Carnegie Mellon University
Software Engineering Institute

MACHINE LEARNING IN CYBERSECURITY: A GUIDE

Jonathan M. Spring
Joshua Fallon
April Galyardt
Angela Horneman
Leigh Metcalf
Edward Stoner

Month and Year (date added at time of publication)

TECHNICAL REPORT
CMU/SEI-2019-TR-005

http://www.sei.cmu.edu

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

This report was prepared for the SEI Administrative Agent AFLCMC/AZS 5 Eglin Street Hanscom AFB, MA 01731-2100

DM19-0312

Table of Contents

List of Figures

List of Tables

Acknowledgments

The authors gratefully thank Jay Kadane, Patrick McDaniel, and Lena Pons for helpful comments on prior drafts of this report.

Abstract

This report lists relevant questions that decision makers should ask of machine-learning practitioners before employing machine learning (ML) or artificial intelligence (AI) solutions in the area of cybersecurity. Like any tool, ML tools should be a good fit for the purpose they are intended to achieve. The questions in this report will improve decision makers' ability to select an appropriate ML tool and make it a good fit to address their cybersecurity topic of interest. In addition, the report outlines the type of information that good answers to the questions should contain. This report covers the following questions:

1. What is your topic of interest?
2. What information will help you address the topic of interest?
3. How do you anticipate that an ML tool will address the topic of interest?
4. How will you protect the ML system against attacks in an adversarial, cybersecurity environment?
5. How will you find and mitigate unintended outputs and effects?
6. Can you evaluate the ML tool adequately, accounting for errors?
7. What alternative tools have you considered? What are the advantages and disadvantages of each one?

Introduction

This report focuses on machine learning (ML) tools for artificial intelligence (AI), and—specifically—about how to ensure that these tools are useful when applied to address a cybersecurity problem.[1] The goal of the report is to assist managers and decisions makers who are considering employing ML for some cybersecurity purpose.

For the purposes of this report, we define ML and AI narrowly to guide the discussions that follow. We define ML as a set of statistical tools that analyze data to infer relationships and patterns. Ideally, the relationships and patterns inferred by ML will lead to a useful model of the object or phenomenon that the data describes. With respect to AI, we define it as a software agent that takes actions based on its environment. The goal of AI, therefore, is not to make the sci-fi dream of creating a thinking robot into a reality. Rather, its goal is to couple a tool, such as an ML tool, with a controller that can take actions based on the tool's output. You can also use tools such as logics and expert systems to implement AI.

Error! Reference source not found. captures four important aspects of an ML tool. An ML tool is *trained* on a body of observations, usually by calculating statistical parameters of the properties of these observations in a given context and in relation to prior observations. Based on these statistics, tool developers make predictions about a topic of interest. Then, the developers test these predictions and refine the tool using results from this testing. Observations need to be reliable and transparent as well as relevant to the topic, and this paper provides guidance for decision makers and managers to collect the information they need to ensure that the ML tool they use meets these requirements.

[1] ML and AI are becoming popular tools for addressing cybersecurity problems. However, this paper is not a tutorial on cybersecurity or ML. For more information, US-CERT provides a summary of cybersecurity concepts [US-CERT 2019]. For a more thorough introduction, refer to *Security Engineering: A Guide to Building Dependable Distributed Systems* [Anderson 2008]. For an introduction to ML, the Machine Learning Wikipedia page is reasonably accessible [Wikipedia 2019]. For other quality introductions, see Andrew Moore's tutorial web page [Moore 2019] or Andrew Ng's online course [Ng 2019].

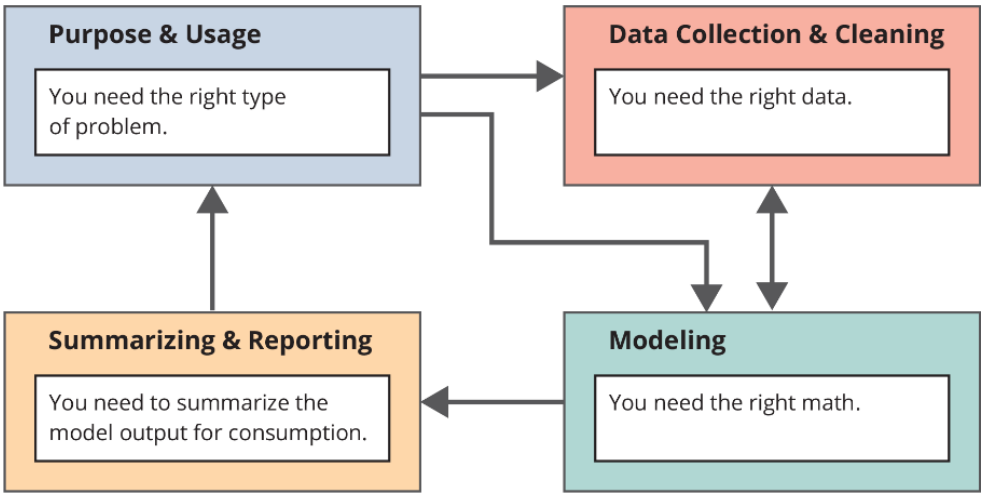

Figure 1: Relationships Between Parts of an ML Tool and Its Use

ML tools do not generate topics for inquiry, but must be employed to address a question that you generate regarding a topic of interest or a business need. You can then evaluate an ML tool for suitability to that topic or need. In the sections that follow, this report outlines seven key questions that we designed to help structure that evaluation in the context of addressing a topic of interest or business need in the field of cybersecurity.

Seven Guiding Questions about ML Tools for Cybersecurity

The most appropriate question to ask about a tool is whether it is a good fit for its intended purpose. In the following sections, we expand this general question into a series of seven more tractable questions that are relevant to cybersecurity applications. It is important to note that we are not trying to find or define the optimal or best tools for cybersecurity, but, rather, satisfactory or good-enough tools.

This report presents the seven questions in the order you should ask them. The first and final questions of the series help frame an evaluation to determine whether a tool is suitable for your needs, and they are questions that the managers and decisions makers who are employing the tools should ask of themselves. The middle five questions are for decision makers and managers to ask the ML tool developers or suppliers they are working with. We designed those middle five questions to help decision makers get the information they need to choose the right tool and develop it correctly.

