embedded CPU, perhaps a GPU for offloading machine vision and other AI tasks, and one or more embedded controllers that take care of lower-level systems like directly commanding the motors, coordinating the behavior of the battery, and configuring the software-defined radios. These systems run their own software, which can be changed through legitimate software updates or potentially malicious activity, and yet, as previously mentioned, they cannot be managed by traditional IT cybersecurity infrastructure.

This presents a problem for some traditional organizational IT policies, which mandate that, for example, all IT systems be centrally managed. When an IT policy that forbids unmanaged devices collides with the reality that the organization needs to deploy DFR, which often does not integrate into conventional central IT management systems, there is the danger that the new attack surfaces introduced by the DFR will simply be ignored. Instead, a more prudent organization would re-evaluate its IT policy, characterize these and other new attack surfaces, and deliberately allow these new devices onto the network while employing other forms of risk management.

Where is your AI?

A key step in managing AI risk, as highlighted in the Map function of the AI RMF, is to determine where AI presents a risk. For modern uncrewed systems, this is a non-trivial task, made all the more challenging by the often opaque, proprietary, and unusual nature of computing equipment in many of these systems. The complex nature of software supply chains, where even the vendor communicating with the end user may not be fully aware of the nature of the AI within their product, makes identifying risks more challenging.

Properly controlling this risk requires transparency on the part of the entire supply chain. In the absence of transparency, the first step is for all stakeholders to begin asking the right questions. Analyzing the Uncrewed System at each level of autonomy, from the lowest control levels to mission levels of autonomy, can assist in enumerating the places where AI might exist in a system. New features, particularly those that come with software updates and promise the ability to adapt behavior or optimize performance, often also have AI components. As we discuss in the last section, Summary and Future Work, we are also developing guidance and resources to help stakeholders to better understand where AI may be present in their systems and the ways to manage their risks.

Failures, Correctability, and Root Cause Analysis

Along with enumerating the different forms of autonomy versus automation at various levels, as we did in Section 2, it is also useful to classify the forms of autonomy at each level according to how transparent or opaque the underlying AI is, using our aforementioned definitions. This discussion becomes a sliding scale, where automation could be considered the most transparent

[18] Veillette, P. "The Serious Threat Of GPS Spoofing: An Analysis," Aviation Week, 2023-10-09, https://aviationweek.com/business-aviation/safety-ops-regulation/serious-threat-gps-spoofing-analysis

form of autonomy, autonomy based on transparent AI being the next most transparent, and autonomy based on opaque AI being the least transparent and least conducive to root cause analysis.

Knowing why an autonomous system performs a specific action and being able to predict what it might do can be vital in safety-critical applications. Sometimes a trade-off must be made. In some but certainly not all situations, autonomy based on opaque AI can outperform that based on transparent AI over some set of environments and tasks. This is because the environment and/or task contains important complexities that could not be modeled analytically or with more transparent techniques.

The obvious trade-off is that we do not know why it does or does not work. Less obviously, we also have a much more difficult time finding other situations where similar failures may happen. This difficulty poses a unique risk management challenge. A central requirement of risk management is the ability to perform root cause analysis on failures, determine situations where similar failures with the same root causes might occur, and put in place controls to manage that risk in the future.

For example, an autonomous vehicle might exhibit a behavior whereby it fails to see certain objects on the road. For more transparent AI systems, it may be possible to trace the reason for this decision and determine with some degree of certainty the extent of situations where similar errors might occur. Ideally, the system's behavior could be corrected in a verifiable manner. Even if this is not feasible, knowing the extent of the possible similar errors that the system could make enables other risk management controls to be applied.

While more opaque systems can be difficult or impossible to trace, many of them are able to learn from such mistakes. For example, they could feed the sensor data generated from the incorrectly missed objects back into the learning system so that the next time they are encountered, the system is more likely to perform the correct action. On the surface, this seems like a reasonable risk management step. However, it is important to know the extent to which such a "patch" corrects the underlying problem and that the "patch" has not caused new errors elsewhere in the system. For AI systems that trend towards the more opaque end of the spectrum, the answers to these questions can be impossible to know.

From a risk management perspective, it would be unwise to simply patch the behavior of the system for those specific objects. Without understanding exactly why the system could not see those objects, it will be impossible to know for sure what other objects it might also miss. This risk is one that is particularly unique to opaque AI systems.

