All.Net Analyst Report and Newsletter

Welcome to our Analyst Report and Newsletter

The dimensions of the problem space

I seem to run into a lot of papers about security issues, and most of the better ones start with an introduction to the space. They sound pretty good and might even have that authoritative look and feel.

But what most lack, in my view, is a truly systematic approach of addressing the dimensions of the problem space. When they do that they seem to miss major parts of the issue that come back to haunt them.

As an example

> ### Executive Summary
>
> This NIST Trustworthy and Responsible AI report is intended to be a step toward developing a taxonomy and terminology of adversarial machine learning (AML), which in turn may aid in securing applications of artificial intelligence (AI) against adversarial manipulations of AI systems. Broadly, there are two classes of AI systems: Predictive and Generative. The components of an AI system include – at a minimum – the data, model, and processes for training, testing, and deploying the machine learning (ML) models and the infrastructure required for using them. Generative AI systems may also be linked to corporate documents and databases when they are adapted to specifc domains and use cases. The data-driven approach of ML introduces additional security and privacy challenges in different phases of ML operations besides the classical security and privacy threats faced by most operational systems. These security and privacy challenges include the potential for adversarial manipulation of training data, adversarial exploitation of model vulnerabilities to adversely affect the performance of the AI system, and even malicious manipulations, modifications or mere interaction with models to exfltrate sensitive information about people represented in the data, about the model itself, or proprietary enterprise data. Such attacks have been demonstrated under real-world conditions, and their sophistication and potential impact have been increasing steadily. AML is concerned with studying the capabilities of attackers and their goals, as well as the design of attack methods that exploit the vulnerabilities of ML during the development, training, and deployment phase of the ML lifecycle. AML is also concerned with the design of ML algorithms that can withstand these security and privacy challenges. When attacks are launched with malevolent intent, the robustness of ML refers to mitigations intended to manage the consequences of such attacks.[1]

Now this sounds really lucid and clear in its presentation, well thought out, etc. So what's my problem? I will break it down a bit.

- *Broadly, there are two classes of AI systems: Predictive and Generative.*
  - Sorry – besides prediction and generation, there is also, at least, explanation. Which is to say there are likely others

1 https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-2e2023.pdf

- *The components of an AI system include – at a minimum – the data, model, and processes for training, testing, and deploying the machine learning (ML) models and the infrastructure required for using them.*
  - This is for machine learning, and that is not the only sort of AI.
- *Generative AI systems may also be linked to corporate documents and databases when they are adapted to specific domains and use cases.*
  - They may also be linked to and do many other things even if not adapted to specific domains and use cases.
- *The data-driven approach of ML introduces additional security and privacy challenges in different phases of ML operations besides the classical security and privacy threats faced by most operational systems.*
  - The notion that some threats are “classical” and others are not is problematic in several ways, including without limit:
    - Threats are generally thought of in different ways, including without limits, threat actors, activities that may cause harm, etc. And the term threat is poorly defined in this context.
    - “security” is undefined in this context and a very broad and widely differently used term.
    - “privacy” is presumably identified as other than “security”, limiting one small chunk out of “security” for no apparent reason.
    - Both of these terms are often associated with unauthorized disclosure, which is only one of many serious issues with such systems. Let us add, as a staring point (and continue from there):
      - Integrity, availability, confidentiality, use control, accountability, transparency, custody (http://all.net/Arch/index.html):
      - And from archival science: reliability and authenticity and various elements of those and other concerns. (https://interparestrustai.org/terminology)
      - And from the psychological and sociological literature, various cognitive issues with systems, individuals, groups, organizations, and societies.
- *These security and privacy challenges include the potential for adversarial manipulation of training data, adversarial exploitation of model vulnerabilities to adversely affect the performance of the AI system, and even malicious manipulations, modifications or mere interaction with models to exfltrate sensitive information about people represented in the data, about the model itself, or proprietary enterprise data.*
  - The lack of a comprehensive approach to the dimensions of the problem and the previous items identified has limited their perspectives here pretty severely.
- *Such attacks have been demonstrated under real-world conditions, and their sophistication and potential impact have been increasing steadily.*
  - Many other such attacks have been underway across the wider spectrum as well.

- *AML is concerned with studying the capabilities of attackers and their goals, as well as the design of attack methods that exploit the vulnerabilities of ML during the development, training, and deployment phase of the ML lifecycle.*
  - Of course they can define AML as whatever they like, but by so constraining it they limit their perspectives and essentially guarantee that they will not cover the space effectively. A term we use for this is selective blindness. By so limiting, they blind themselves and their readers to the broader implications and thus, in some sense, guarantee at best limited success.
- *AML is also concerned with the design of ML algorithms that can withstand these security and privacy challenges.*
  - Apparently they are only interested in the algorithms and not the other aspects of even the limited components of the mechanisms they themselves have previously identified.
- When attacks are launched with malevolent intent, the robustness of ML refers to mitigations intended to manage the consequences of such attacks.
  - The term “attacks” and “malevolent intent” are of course problematic as well. As a starting point, one person’s malevolence is another person’s benevolence. It’s a matter of perspective. As definition, the additional problem is that since “malice” and “attack” are not differentiable terms, it seems to all come down to the mind of the observer, which defied scientific principals of observer independence.

I do not intend this to be harsh

This is not intended to be a critique of the people and their intent to do good things with their write-up. And it is not intended to critique their particular approach to the problem they seek to solve. It was merely an example in front of me. The underlying problem is a lack of clear definitions and defined dimensions of the problem space. With such definitions and dimensions, we can take comprehensive approaches, or if we wish to only address a subspace, be clear on what that subspace is.

And to be clear, this is a very common problem in the cyber security space, today as it has been for a long time. We seem to automatically choose common terms and adopt them as if everybody knew we were talking about something else. But the side effect is we grind these terms into our thought patterns and start to not realize what we are missing.

Which is why it is so important to start by recognizing the dimensions of the space we are working in and through and make them explicit, well defined, and carefully tended.

Conclusions

The dimensions of the problem space are key to solving the problems in the space. As this example shows, even well meaning people who fail to address the dimensions of the problems they seek to work to solve will miss the forest for the trees. We don’t need to try to boil the ocean to make progress, but we need to recognize the ocean and what part of it we are seeking to bring light to, or we are simply spitting into the wind and digging ourselves deeper. And if you didn’t enjoy the mixed metaphors here, it’s never too late to do the right thing by taking your time and seeking the truth.