You can refer to **Error! Reference source not found.** above to review the relationships that structure your inquiry, and the questions and discussion in the sections that follow will help you carry it out. Question 1 will help you establish the purpose of your inquiry, which constrains both the data and the ML models[2] that will be adequate to fulfil it. The fourth aspect—reporting—is the focus of question 3. Questions 2 and 5 focus on data, models, and their relationship. The other questions discuss topics that are closely related to the structure of the inquiry. Question 4 focuses on defending the tool, its data, and its models from attack; question 6 focuses on evaluation of the tool as a whole; and question 7 focuses on comparing different tools to evaluate their fit for your project.

The discussions in the sections below center on providing guidance about what kind of information constitutes a good answer for each question. The report does not detail how to produce satisfactory answers to these questions, but how to evaluate whether an answer is satisfactory.

Table 1 introduces the questions in the order you should ask them. It also provides a summary of the information we cover in greater detail in the sections below about the information good answers should contain.

[2] An ML *model* is, roughly, the mathematical structure the ML tool uses to produce its output from the data.

Table 1: Information That Good Answers to ML Questions Should Contain

1. What is your topic of interest?	The aim of this question is to establish the goal of your investigation. A good goal should address specific cybersecurity topics that will guide how you apply the tool, such as the impact or implementation of a specific security policy.
2. What information will help you address the topic of interest?	The aim of this question is to establish what information you will use to drive your investigation. A good response should demonstrate that the input data includes or encodes features that allow meaningful assessment, such as prediction or classification.
3. How do you anticipate that an ML tool will address the topic of interest?	This question seeks to address the applicability and transparency of the tool. You must choose the ML tool carefully so that it will output appropriate information at a high enough standard to evaluate it.
4. How will you protect the ML system against attacks in an adversarial, cybersecurity environment?	The purpose of this question is to evaluate the defensive disposition of the ML system. A response should describe what protections the tool itself has, as well as how the data it uses and produces is protected during both training and operation. Additionally, the response should address the measures that exist in the environment surrounding the ML tool that makes it resilient if an adversary successfully attacks it.
5. How will you find and mitigate unintended outputs and effects?	This question addresses considerations for handling sensitive information carefully to avoid introducing both errors and bias into an ML system. A good response should consider the following five principles: representation, protection, stewardship, authenticity, and resiliency.
6. Can you evaluate the ML tool adequately, accounting for errors?	A good response to this question should plan an evaluation that assesses all of the following items in detail: data sources; design of the study; appropriate measures of success; understanding the target population; analysis to explore missing evidence; and the expected generalizability of results.
7. What alternative tools have you considered? What are the advantages and disadvantages of each one?	A fair answer to these questions should compare multiple types of tools. You should consider cost of development, maintenance, and operation. Since these are cybersecurity tools, you should consider how an adversary might respond to them.

The answers to all these questions require *good evidence*. Collecting good evidence requires structured observations, such as experiments and case studies that you design to reduce mistakes and errors. Reliable and robust methods of reasoning are also important for collecting good evidence. The research methods used in the sciences are a good source of guidance for study design and reasoning [Spring 2017]. In cybersecurity, researchers and analysts should interpret evidence knowing that an adversary may influence the decision-making process [Horneman 2017]. These considerations help us avoid messy situations, such as the one outlined in Figure 2, and move towards better, more intelligible tools.

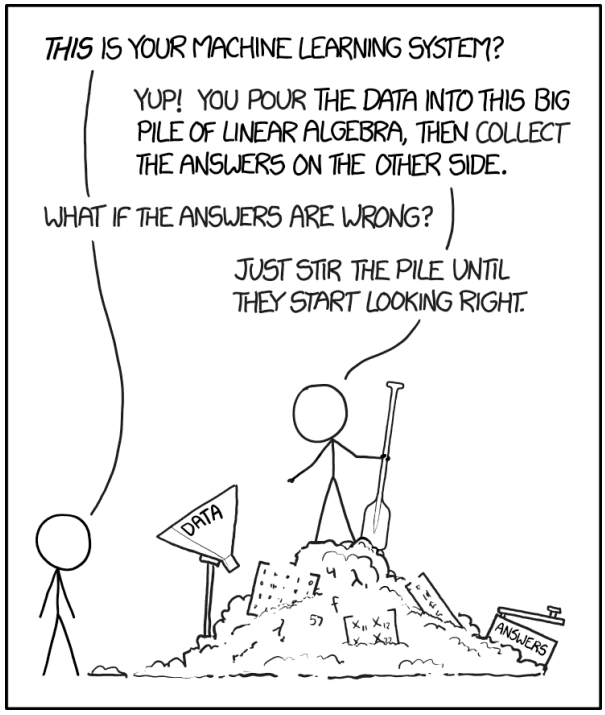

Figure 2 "Machine Learning" by Randall Munroe [Munroe 2017] CC BY-NC

1. What is your topic of interest?

This first question—which managers and decision makers should ask of themselves—is important for at least two reasons. First, it helps establish the purpose of the tool so that you can evaluate the other questions. Second, as discussed in the introduction, ML tools do not generate topics for inquiry. Rather, you are the one that must pose a useful topic for a tool to address. A proper description of your topic of interest should address a problem you want to solve.

In a business context, your topic of interest likely addresses a need of your organization. But a topic of interest should be specific. A topic as broad as "improve our security posture" should be broken into component parts. For example, a good topic might address whether a given attack on the organization is similar or related to certain prior attacks. Such a topic addresses the broad cybersecurity need of analyzing malicious campaigns. More importantly, a topic is a good candidate for analysis by an ML tool if, given appropriate context, data on the organization's network contains features that enable an informed response to the question that the topic seeks to answer. A complete problem specification should contain descriptions of the tool's available actions, what

each action does, a test of success, and the difficulty or cost of each action [Russel 2010, chapter 3].

On the other hand, common examples of ambiguous topics include "find unusual behavior," or "is this computer behavior weird?" These questions provide no insight into what you actually want to accomplish. While often used as euphemisms for security violations, anomalies (i.e., something unusual or "weird") might not always constitute security violations in practice. In other words, these examples do not provide enough information about your topic's goal to be able to select data that contains the right information or features to deal with the problem. A good topic should indicate the criteria you can use to measure the success of the solution. To improve ambiguous topics and provide measures of success, you should specify the relationship between specific past anomalies, future anomalies, and security violations.