Appropriately managing the risk of more opaque forms of AI typically requires additional controls beyond those used for more transparent systems. Examples include monitoring for such changes in the system's behavior (especially systems that learn and adjust their behavior regularly), with extensive testing and roll-back of policies, as warranted. Such controls are unlikely to be possible without extensive resources and the cooperation of significant parts of the software and data supply chain. It is an open question as to whether it is appropriate to deploy more opaque AI systems if resources and cooperation are unavailable.

As technology progresses, policies around the use of AI with different levels of transparency cannot be universal or static. The goal should be to manage risks relative to each other. On the one hand, it may be appropriate to take on a new, possibly less well-understood risk in the short term if the trade-off is a dramatic improvement in performance that reduces risk elsewhere. However, this decision needs to be flexible. If future systems can yield the same improved performance but with risks that are better understood, this decision may be worth revisiting even if there is a substantial cost in switching.

https://doi.org/10.52202/075106-0003

Misunderstood Explanations

Explanations and explainable AI systems are often touted as solutions to many risks inherent in AI systems. On the surface, this thinking might seem reasonable; however, it is important to carefully consider how explanations are used in the context of risk management and if the particular types of explanations offered by an AI system are compatible with the use case. In this context, we consider explanations that provide human users and investigators with accurate information that they can understand and is technically actionable. Indeed, the behavior of any computer program, including any AI system, can be explained at some level. However, for many ML systems, this explanation is not at a level that is understandable or technically actionable. Furthermore, many ML techniques, such as genetic algorithms and deep learning, depend on random number generators for their operation, thus their explanations inevitably have a probabilistic component to them.

There are many ways to divide the space of different types of explanations[19]. Perhaps the most important when it comes to AI systems are explanations that reflect the reason for the underlying decision, as compared to explanations that are consistent with the behavior of the system. The former might be considered "the" explanation for the decision, meaning that it is the singular true explanation. The latter might be regarded as "an" explanation for the decision, being an explanation out of many possible explanations that are consistent.

This distinction is vital when it comes to risk management. An explanation that is based on the underlying decision-making process, expressed in a manner that is understandable to humans at a suitable level of abstraction, allows the human to predict what the system will do and perform root cause analysis of failure.

In contrast, while there is only one explanation that reflects a system's underlying decision-making process, there are an infinite number of explanations that are consistent with any finite set of the system's decisions. While such explanations might be satisfying to human curiosity, it is dangerous to rely on them to predict the behavior of the system. Such explanations also do not allow investigators to perform root cause analysis (beyond a statistical analysis) or prevent future failures. Unfortunately, many AI systems that are sold on the promise of having "Explainable AI," particularly those that make use of gradient methods like Locally Interpretable Model-agnostic Explanations (LIME)[20] and Shapley Additive Explanations (SHAP)[21], produce "an" explanation and are incapable of producing "the" explanation.

The obfuscation of the distinction between "the" explanation and "an" explanation poses a considerable risk because, from the perspective of a human observer, it is impossible to determine if a given explanation is "the" explanation or "an" explanation. Risk management controls that are developed based on "an" explanation for how the system works run the risk of both unnecessarily restricting the behavior of the system and being blind to real risks. For example, an autonomous vehicle might have the ability to detect stop signs. There are many aspects of the sign that the system might detect, such as its shape, color, and markings. If the system fails to see a particular

[19] Sheh, R., and Isaac, M. (2018) Defining Explainable AI for Requirements Analysis. In KI-Künstliche Intelligenz, V32, PP 261-266, Springer Berlin Heidelberg

[20] Ribeiro, M. T., Singh, S., Guestrin, C., ""Why Should I Trust You?": Explaining the Predictions of Any Classifier," arXiv, revised 2016-08-09, https://arxiv.org/abs/1602.04938

[21] Lundberg, S., Lee, S., "A Unified Approach to Interpreting Model Predictions," arXiv, revised 2017-11-25, https://arxiv.org/abs/1705.07874

stop sign, it is vital to know exactly which aspect of the sign was missed. An explanation that happens to be consistent with the observed failures is not sufficient.

Human equivalence is sometimes cited as a reason for "the" explanation being too high a bar to expect AI systems to clear. It is true that humans are sometimes unable, or unwilling, to provide "the" explanation for why they do something. However, a primary control that society uses to manage the risk of not being able to perform root cause analysis on human behavior is to remove the human's ability to perform the task. For example, a driver who crashes often enough will lose the ability to drive, be it by having their license suspended or losing their insurance coverage. This control works because removing the ability of the singular human to drive does not prevent other humans from driving. However, for an organization that has deployed AI on a fleet of uncrewed systems that effectively have the same AI "driver," the equivalent would be to "ground" the entire fleet. The grounding of a UAS fleet is not a viable risk control to impose.

