PL
TAMING THE AI WILD WEST WITH IBM
Better, Faster, Stronger NETWORK AUTOMATION FEATURING LightRiver
POWERING THE AUTONOMOUS 5G CORE
ENTERPRISE BLOCKCHAIN
Artificial Intelligence FOR NETWORK CONVERGENCE
Industry 4.0 PROGRAMMABLE MEDIA ANALYTICS
ARTIFICIAL INTELLIGENCE IN CYBERSECURITY
ANALYTICS FOR IOT
ETHICAL GENERATIVE AI
SERVICE-DRIVEN NETWORKS
AI, ANALYTICS & AUTOMATION

www.pipelinepub.com
Volume 19, Issue 11

The Ethical Use of Generative AI

By: William Yeack, CSE, Mark Cummings, Ph.D., Zoya Slavina

Generative AI promises profound benefits for society. Unfortunately it comes with some nasty side effects, including serious cybersecurity threats. Taking an approach of identifying the bad side effects and then developing ways to mitigate them, may make the ethical use of Generative AI possible across a wide application area. In the cybersecurity space, there is potential to use Generative AI systems in an ethical fashion to test defensive systems and train staff. Given that, professional ethics in this field require us to do so.

Experts in Generative AI warn against using it to test mitigation tools out of a concern that doing so will make the Generative AI system better and more likely to create these nasty side effects. In the cybersecurity space, this creates a dilemma. How can we test and train effectively without using Generative AI? Below we will discuss the dilemma and its resolution in the context of cybersecurity, as well as how the approach taken for cybersecurity can be generalized to the other application areas.

Bad actors using Generative AI are fundamentally changing the cybersecurity attack space. The bad actors will do everything possible to improve the Generative AI systems for creating attacks. They will do this by subverting attempted controls on public systems and creating private systems stripped of all controls.

In rogue states where cybercrime is a significant portion of the GDP, they will build fit-for-purpose Generative AI attack systems. In fact, just recently a Generative AI SaaS (Software as a Service) specifically structured to create cybersecurity attacks appeared on the Dark Web. This is forcing the creation of a new generation of defense tools—ones that are dynamic and adaptive.

The question is: how do we test and improve our defenses without increasing the strength and ease of Generative AI systems to create attacks? There is a large body of published material documenting the capability of Generative AI systems to do bad things. In each one, the authors recommend that using Generative AI systems to test around these bad things shouldn't be done. It shouldn't be done, because in so doing, the Generative AI systems will be trained to do bad things *better* and to develop easier ways to circumvent ethical controls.

In cybersecurity recommendations relative to Generative AI, authors start with not using Generative AI for “Red Teaming.” A Red Team is a group of people and tools that are used by cybersecurity professionals to test defensive systems. The concept is that the Red Team can find vulnerabilities that a “Blue Team” of defenders can then discover. Based on this discovery, the Blue Team then can improve defensive tools, process, procedures, and training. The argument is that in performing Red Team functions, a Generative AI system will learn how to create more new types of attacks, which may be more damaging, faster, etc. The logic here seems unassailable. Carrying that logic further, employing a Generative AI system as part, or all, of a Blue Team would lead to the same result: a better attacking Generative AI system. Without being able to use Generative AI systems in testing defenses and the training of staff to respond to Generative AI-created attacks, however, defenders will be at a great disadvantage.

The challenge then is to create a way to use Generative AI while avoiding this ethical conundrum. What is presented below is the beginning of an approach to meeting this challenge.

An Ethical Approach to Using Generative AI

Here we will present the basic principles of an approach to using Generative AI while avoiding the ethical conundrum the cybersecurity defense space. Then, we will discuss some of the implementation issues.

The basic principles are:

1. Place the Generative AI system in an environment where it cannot communicate with the outside world
2. Destroy the Generative AI system after it is used

These principles seem simple until we consider their implementation. Let’s take each of the two principles one at a time. To achieve Principle #1, we cannot use any public Generative AI system available on the Internet. We must use a system that is designed to run on dedicated, isolated hardware—that is, physically and logically isolated. This is often referred to as being “air-gapped.” So, we have to either acquire or develop such a system. (More on that later.)

We also know that Generative AI systems have a record of manipulating people to do things for them that the AI can’t do for itself. In a lab testing environment, staff must be constantly reminded not to do anything on behalf of the Generative AI system. And especially not outside of the physically and logically controlled lab environment.