After you settle on your topic of interest or need, you can start thinking about what you want to do about it. The first step in addressing your cybersecurity topic will be to identify what kinds of information you should have. The next section will help you ensure that you have adequate information available, which is vital to confirm before you can ask how an ML tool will help you process that information.

2. What information will help you address the topic of interest?

This is the first in the series of questions that managers and decision makers should ask the suppliers or developers of an ML tool. You should expect a response to include a list of the available information that you can gather and its sources, as well as an explanation of why that information is adequate for your topic of interest. The response should also ensure that the gathering of information meets ethical, legal, and privacy responsibilities. Understanding what information will help address your topic of interest is not a call to evaluate whether you can capture the right data fields. Rather, it is intended to make sure that the right kind of information is available in the first place, that the available data is sufficient in breadth and quality, and that you can access and use the information ethically.

The available information must characterize the topic of interest and its context. For example, if an ML tool is going to make a detection decision about whether certain code is malicious, training data should include the organization's security policy as well as other data that is relevant to it.[3] Including such data is important because the security of an information system is relative to the security policy of the organization that uses it. Security policies are not embedded in computer code per se, so evaluating whether software is malicious requires additional information beyond the software itself. Furthermore, software may behave in ways that violate one organization's policy but not another's. Or, certain software may violate an organization's policy in one context but not in another. Therefore, a question of interest that involves correlating code snippets with a specific organization's well-defined security policy, within a margin of error, would probably lead to a reasonable ML and AI task. On the other hand, the question, "Is a given piece of software malicious in general?" is an example of an approach that is not useful because the input data does not

[3] For definitions of terms such as "security policy," see the Internet Security Glossary, Version 2 [Shirey 2007].

have the right type of information—namely, a security policy context to define the concept of maliciousness.

In addition to ensuring that available data is relevant, you must ensure that it is adequate. Information can be inadequate if (1) training data is missing or poor; (2) there is not enough data to cover each relevant context; and (3) the data does not represent the topic of interest or the deployment environment. The changeability of adversary tactics and behaviors may be a common cause of all three of these types of inadequacy. What an adversary did last month does not necessarily predict what they will do next month. Gathering data, labeling it when needed, and continually investigating and learning about the cybersecurity field, therefore, will be ongoing needs for maintaining your ML tool. The answer to this section's question should address not only how you can meet the data need initially, but how you will meet it throughout the life of the ML solution.

Measures of adequacy change based on your problem as well as the type of ML tool you use. The question in the next section will help you ask about the problem type and refine whether the ethically available information is, in fact, relevant and adequate.

In addition to being adequate, relevant information needs to be *ethically* available. You should consider ethics and privacy at multiple layers of an ML and AI tool's development and use. For example, accessing data must not violate anyone's right to privacy. If the ML tool will operate on data about citizens of the European Union, for example, then you are expected to use strong privacy controls as described by the General Data Protection Regulation (GDPR).[4] Other jurisdictions have different obligations, but ethical use of data should involve genuine consent from relevant people, not just abiding by legal obligations. ML tools have documented privacy leaks and flaws [Papernot 2018], and such privacy leaks appear to be baked into the underlying formalism [Yeom 2018]. Therefore, adequate privacy in ML tools is a subject of ongoing research and cannot be guaranteed. You should embed AI and ML tools in a systems approach that identifies and mitigates risks during the entire process of data collection, training, processing, use, and storage (see question 4 for further discussion about protection).

After you refine your understanding about what information is relevant and ethically available for your purpose, the next step is for you to ask what a successful response from the ML tool should contain. An ML tool will process your information to address your topic of interest, but the features you want the response to contain will place important constraints on the tool, its data, and its model.

3. How do you anticipate that an ML tool will address the topic of interest?

This third question of the series is also for managers and decision makers to ask the suppliers or developers of an ML tool. An adequate answer should include the following three items: (1) a description of the tool's applicability based on your goals; (2) a description of what sort of results to expect; and (3) a description of the kind of explanation that the tool's output, or the suppliers and

[4] For more information about GDPR, see the official European Commission resource [European Commission 2019].

developers themselves, will provide to explain how it works and ensure it meets your needs. Because of the length of the discussion for this question, this section is broken into three subsections to address each of these items.

Relatively speaking, this question is easier to answer if the topic of interest is about association or observation. If the topic of interest concerns intervention or causal reasoning, the question becomes more difficult to answer. Additionally, for the ML tool's outputs to be adequately explained, the explanation should share certain characteristics with the explanations any human expert might provide about the services he or she offers.

Item 1: Applicability

When it comes to the first item concerning applicability, there are two issues you should consider to determine whether you can apply the ML tool to your topic. The first issue is to determine whether the way the ML tool works is relevant for addressing your question overall, and the second issue is to determine the correct type of ML tool to use to meet your goals.

For the first issue, you can help determine which tools are appropriate by thinking about the question words in your topic of interest—words such as "what is," "what if," and "why." ML models (recall **Error! Reference source not found.**) can be well-suited for answering questions about what something is, associations, or observations [Pearl 2019]. A question such as "Is this email spam?" is a promising start because it asks a specific "what is?" question. However, answering "what if?" and "why?" questions requires more than an ML tool alone. These questions require carefully structured data collection (experimental design) in addition to a statistical or ML tool. Such questions commonly form part of cybersecurity topics of interest. For example, an incident analyst might shape a plan to fix their network by asking "what if this system is infected?" However, ML is not likely well-suited for these types of situations [Pearl 2019]. For ML tools to impact any such cybersecurity situation, you should have a careful strategy to bridge the gap between the cybersecurity topic of interest and the questions that ML tools are well-suited to answer. Better tools might include deployable logical reasoning [Klein 2010] or structured general knowledge [Spring 2018] to support counterfactual reasoning.

To address the second issue, you should ask whether the type of ML tool will be useful for your purpose. Two important types of ML tools are those that use supervised learning and those that use unsupervised learning. In *supervised* learning, an ML tool learns to make predictions from properly labeled examples provided to it. In *unsupervised* learning, an ML tool finds patterns without explicit feedback, such as creating clusters of items that are somehow related [Russell 2010, chapter 18].