Risks of AI System End of Life

When commissioning any mission-critical system, one must consider managing organizational process continuity when the system reaches end of life, both planned and unplanned. Work must continue as the organization transitions to a new, future system, which might include migrating data and configuration, as well as training personnel. Measures often taken to ensure a smooth transition include having configuration and data documented and stored, ready for the new systems, and having organizational processes in place to develop and deploy the new systems. Local backups of critical information help to protect against disruption caused by unexpected end of life of the system, such as failure of the computer that it is running on or a cloud provider that is critical to the operation of the system going out of business.

AI systems that include ML components pose unique challenges. An AI system that has become an important part of the organization has, by definition, incorporated organization-critical information into its learned model. However, ML models are highly specific to a particular learning technique and often even to a specific system. If the system experiences a failure, there can be a significant risk that the learned model will not be useful for a replacement system.

In the context of autonomous systems, this risk can appear at any level of abstraction. For example, an autonomous vehicle may include a management computer that adjusts various control parameters as the vehicle ages and its underlying behavior changes. The computer compensates for these changes and makes the vehicle behave as close as possible to a newer vehicle. If the management computer fails, the replacement computer will not have this learned information unless the old model was backed up prior to failure and is compatible with the replacement computer. If restoration is not possible, there is a risk that the vehicle may behave in an unexpected manner until it has had a chance to relearn its various parameters. This risk is particularly pronounced for situations where the learning controller is used to address relatively unusual situations, such as operating close to mechanical limits, that may not be encountered in post-repair testing.

In the ideal case, this AI risk might be controlled by dictating that equipment that contains ML models must have procedures for such learned models to be migrated to replacement equipment. As a fallback, visibility into this risk means that procedural controls can be put in place, such as ensuring that systems being brought back online after repair also have appropriate retraining. It also becomes especially important for post-repair testing to fully exercise the system. This can take considerable effort, particularly for systems that rely on ML models to address highly complex environments and tasks.

https://doi.org/10.52202/075106-0003

Another example is an autonomous vehicle that relies on navigational path planning from a cloud provider which routes the vehicle according to weather and traffic. The cloud-hosted path planner might also learn optimal paths based on organizational usage patterns. If the cloud provider experiences a disruption or goes out of business, it is unlikely that the learned data can be migrated to a new service. At the very least, comprehensive requirements for the cloud-hosted service would be needed to find, develop, and/or configure a replacement service. There may also be a significant period of degraded performance as the new service relearns the usage patterns that may also pose risks to the organization.

Greater than the Sum of their Parts

Thus far, we have considered cybersecurity and AI risks individually. It is also essential to consider them together, including how AI can change the management of cybersecurity risks and vice versa. This field is still nascent with little existing guidance considering both factors. In this section, we highlight a few examples.

The Dangers of Inexplicability. In today's world of complex IT systems, it is almost expected that IT systems will behave in ways that users do not expect or understand. With the increasing adoption of AI systems, particularly opaque AI, not only can the system behave in a way that is unexpected, but it can also be impossible to determine why the system behaved in that manner. This property is sometimes referred to as inexplicability.

This presents an AI risk in its own right, for it becomes impossible to perform root cause analysis on the failure or determine how to avoid it in the future. For instance, an uncrewed system that is performing surveillance might consistently miss unusual events in a particular sector. In some cases, it may even be impossible to determine if the system failed or if it simply had more information than the user. For example, an uncrewed vehicle might autonomously choose to take a route with which the operator disagrees.

It also presents a cybersecurity risk as detection of malicious activity is a cornerstone of cybersecurity risk management. If the system is known to behave in an unexpected manner and, in some cases, changes its behavior over time as it learns its environment and task, it becomes much more difficult to distinguish legitimate activity from malicious activity.

Worse yet, the presence of systems that behave in inexplicable ways raises the tolerance of people in the organization to inexplicable and unusual behavior more generally. An attacker who is aware of this, particularly on the inside, can more easily deflect questions if unusual activities are detected by blaming the behavior "on the AI," potentially even for systems that have nothing to do with AI.

Attacking the AI. The underlying algorithms of the AI systems present a wide range of attack surfaces that have only recently been studied. These attacks come about largely because AI systems understand the world in a way that is quite different from the way most humans do. This difference in understanding means that an attacker can create situations that seem normal to a human observer and yet can trick an AI into making the decision that the attacker wants it to make.