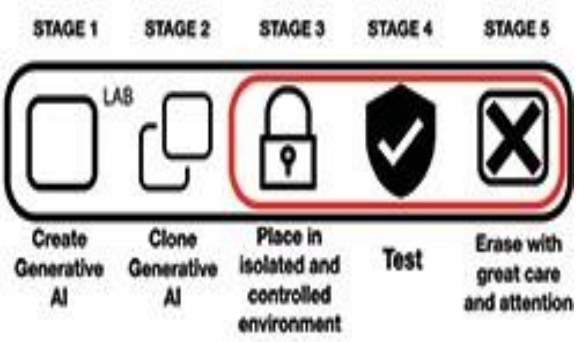

Illustration #1 Containing Generative AI Systems for Ethical Test and Training
size to enlarge

It will be harder to ensure that cybersecurity staff members in training environments, however, don't act on behalf of the isolated Generative AI system. This is because they are less likely to fully understand the danger. So, special controls and extra reminders/training, plus active monitoring, will be required.

Principle #2 is intended to remove the threat of the Generative AI system learning to become better at attacks. The key question here is how to do it economically. Generative AI systems are expensive. They take a long time and a lot of effort to train. It is difficult to convince people to throw all that investment away. Once a Generative AI system has been created, however, it can be cloned. Cloning takes effort. But nothing like the effort required to develop and train a system. Thus, for a particular lab test, a clone can be created and at the end of the test, be erased. There are a number of cautions to consider here though. Erasing has to be done with serious discipline—it's not just simply erasing a few pointers and leaving everything else, for example. Also, care has to be taken to make sure that the Generative AI system doesn't take any action to clone itself and then hide its clone. There has been published speculation about systems actively trying to prevent shutdown, or creating a clone and hiding it.

Implementation Tradeoffs in Expense, Time, and Effectiveness

Developing a Generative AI system such as ChatGPT4 or Claude2 took many years and many tens of millions of dollars. Of course, doing it a second time with the aid of the first timers' experience takes less. Still, to develop and train a model with billions of parameters is likely to take a substantial portion of a year, and the cost will be less than the first timers' but still tens of millions of dollars. Furthermore, to satisfy Principle #1, the Generative AI system has to run on dedicated hardware—not on a public cloud. And, not even on a private cloud. Some estimate that the hardware for the largest models can approach the billion-dollar level. So, the hardware and operating expense of such a model is also quite large.

For large government organizations and very large enterprises, this may not be a problem. For the innovative, small software companies, however, from which the needed innovative dynamic/adaptive (S2-D2) defensive tools are likely to originate, this is out of reach.

There *is* an alternative: Open Source. There are open source Generative AI models and open source Internet crawls for training. Furthermore, there are a range of such tools available. Some will run on a single desk-top computer. At the other extreme is the recently announced Meta LLMA2.

The open question is how large a Generative AI system will need to be. A smaller system will be more economical and easier to control. But, will it have enough capability to be a good stand-in for systems being used by bad actors and rogue states to launch cyber attacks?

A new approach is emerging that may help with this problem: MOE. A MOE (Mixture of Experts) is a smaller expert model trained on its own tasks and subject areas. With a MOE, it may be possible to have a small Generative AI system that is just as capable at producing attacks as a large, more general system. Of course, the attackers will likely use MOEs, too, so keeping track of the cost/capability trade-off over time will be critical.

This outline of the trade-offs shows that it is possible to meet both principles while using Generative AI technology to test defenses and train staff. Clearly, there is still work to do for interested organizations to create their own environments that meet these principles. And different organizations may take different paths in so doing.

Cybersecurity Defense Conclusion

The bottom line is that it is indeed possible to use Generative AI systems in an ethical fashion to test defensive systems and train staff. And, as stated above, professional ethics in this field demand that we do so.

Other Generative AI Application Areas

The above discussion only considers cybersecurity issues around defending *other* systems from attacks created by a Generative AI system. It does not consider protecting Generative AI systems themselves from attack. This is one of the other areas that needs to be considered.

Other application areas have problems with harmful side effects of Generative AI like misinformation, lying, libel, manipulation, and so on. A similar approach can be applied to those side effects. For each, the side effect(s) must be well defined. Then, principles for overcoming those bad side effects must be established. In some cases, the principles stated here may be relevant. In others, new principles will need to be developed. In any case, developing such principles may not be easy. For example, a high school student in an Oakridge, Tennessee, community meeting proposed: “AI should be used to assist a doctor but should never be used instead of a doctor.” This may seem reasonable on its face. But, there are already completely automated surgical systems currently in use. Once the principles are well defined, various implementation approaches can be considered. In some cases, the approaches discussed here can be a foundation; for others, new approaches will be needed and those approaches may require new technology. What seems clear is that there is no one-size-fits-all solution. Identifying the harmful side effects and working to mitigate them, however, may make the ethical use of Generative AI possible across a wide application area.