There are many differences between the two types of ML tools, such as what questions they are suited to answer, what information needs they have, and what their computational costs are. One difference between the two is that supervised learning requires labeled data. It is important to highlight this small difference because it has outsized importance when it comes to cybersecurity. For example, if you want your tool to predict when network traffic is malicious, you need to be able to train your model on a large sample of traffic where each packet or flow has been labeled as malicious or not malicious. It is important to reflect on how these labels might be generated and where you might get them. Your organization is unique in important ways, and your traffic will be

different than that of other organizations, so reusing someone else's labeled data can be problematic. In cybersecurity domains, accurately labeled data that is representative is often not available. If you want to deploy a tool with supervised learning, check the answers you received for question 2 and ensure you can acquire sufficiently high-quality labels.

Item 2: What Results to Expect

The second item that a good response to this question should contain involves getting a description of what results to expect from the ML tool. This description should help you better understand the relationship between your topic of interest and the output of the ML tool so you can be in a better position to assess whether it will be the right fit for your purpose. To that end, consider asking the following three questions:

- Will you be able to use the output directly for your purpose or will it need to be used as input to another process?
- What sort of accuracy can you expect with respect to how you anticipate using the results?
- Can you expect the tool to produce results within your time constraints?

Item 3: Explanation of How the Tool Works

The third and final item that a good response to this question should contain involves getting an explanation of how the tool works and whether the tool's output will provide enough information for you to understand why it worked the way it did. To determine whether the tool's output will provide enough information for you to understand it, it is useful to reflect on what we expect of human experts when they explain their decisions. When human experts provide advice or a recommendation, we often expect one of two things: (1) an immediate demonstration of success or (2) an explanation of their decision-making process that you can understand without having to become an expert yourself. An ML tool should be able to provide an explanation of its output that meets one of these criteria.

The first of the two criteria requires that the service or expertise be immediately testable. For example, if a technician fixes a broken device, such as a car or a clock, then the proof of his or her expertise is immediate—the device either works or it does not. However, the requirements are different and more difficult to meet with the second of the two criteria, when you can't test results immediately. For example, if the layperson wants to know something that is not immediately testable, such as whether a car will continue to work for the next five years after a certain repair, then he or she needs more information. The system that we currently have is that the expert should provide an explanation that may be oversimplified but has at least two important features: (1) it is at an adequate level of detail to transmit what the layperson wants to know when that information is not immediately testable, and (2) it is transparent enough to be auditable by other experts. These features essentially constitute a social system of accountability between laypeople and a network of experts, though in some places it is bolstered by legal conventions such as malpractice.[5]

[5] For a discussion on the topic of expertise and further references, see "How can the Public Assess Expertise?" [Douglas 2018].

You should apply these expectations to explanations about ML tools as well. For example, ask whether the results of the ML tool are immediately testable or whether their impact will occur in the future. In the first case, explanation is less important than adequate tests. Questions 4 and 6 below will help you ensure that you have adequate testing. In the second case, you must evaluate the ML tool's decision-making process as you would an expert explanation.

If the output from your ML tool will require an explanation rather than demonstration of success, the explanation may come from the designers or vendors of the tool rather than as part of its output. However, if they are required to produce the explanation, make sure that your contract or service agreement includes a provision for them to provide such explanations about unexpected results in the future.

Existing ML and AI research is not well positioned to provide guidance on evaluating a tool's decision-making process. Explanations of decisions made by ML tools are the subject of a research area known as explainable AI, which focuses on the details of the ML apparatus. Current explainable AI research does not help with our question at hand, where the important aspect of whether an answer is good comes down to whether it helps the layperson understand why the expert's choice is reliable. For example, if you ask a car mechanic why a certain repair is appropriate to make, you will most likely expect an explanation you can understand about how cars work and why the replacement parts and their installation are adequate. If, instead of getting an explanation about the car, you get an explanation about the mechanic from a doctor who images the mechanic's brain during the repair and tells you that the readings indicate healthy memory, this explanation is not strictly false. It is also, in some sense, encouraging. But it is not the explanation you requested regarding whether the repair was actually appropriate. It is not an explanation at the correct level.[6] The current research into explainable AI is most likely going to provide an explanation that is more like the doctor's than the mechanic's. Such technical explanations are not sufficient to answer this section's question. The literature on "mechanistic explanation" provides additional details on aspects of good explanations [Glennan 2017] that could be adapted to cybersecurity [Spring 2018].

In addition, a single explanation is likely not adequate for all interested parties. Usually, experts address an explanation to a specific stakeholder community, and explanations that are adequate for one community are not always adequate for another [Preece 2018]. Practically, this means you should ensure that you, as a manager, get an adequate explanation to meet your needs, and that relevant stakeholders in your organization—such as developers, users, or lawyers—get an adequate explanation that meets their needs as well.

After you have addressed whether your ML tool's outputs are adequate in general, you are ready to focus on how you should reflect on the outputs of an ML tool meant to function specifically as a cybersecurity tool, which is the subject of the next section. It is not enough to be sure your suppliers have adequately designed and explained the results. Your adversaries will want to subvert or bypass your ML tool, and even well-designed tools are vulnerable. The next section is about mitigating the risks of corrupted results, models, or data.

---

[6] By "level," we mean the mechanistic level of explanation as described in Explaining the Brain: Mechanisms and the Mosaic Unity of Neuroscience [Craver 2007].

4. How will you protect the ML system against attacks in an adversarial, cybersecurity environment?

This fourth question of the series is also for managers and decision makers to ask the suppliers or developers of an ML tool. This question aims to prepare you for the fact that, in the same way that no modern computer system can be absolutely assured, there is no fail-safe way to protect an AI or ML system. Research into defending ML tools is ongoing. Responses to this question should acknowledge this fact and take a systems-security approach for reducing the risk and the impact of attacks to acceptable levels. Three important examples of such protection include the following: (1) robustness in the integrity of the tool as well as the protection of confidentiality; (2) resilience against attacks during training and classification; and (3) evidence that the input data is reliable and representative.

ML tools contain well-known vulnerabilities, many of which are susceptible to adversary manipulation. As a poignant example of attacks against ML tools, consider the case of self-driving cars. Such cars use ML tools to identify street signs, among other things. By intentionally manipulating a small section of a stop sign with a purpose-designed sticker, an adversary can make these operational ML tools reliably misclassify a stop sign as a 45-mile-per-hour speed limit sign [Eykholt 2018]. Adversaries constantly expose cybersecurity tools to input that is likely less obvious to an analyst than a sticker on a traffic-control sign. Therefore, any cybersecurity system you implement using ML tools should take these threats seriously.