Adversarial Machine Learning[22] refers to the field of study concerning tricking AI systems containing ML components, by influencing the inputs into the system, so that it makes a decision

[22] Vassilev, A., Opera, A., Fordyce, A., Anderson, H., "NIST Trustworthy and Responsible AI, NIST AI 100-2e2023, Adversarial Machine Learning, A Taxonomy and Terminology of Attacks and Mitigations," National Institute of Standards and Technology (NIST), 2024-01-02, https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-2e2023.pdf

that it should not make. For example, uncrewed ground vehicles on public roads must read a wide variety of street signs. In modern systems, this is often accomplished through deep learning, which learns patterns that correspond to symbolic concepts such as stop signs. Crucially however, the learning takes place in a way that cannot be comprehensively verified. It has been shown such systems can be tricked into erroneously identifying one sign as another, with simple, unobtrusive modifications to the sign, such as applying small stickers. A human observer would not interpret the stickers as a change in the meaning of the sign; however, an ML system may misread the image.

Certainly, this is not a property that is unique to ML systems or AI systems in general. Humans are also susceptible to inputs that have been manipulated to make us think or see something that is different from what is there. In the human field, these are known as illusions. Adversarial Machine Learning could be considered the field of study concerning finding illusions that ML systems suffer from and particularly ones that are not immediately obvious to humans who might be monitoring the system.

This parallel with human illusions also suggests a potential preemptive risk management control. For safety-critical tasks, such as humans operating motor vehicles, the environment is designed to minimize the probability and impact of illusions. For example, traffic cones are a bright color and tend to not resemble anything else that one expects to see and might find confusing. Even if someone paints something on the traffic cone, a human is unlikely to confuse it for something else, and even if they did, there are generally many cones marking an area. For ML systems, the equivalent would be determining where such "illusions" might occur and structuring the environment accordingly. This can be difficult as different ML systems can suffer from very different illusions. More transparent AI systems, while still likely to suffer from illusions, may permit greater analysis as to why the illusions happen and how they might be proactively prevented.

There are also attacks on AI that do not involve machine learning but can be equally impactful. One well-known example relevant to both crewed and uncrewed systems was demonstrated in 2020 when an artist dragged 99 mobile devices in a wagon at slow speed across a bridge, tricking Google Maps into reporting a traffic jam on the bridge[23]. While humans would understand a traffic jam to be numerous cars moving slowly, Google Maps can only detect the position of users' mobile devices and makes the assumption that their motion near roads corresponds to the motion of the cars. Crucially, this difference is relatively unknown to users of Google Maps, making it possible to then trick users, including any autonomous uncrewed system that incorporates traffic data from Google Maps, into thinking that there is a traffic jam. An attacker, knowing this, could fake such traffic jams to discourage vehicles from passing through a particular area.

Governance, Social and Regulatory Consequences

Risk management governance is a control that ties organizational objectives to processes, protective assets, and configurations. As such, it must consider and manage the specific technologies used within an organization in the context of that organization itself. Industry best practices and frameworks available, including those of NIST, are excellent resources so that organizations have a known baseline. However, appropriate application is still required for the individual context, particularly when it comes to these new technologies.

In the context of uncrewed systems, there are wide-ranging and novel applications. The challenges already discussed with the application of an existing framework, toolset, or otherwise to a new and novel application also exist in the domain of governance. The governance flavor to

[23] Barrett, B., "An Artist Used 99 Phones to Fake a Google Maps Traffic Jam," Wired, 2020-02-03, https://www.wired.com/story/99-phones-fake-google-maps-traffic-jam/

https://doi.org/10.52202/075106-0003

these includes different societal expectations of new technologies, which flow into additional expectations on an organization using technologies (e.g., AI manipulation of images). New technology also increases the scope of impact of a consequence far beyond the traditionally expected bounds of an organization (e.g., accidental poisoning of a shared AI data source).

Without governance review cycles factoring in technologies such as AI and automation, there is also the risk that processes and procedures designed to allow for sound procurement decisions, consistent technology use, incident response and review, and other standard cyber practices are no longer possible because the technology is inherently different 'under the hood.' At the operational level, processes, procedures, and plans need to factor in the capabilities of the new technologies and changes to the operating environment to allow the technologies to function fully.