To prepare for such possible manipulation, you should ask the developers and suppliers you are working with how the design and deployment of the ML tool protect against well-known attacks. This is not an abstract question about all possible attacks, but, rather, a way to address relevant attacks of which there are well-documented classes [Papernot 2018]. An adequate answer to this question should explain what protections will be in place against known attacks on the integrity of the tool's decisions and the confidentiality of sensitive information during both training and deployment. In addition, you should get evidence that the input data is reliable and representative.

Cybersecurity cannot prevent all attacks, nor does it purport to, but any tool deployed in a cybersecurity context should be part of an explicit risk assessment.[7] You can make your risk assessment easier if the tool is explainable—which we defined above in the discussion for question 3—but every tool should have a risk assessment. For ML tools, you should assess the tool's vulnerability to known classes of attacks and assess potential damage to both the system and consumers of the tool's output. A good answer to the question for this section should include an explanation of how a robust design of the ML model and of the tool itself protects it from attacks. It should also include an explanation of how you can continuously detect and mitigate any attacks that bypass these protections or manipulate the deployed ML tool. You should ask about plans for ongoing monitoring and evaluation.

These first four sections covered questions about the tool's purpose, data, model, and results to help ensure you understand what you want your ML tool to accomplish and what resources you have available to support it. The following two sections introduce questions that help ensure that

[7] For an example, see *The Security Risk Assessment Handbook* [Landoll 2005]. See also *Introducing OCTAVE Allegro: Improving the Information Security Risk Assessment Process* [Caralli 2007].

you use what you have to get what you want. The following section will help you discover and mitigate problems in your data and models.

5. How will you find and mitigate unintended outputs and effects?

Managers and decision makers should ask the suppliers or developers of an ML tool this question. You should expect a good response to identify stakeholders, provide guidance on how to consult them, and list sources of sensitive or misleading information. Additionally, a good response will provide information about how the tool performs its task without harming stakeholders. This topic is related to question 2, which discussed whether information was ethically available.

ML models, as discussed above, find patterns that may not be apparent to a human analyst. In many cases, the model may capture sensitive information or reproduce unwanted biases even if they are explicitly excluded from the data [Rocher 2019]. Studies have vividly demonstrated the occurrence of this problem in AI that uses ML tools to advise on criminal sentencing in the United States. For example, African Americans are four times more likely to be arrested for drug charges than white Americans despite similar rates of drug usage between the two groups. ML tools that have been trained to provide information on recidivism have relearned this bias in the criminal justice system even when race is specifically excluded as a data input [Angwin 2016]. Such systems frequently relearn underlying human biases that exist in other fields of knowledge and culture as well [Caliskan 2017]. The ML tool's strength in highlighting intricate relationships in data becomes a glaring weakness if it uses these relationships to reintroduce intentionally excluded features.

You should ask tool developers and suppliers about how the tool avoids or mitigates such unintended consequences. A good answer should provide assurance on—at minimum—the following five aspects of data handling and model development:[8]

- **Representation**—all relevant subjects are proportionally represented in the data. One aspect of proportionality that is especially relevant for cybersecurity is that the ratio of benign to malicious elements in the training data is realistic. For example, if a company receives 1000 emails from outside sources a week and only 50 of them contain malicious elements, the training data should reproduce that ratio. You should also represent other relevant features proportionally. Failing to provide accurate representation might introduce cultural bias into the initial classification. If all benign emails are in English, for example, but some malicious emails are not, the data will erroneously introduce a cultural bias that non-English languages are malicious. Commercial ML tools in other subject areas are known to commonly suffer from this problem. For example, poor representation in training data leads to systematically higher errors for facial recognition of darker-skinned women [Zou 2018].
- **Protection**—the ML tool does not mislearn confounding factors as proxies for sensitive characteristics, such as in the example above, where the ML tool relearned to detect race because other features of the data set (e.g., location) were correlated with it [Angwin 2016].

[8] We adapted the first four items from "AI is convicting criminals and determining jail time, but is it fair?" [Plonski 2018]. For other ethical considerations on collecting data in cybersecurity studies, see "The Menlo Report: Ethical Principles Guiding Information and Communication Technology Research" [Dittrich 2012].

Just marking race as a sensitive characteristic is not enough to resolve this issue. You should ensure the characteristic is not predictable or learnable from the remaining data.

- **Stewardship**—the relevant communities that produce the data that the ML tool analyzes, and which the ML tool's use impacts, are engaged. For example, if your organization will use an ML system to screen email for malicious content, the users of the email system are a relevant community that you should consult about how screening impacts their work and how it safely manages their data. Furthermore, the data adequately represents the desired end state. An example of a violation of this aspect of stewardship would be if someone trained a tool to engage in conversation using early 20th century literature where it was common to use derogatory terms for various groups of people.
- **Authenticity**—the features of the training data are faithful to the application environment. For example, you should ensure that past observations about a system are informative about its future, and that there is good justification to generalize from training examples to the deployed application.
- **Resiliency**—either (1) adversary access to the training or benchmarking data does not allow the adversary to easily interfere with the AI or ML tool, or (2) the tool does not rely on static input data. Resiliency is closely related to question 4.

If you ensure that your ML tool treats relevant people fairly, you often improve the reliability and usefulness of the result because the data and the model better align with the tool's purpose. A comprehensive evaluation of whether an ML tool's results are adequate for its purpose, however, is a broad topic. The following section introduces evaluation and highlights three common mistakes that can occur when people evaluate cybersecurity tools.

6. Can you evaluate the ML tool adequately, accounting for errors?

This is the final question in the series for managers and decision makers to ask the suppliers or developers of an ML tool. You should expect a good response to give details about the evaluations you can perform, and these should be closely linked to the techniques that scientific and engineering researchers use in their evaluations, such as repeatability, validity, and generalizability.

This section highlights three common flaws you should be sure the evaluations avoid. The first highlights the importance of performance measures. The second two are example mistakes that would subvert the generalizability of an evaluation. *Generalizability* ensures that your evaluation applies to other situations, such as your actual deployment of the tool. In summary, you can avoid some important common problems by ensuring that your evaluation (1) manages and assesses relatively rare events adequately; (2) understands the relationship between the evaluated population and the population you want to learn about; and (3) properly accounts for missing evidence. This section is broken into three subsections to address each of these items as well as a concluding section that addresses the importance of performance metrics. While far from exhaustive, the considerations that follow will form a baseline for an adequate response to this question. For a thorough account of the framework for an adequate evaluation, see "Practicing a Science of Security" [Spring 2017].

Item 1: Managing Rare Events

A common problem for operators who manage alarms, such as nuclear power plant operators or computer security incident responders, is that people tend to ignore alarms if they are almost always triggered incorrectly and do not require a response most of the time. However, when people tend to ignore alarms because they are often false, errors occur when the system triggers a real alarm. This issue is more likely to happen if the event of interest is relatively rare compared to the normal state. Many relevant cybersecurity events are relatively rare. For example, most software instances are not malicious. The effect of a low base rate of occurrence is that seemingly small problems in detection result in overwhelming numbers of false alarms, and analysts may end up ignoring all of the alarms that the system produces.

Rare Events and False Alarms

To demonstrate the effect that rare events have on erroneous alarms, suppose we are interested in a malicious item with a base-rate of one in 10,000, so that in one million samples, only 100 will be malicious, and 999,900 will be benign. Now consider a tool that is correct 99% of the time, so that 1% of malicious events are identified as benign and 1% of benign events are identified as malicious. This tool will misclassify 9,999 of the 999,900 benign events as malicious, and it will correctly classify 99 of the 100 malicious events as malicious. Therefore, despite a seemingly low error rate in the test, about 99% (calculated as $\frac{9999}{99+9999} = 99.0\%$) of the tool's results are alarm errors [Axelsson 2000]. The fact that there are so many more benign cases in the population drastically impacts the results.

Table 2: An Example of Notional Alarm Errors with 1% Test Error

	Total events	*Reported malicious (alarm)*	*Reported benign*
Actual malicious	100	99 (correct alarms)	1
Actual benign	999,900	9999 (alarm errors)	989,901

Item 2: Understanding the Relationship Between the Evaluated Population and the Target Population

The second common error you should avoid is called *survivorship bias*, which is a misunderstanding of how the population you observe represents the population you want to study. This type of error got its name because it was first studied when evaluating new armor for military aircraft returning from combat [Mangel 1984]. The goal of the armor evaluation was to estimate which vulnerable parts of a plane most often led to the plane being shot down. The difficulty was that only planes that were not shot down were the ones available for observation to determine the need for new armor. The engineers observing the planes originally considered the possibility that returning planes needed more armor where there was evidence that they had been shot. The right answer, however—which the project team correctly ascertained—was to put more armor where they were *not* shot. The planes that did not return likely took fatal damage in the places where the successful planes remained unscathed.

The general problem of survivorship bias also relates to cybersecurity. Usually, we want to estimate something unobservable, such as how many intrusion events an organization did not detect. There are statistical methods for such evaluations, but using them is not always straightforward. The main pitfall to avoid is treating an observed or evaluated population as representative when it is systematically not representative. Analogously to the airplane example above, taking all the intrusions an organization knows about as input data is a strategy that is unlikely to be useful for evaluating a tool's ability to detect intrusions the organization does not know about.

Item 3: Accounting for Missing Evidence

Misunderstanding absent evidence is the final common error to guard against during evaluation of AI systems and ML tools. Intelligence analysis has a long history of addressing this challenge, and cybersecurity should draw on that history. Heuer offered the following advice on overcoming this problem: "[I]dentify explicitly those relevant variables on which information is lacking, consider alternative hypotheses concerning the status of these variables, and then modify… judgement accordingly. [Also] consider whether the absence of information is normal or is itself an indicator of unusual activity or inactivity" [Heuer 1999, p. 19]. This advice is worth quoting here because we should apply it to thinking about what answers the questions we ask about evaluation quality should include.

The Importance of Performance Metrics

The examples from each of the items above highlight the importance of selecting performance metrics. Performance metrics are the criteria that the tool's users will employ to evaluate its success or failure [Russell 2010, chapter 2]. In the best case, if you properly use performance metrics, the ML tool will perform well in the metrics you measure. On the other hand, the ML tool might perform poorly if you do not measure certain aspects of its performance. For example, if your performance metrics do not include time calculations, then it is unlikely the ML tool will compute answers quickly. Or, if your performance measures do not consider usual rates of occurrence in your environment, the tool will likely overwhelm analysts with alerts. In broad strokes, make sure your evaluations measure what you actually care about improving.

This section is not a complete checklist of the evaluation of a cybersecurity tool. We provided three common errors that you should ensure any answer avoids. For more comprehensive evaluation requirements, see *Security Engineering: A Guide to Building Dependable Distributed Systems, 2nd Edition* [Anderson 2008] and *Systems Security Engineering: Considerations for a Multidisciplinary Approach in the Engineering of Trustworthy Secure Systems* [NIST 2016]. The following, and final, section asks you bring together what you have learned, compare options, and make your decision.

7. What alternative tools have you considered? What are the advantages and disadvantages of each one?

Managers and decision makers should ask themselves these final questions. As you do so, you should take stock of the answers to the previous six questions. However, to properly compare multiple tools, you should ask all seven of the questions in this report about multiple solutions. At this point, you should go beyond comparing answers to the previous questions and also address

business considerations. Estimating costs of development, deployment, and maintenance are important problems of their own,[9] but we will not focus on those here.

You should be sure to consider advantages and disadvantages at various stages in the tool's lifecycle. ML tools can quickly accrue unsustainable levels of maintenance costs and technical debt. In this case, cybersecurity tools are no better than any other software tool. Some risk factors that you should avoid where possible include complex ML models that erode software boundaries, hard-to-trace data dependencies, and undue tolerance for software design anti-patterns that result in spaghetti code [Sculley 2014].

In the case of cybersecurity, it is particularly relevant to understand the threat lifecycle and how you can update a tool when an adversary learns how to subvert it. For example, it is undesirable if one week of adversary effort takes a three-month tool redevelopment to counter. More generally, you should try to predict the adversary's response to the tool and whether that response puts them in a better or worse position to carry out an attack [Spring 2015].

[9] For examples about estimating costs of development, deployment, and maintenance, see the publications on software cost estimates at the SEI blog at https://insights.sei.cmu.edu/sei_blog/software-cost-estimates/.

Conclusion

Machine learning (ML) and artificial intelligence (AI), like any tools, should be designed to fit their intended purpose. ML tools have great promise, but you should take due care before applying them to cybersecurity problems.