Using the previous example of allowing additional public network connectivity from a UAV, it may be necessary to change monitoring processes for network intrusions and review risk registers, metrics, and reporting elsewhere in the organization. These actions allow for the application of risk management controls that factor in the new and expanded threat surface related to external connectivity and the possible malicious actions that an adversary could perform, throughout the organization, should they compromise the UAS. Governance controls related to continuity and disaster recovery would also need to be reviewed for allowing connectivity to public networks.

A further aspect that governance needs to consider is the potential capabilities of the new technology in the context of social and regulatory consequences. Just because an organization elects not to use the full capabilities of a technology does not mean it is protected from public scrutiny with potential reputational damage. Nor does it protect the organization from any regulatory controls related to that capability, even if not in use. Organizations must also consider these environmental factors when reviewing governance when integrating new or novel systems into their technology ecosystems.

On the bright side, this review of governance is also beneficial for an organization in achieving improvements in processes and procedures related to the use of the new technology. The process of reviewing the capabilities of a new technology (or novel application) against an organization's objectives for the purpose of identifying and managing risks also allows an organization to develop the most effective and efficient processes for operating and utilizing available capabilities. This improves efficiencies and return on investment (or confirms that the product is not suitable).

Immediately Actionable Risk Management

Cybersecurity

The CSF 2.0 and associated cybersecurity guidelines, such as SP 800-53, help vendors and end users to identify systematic risks in uncrewed systems. It helps users and vendors to divide the system up and, in the context of the six CSF 2.0 functions, determine at a general level where risks lie and how risks might be addressed. This helps to create a healthy cybersecurity ecosystem. However, the guidelines only help if organizations reevaluate their attack surfaces in the presence of these new technologies.

While more specific guidance is being developed, users can employ general NIST guidelines, such as the CSF 2.0, to get a head-start on identifying cybersecurity implications that they may encounter when adopting new technologies. End users with existing or new information technology infrastructure can systematically identify each of the NIST SP 500-53 controls and assign them to responsible individuals.

Organizations can also use this as an opportunity to take a broader view of their cybersecurity policies and procedures to ensure that they are harmonized or, at the very least, non-contradictory.

Organizations can delineate clearer boundaries between their different policies to avoid confusion as to which policy might apply to the "gray zone" that often exists with these increasingly connected devices.

Artificial Intelligence

Organizations can similarly begin to evaluate their AI risk exposure using the AI RMF, even in the absence of uncrewed systems-specific guidance. Until the supply chains for data and software become more accustomed to rigorous AI risk management, there will be a high degree of uncertainty around which particular systems expose an organization to risk and how this risk is manifested. Agencies like NIST are working to provide additional resources and guidance to all stakeholders, with the goal of providing end users with more detailed information to make risk management easier.

In the meantime, organizations will need to do their own research and investigations to determine how to apply these general guidelines to their uncrewed systems. Later in this section, we present a list of questions derived from a workshop of uncrewed systems stakeholders, held in February 2024. Organizations may use these questions to assess their AI risks and tolerances, as well as to educate supply chain players on the need for greater transparency on the underlying AI systems. This increased transparency better informs the end users' risk management decisions.

Organizational Governance

Organizations may leverage existing governance to improve risk management visibility. and therefore, management for UAS.

Organizations, with or without an uncrewed systems cybersecurity or AI subject matter expert, can perform a minimal assessment of this technology through any technology risk management governance process. These processes include third-party risk management governance, appropriate procurement governance, appropriate change management governance, access control and authorization (including non-humans, such as the Uncrewed System itself), and other applicable governance processes. Using existing governance processes would go a long way to improving the visibility of AI risks while waiting for more specific guidance.

Another way to consider uncrewed systems in the context of existing governance processes is to view the AI system as an "unknown" human performing those tasks. While an imperfect analogy, it would be a useful process to run any known human factor risk governance process over the use case of an untrained human performing those tasks to identify and then model potential risks that may require additional controls or management.

Identifying Risks

During the Workshop on Cybersecurity and AI Risk Management for Uncrewed Aircraft Systems in Public Safety, held in February 2024, an interactive session was held to solicit the top concerns that the attendees had regarding cybersecurity and AI risk. Attendees were first invited to submit suggestions for questions that they felt most important. These were collated and then, in a second round, the attendees were asked to vote for their top questions. Most of the attendees were public safety end users, followed by government representatives and researchers. The resulting list of questions is presented below. We stress that this list is not comprehensive. It represents areas of greatest concern to the attendees, however it also highlights possible blind spots in the end users' understanding of these risks.