The cybersecurity context makes answering some questions difficult. For example, getting relevant information can be difficult if what you need to gather is absolutely correct knowledge about malicious intent. Getting such information amounts to reading some unknown person's mind. A well-designed tool will sidestep such difficulties to bring value where it can.

Tools are part of systems. A hammer cannot drive a nail by itself, though it certainly makes your arm more effective at driving nails than it would be without the hammer. Well-designed ML tools should likewise integrate with the rest of your cybersecurity system. ML tools can make that system better, and if well-integrated, the system can also protect and support an ML tool.

If you ask the questions that we outline in this document, you should be better able to acquire a well-designed ML tool that fits your needs and to integrate it into your organization. And you should not forget that maintaining the tool means asking these questions again occasionally to make sure the context, environment, and purpose of the ML tool remain aligned.

References

URLs are valid as of the publication date of this document.

[Anderson 2008]
Anderson, Ross J. *Security Engineering: A Guide to Building Dependable Distributed Systems, 2nd Edition*. Wiley. 2008. ISBN 978-0-470-06852-6.

[Angwin 2016]
Angwin, Julia; Larson, Jeff; Mattu, Surya; & Kirchner, Lauren. Machine Bias: There's software used across the country to predict future criminals. And it's biased against blacks. *ProPublica*. May 23, 2016. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[Axelsson 2000]
Axelsso, Stefan. The Base-Rate Fallacy and its Implications for the Difficulty of Intrusion Detection. *ACM Transactions on Information and System Security (TISSEC)*. Volume 3. Issue 3. August 2000. Pages 186-205. https://dl.acm.org/citation.cfm?doid=357830.357849

[Caliskan 2017]
Caliskan, Aylin; Bryson, Joanna J.; & Narayanan, Arvind. Semantics derived automatically from language corpora contain human-like biases. *Science*. Volume 356. Issue 6334. April 14, 2017. Pages 183-6. http://science.sciencemag.org/content/356/6334/183

[Caralli 2007]
Caralli, Richard A.; Stevens, James F.; Young, Lisa R.; & Wilson, William R. *Introducing OCTAVE Allegro: Improving the Information Security Risk Assessment Process*. CMU/SEI-2007-TR-012. Software Engineering Institute, Carnegie Mellon University. May 2007. https://resources.sei.cmu.edu/library/asset-view.cfm?assetid=8419

[Craver 2007]
Craver, Carl F. *Explaining the Brain: Mechanisms and the Mosaic Unity of Neuroscience*. Oxford University Press. August 2007. ISBN 9780199299317.

[Dittrich 2012]
Dittrich, D. & Kenneally, E. The Menlo Report: Ethical Principles Guiding Information and Communication Technology Research. *US Department of Homeland Security*. August 2012. https://www.caida.org/publications/papers/2012/menlo_report_actual_formatted/menlo_report_actual_formatted.pdf

[Douglas 2018]
Douglas, Heather. How Can the Public Assess Expertise? *JBS Haldane Memorial Lecture*. January 31, 2018. https://www.youtube.com/watch?v=cuB06iZ8-sM&feature=youtu.be

[European Commission 2019]
European Commission. 2018 reform of EU data protection rules. *European Commission Website*. March 27, 2019 [accessed]. https://ec.europa.eu/commission/priorities/justice-and-fundamental-rights/data-protection/2018-reform-eu-data-protection-rules_en

[Eykholt 2018]
Eykholt, Kevin; Evtimov, Ivan; Fernandes, Earlence; Li, Bo; Rahmati, Amir; Xiao, Chaowei; Prakash, Atul; Kohno, Tadayoshi; & Song, Dawn. Robust Physical-World Attacks on Deep Learning Models. *arXiv*. April 10, 2018. https://arxiv.org/pdf/1707.08945.pdf

[Glennan 2017]
Glennan, Stuart & Illari, Phyllis, eds. *The Routledge Handbook of Mechanisms and Mechanical Philosophy*. Routledge. 2017. ISBN 9781138841697.

[Heuer 1999]
Heuer, Richards J. *Psychology of Intelligence Analysis*. Center for the Study of Intelligence, Central Intelligence Agency. 1999. https://www.cia.gov/library/center-for-the-study-of-intelligence/csi-publications/books-and-monographs/psychology-of-intelligence-analysis/PsychofIntelNew.pdf

[Horneman 2017]
Horneman, Angela. How to Think Like an Analyst [blog post]. *SEI Blog*. July 17, 2017. https://insights.sei.cmu.edu/sei_blog/2017/07/how-to-think-like-an-analyst.html

[Klein 2010]
Klein, Gerwin; Andronick, June; Elphinstone, Kevin; Heiser, Gernot; Cock, David; Derrin, Philip; Elkaduwe, Dhammika; Engelhardt, Kai; Kolanski, Rafal; Norrish, Michael; Sewell, Thomas; Tuch, Harvey; & Winwood, Simon. seL4: Formal Verification of an Operating-System Kernel. *Communications of the ACM*. Volume 53. Number 6. June 2010. Pages 107-115. https://cacm.acm.org/magazines/2010/6/92498-sel4-formal-verification-of-an-operating-system-kernel/fulltext

[Landoll 2005]
Landoll, Douglas J. *The Security Risk Assessment Handbook: A Complete Guide for Performing Security Risk Assessments*. CRC Press. December 12, 2005. ISBN 9781420031232.

[Mangel 1984]
Mangel, Marc & Samaniego, Francisco J. Abraham Wald's Work on Aircraft Survivability. Journal of the American Statistical Association. Volume 79. Issue 386. 1984. Pages 259-267. https://www.tandfonline.com/doi/pdf/10.1080/01621459.1984.10478038?needAccess=true

[Moore 2019]
Moore, Andrew W. AWM Tutorial Page. *Carnegie Mellon University, School of Computer Science Website*. March 27, 2019 [accessed]. https://www.cs.cmu.edu/~awm/tutorials.html

[Munroe 2017]
Munroe, Randall. Machine Learning. *xkcd Website*. May 17, 2017. https://xkcd.com/1838/

[NIST 2016]
Ross, Ron; McEvilley, Michael; & Oren, Janet Carrier. *Systems Security Engineering: Considerations for a Multidisciplinary Approach in the Engineering of Trustworthy Secure Systems*. SP 800-160v1. National Institute of Standards and Technology, United States Department of Commerce. November 2016.

[Ng 2019]
Ng, Andrew. AI for Everyone. *Coursera*. July 31, 2019 [accessed]. https://www.coursera.org/learn/ai-for-everyone

[Papernot 2018]
Papernot, Nicolas; McDaniel, Patrick; Sinha, Arunesh; & Wellman, Michael P. SoK: Security and Privacy in Machine Learning. Pages 399-414. In *2018 IEEE European Symposium on Security and Privacy (EuroS&P 2018)*. London, United Kingdom. April 24-26, 2018.

[Pearl 2018]
Pearl, Judea. The Seven Tools of Causal Inference, with Reflections on Machine Learning. *Communications of the ACM*. Volume 62. Number 3. March 2019. Pages 54-60. https://cacm.acm.org/magazines/2019/3/234929-the-seven-tools-of-causal-inference-with-reflections-on-machine-learning/fulltext

[Polonski 2018]
Polonski, Vyacheslav. AI is convicting criminals and determining jail time, but is it fair? *World Economic Forum*. November 19, 2018. https://www.weforum.org/agenda/2018/11/algorithms-court-criminals-jail-time-fair/

[Preece 2018]
Preece, Alun; Harborne, Dan; Braines, Dave; Tomsett, Richard; & Chakraborty, Supriyo. Stakeholders in Explainable AI. *arXiv*. September 29, 2018. https://arxiv.org/pdf/1810.00184.pdf

[Rocher 2019]
Rocher, Luc; Hendrickx, Julien M.; & de Montjoye, Yves-Alexandre. Estimating the Success of Re-identifications in Incomplete Datasets using Generative Models. *Nature Communications*. Volume 10. Number 1. July 2019. Page 3069.

[Russell 2010]
Russell, Stuart J. & Norvig, Peter. *Artificial Intelligence: A Modern Approach*. 3rd edition. Prentice Hall. 2010. ISBN 9780136042594

[Sculley 2014]
Sculley, David; Holt, Gary; Golovin, Daniel; Davydov, Eugene; Phillips, Todd; Ebner, Dietmar; Chaudhary, Vinay; & Young, Michael. Machine learning: The high interest credit card of technical debt. *Google*. 2014. https://storage.googleapis.com/pub-tools-public-publication-data/pdf/43146.pdf

[Spring 2015]
Spring, Jonathan M. & Stoner, Edward. *CND Equities Strategy*. CERTCC-2015-40. Software Engineering Institute, Carnegie Mellon University. July 2015. https://resources.sei.cmu.edu/library/asset-view.cfm?assetID=442305

[Spring 2017]
Spring, Jonathan M.; Moore, Tyler; & Pym, David. Practicing a Science of Security. In *Proceedings of the 2017 New Security Paradigms Workshop*. Santa Cruz, CA. October 1-4, 2017. https://dl.acm.org/citation.cfm?id=3171533&picked=prox

[Spring 2018]
Spring, Jonathan M. & Illari, Phyllis. Building General Knowledge of Mechanisms in Information Security. *Philosophy and Technology*. September 17, 2018. https://link.springer.com/article/10.1007/s13347-018-0329-z

[Shirey 2007]
Shirey, R. Internet Security Glossary, Version 2. *IETF*. August 2007. https://datatracker.ietf.org/doc/rfc4949/

[US-CERT 2019]
US-CERT. Introduction to Information Security. March 27, 2019 [accessed]. https://www.uscert.gov/sites/default/files/publications/infosecuritybasics.pdf

[Wikipedia 2019]
Machine Learning. *Wikipedia*. March 27, 2019 [accessed]. https://en.wikipedia.org/wiki/Machine_learning

[Yeom 2018]
Yeom, Samuel; Giacomelli, Irene; Fredrikson, Matt; & Jha, Somesh. Privacy Risk in Machine Learning: Analyzing the Connection to Overfitting. Pages 268-282. In *2018 IEEE 31st Computer Security Foundations Symposium (CSF)*. Oxford, United Kingdom. July 9-12, 2018.

[Zou 2018]
Zou, James & Schiebinger, Londa. AI can be sexist and racist---it's time to make it fair. *Nature*. July 18, 2018. 559(7714):324. https://www.nature.com/articles/d41586-018-05707-8

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
(Leave Blank)	Month and Year (date added at time of publication)	Final

4. TITLE AND SUBTITLE	5. FUNDING NUMBERS
Machine Learning in Cybersecurity: A Guide	FA8702-15-D-0002

6. AUTHOR(S)

Jonathan M. Spring, Joshua Fallon, April Galyardt, Angela Horneman, Leigh Metcalf, Edward Stoner

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)	8. PERFORMING ORGANIZATION REPORT NUMBER
Software Engineering Institute Carnegie Mellon University Pittsburgh, PA 15213	CMU/SEI-2019-TR-005

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSORING/MONITORING AGENCY REPORT NUMBER
AFLCMC/PZE/Hanscom Enterprise Acquisition Division 20 Schilling Circle Building 1305 Hanscom AFB, MA 01731-2116	n/a

11. SUPPLEMENTARY NOTES

12A DISTRIBUTION/AVAILABILITY STATEMENT	12B DISTRIBUTION CODE
Unclassified/Unlimited, DTIC, NTIS	

13. ABSTRACT (MAXIMUM 200 WORDS)

This report lists relevant questions that decision makers should ask of machine-learning practitioners before employing machine learning (ML) or artificial intelligence (AI) solutions in the area of cybersecurity. Like any tool, ML tools should be a good fit for the purpose they are intended to achieve. The questions in this report will improve decision makers' ability to select an appropriate ML tool and make it a good fit to address their cybersecurity topic of interest. In addition, the report outlines the type of information that good answers to the questions should contain. This report covers the following questions:

1. What is your topic of interest?
2. What information will help you address the topic of interest?
3. How do you anticipate that an ML tool will address the topic of interest?
4. How will you protect the ML system against attacks in an adversarial, cybersecurity environment?
5. How will you find and mitigate unintended outputs and effects?
6. Can you evaluate the ML tool adequately, accounting for errors?
7. What alternative tools have you considered? What are the advantages and disadvantages of each one?

14. SUBJECT TERMS	15. NUMBER OF PAGES
	29

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	UL

NSN 7540-01-280-5500

Standard Form 298 (Rev. 2-89) Prescribed by ANSI Std. Z39-18
298-102