This list helps to identify the top concerns these stakeholders have in managing cybersecurity and AI risk. The results represent some of the top questions that those who are managing or procuring uncrewed systems should consider when analyzing their risk exposure, including how

https://doi.org/10.52202/075106-0003

the Uncrewed System might present risks that differ from those of more conventional IT systems. It also represents aspects of the cybersecurity and AI risk management challenge for which there is already some base level of understanding and the greatest likelihood for traction in the immediate term. Knowing the answers to these questions, or knowing that the questions could not be answered, and having discussions inspired by the asking of these questions are important first steps to applying the aforementioned resources, such as the CSF 2.0 and AI RMF.

The top 10 questions are presented below, adjusted slightly for relevance to the broader uncrewed systems domain.

1. How secure is the system? What is the attack surface? How, and in what timeframe, will we be notified of breaches of the various severity levels? What cybersecurity stress testing or "Red Teaming" has been performed, on both the system and systems on which it depends?
2. How and where is the data stored? What is the level of encryption and other security methods for the data and its derivatives, such as metadata, error logs, vendor telemetry, and temporary files, in transit and at rest? Who can decrypt or modify the data?
3. How can we determine if the AI is giving incorrect information? What are the possible mission-critical failure modes of the system? Do we have enough access to the system to determine if a decision that we do not agree with is due to an error on the part of the AI system or to otherwise perform root-cause analysis? How can we correct the behavior of the system?
4. What measures are taken to ensure an appropriate level of data privacy? What are the levels of data privacy that have been chosen and how were those decisions made?
5. In what physical locations is the AI processing being done? For instance, is the AI processing occurring on the Uncrewed System, on the controller, on our own servers, or in a cloud data center in-state, in-country, or elsewhere?
6. What AI and cybersecurity challenges have other users run into with this system and what steps have they taken to mitigate risk?
7. Is information about the confidence that the AI system has in its decision available to the operator and/or investigator? What information is available to assist in interpreting the confidence level of this decision?
8. What will the system do if it encounters a situation that is unusual or unexpected? Is the system capable of detecting if it is operating in a situation that was not represented during its development, training, or testing?
9. Where does additional input data to the system, such as maps and positioning, come from; how is it updated and authenticated; and what measures are in place to address intentional and accidental interference, corruption, jamming, spoofing, or similar?
10. What measures are taken to ensure continuity of critical operations in the event that the system undergoes planned or unplanned downtime or end of life? How likely is it that the information and systems required for recovery are affected by the same cyber-attack as the main system? If the AI system suffers a failure, such as effects from a bad model or input data, what plans exist to roll back to a good model?

Summary and Future Work

Cybersecurity. Artificial intelligence. Uncrewed systems. All three of these represent concepts and technologies that have individually resulted in major disruptions to the field of risk management and governance at different points in very recent history. The risks, particularly in less

common use cases, are poorly understood. Very little guidance or precedent exists for the management of risks when these three come together outside of specific, well-resourced sectors.

We provided an overview of some key definitions, concepts, and resources in these rapidly changing domains. Our coverage of these topics is not comprehensive, as the field is too broad for the scope of this paper. Instead, as a starting point, we described a few key, less obvious risks in more detail. We also provided some actionable steps, mostly around gathering information by asking key questions and initiating discussions with other stakeholders, particularly within the software and data supply chains.

The NIST PSCR Working Group on Cybersecurity and AI Risk Management for Uncrewed Aircraft Systems in Public Safety is working to identify areas of most need for resources. The goal is to provide targeted guidance to bridge the gap between conventional risk management and governance and the needs of the Uncrewed Aircraft Systems in Public Safety application. These resources will also have relevance to uncrewed systems and other applications, particularly where there is already a strong risk management culture.

A high priority is the further development of the list of questions that those in charge of managing the procurement and use of uncrewed systems can ask to better understand risks. Associated with this will be the development of guidance that explains each of the questions in greater detail, along with commonly expected answers and the implications of the various answers for the organization's risk management and governance.

Parallel work will be to continue consultations with a broad range of stakeholders, including users, software and supply chain entities, and those representing societal concerns, to determine what other questions have the most impact on improving the state of risk management for uncrewed systems. This effort will also seek to raise the general level of awareness of these novel and less well-understood risks and to identify additional resources that may be needed. We invite all stakeholders to join us in this work.

These efforts will also feed into the development of joint profiles for both the NIST AI RMF and CSF 2.0. While these efforts focus on the Uncrewed Aircraft Systems in Public Safety application, the guidance will also be relevant to broader classes of uncrewed systems and applications beyond public safety.

https://doi.org/10.52202/075106-0003