www.pipelinepub.com
Volume 22, Issue 1

AI Security and Assurance: Achieving the AI Intelligent Agent Promise

By: Mark Cummings, Ph.D.

GenAI, and intelligent agents using it, have the potential for significant productivity improvements. To achieve the full productivity benefits, we need to learn how to secure, develop, and deploy these technologies. Currently, it appears that we are very low on the learning curve. As an industry, we are just beginning to understand the security vulnerabilities and the challenges of developing effective applications of these technologies. The best way to quickly move up the learning curve and capture the full productivity benefits is to create a way that implementers can share experience and develop best practices. Combining that with similar work on minimizing the downsides of the technology will turbocharge both. The Bace Cybersecurity Institute (BCI) is exploring the creation of such a sharing organization. For discussion purposes, BCI is calling such an organization AIWG (AI Working Group) and creating a place where interested professionals with the necessary types of expertise can discuss challenges and possible responses.

AI Security Challenges

When people first became aware of GenAI, attention focused on how it could be used to increase the frequency and strength of security attacks.

Awareness has slowly developed about AI's own security vulnerabilities. It first focused on corrupting the models during training. The corruption could be intentional or unintentional. This is a real problem. But since training is generally done in a controlled environment, remediation focuses on having adequate controls. Key areas to control are the source and quality of training data and the control of the supply chain for the training data. These controls should be applied to both basic training and fine-tuning. Although the controls are fairly well understood, ways of implementing them are still in a learning process. For example, there is some early work being done on what becomes a standard corpus of training data.

Similar to concerns about training vulnerability, there are concerns about what goes into the Context Window of an LLM. Since the frontier model context windows have become so large, it is easy to miss a small piece of corrupted data in a Context Window. There are two ways that corrupted data can get into a Context Window: as part of the background data that is loaded into the Context Window, or through an inference request. Both are generally called prompt injection.

There is a growing recognition of how AI systems can be compromised by prompt injection attacks. Currently, prompt injection attacks can work in one of two ways. They put corrupting information directly in inference requests, or in "background" information such as photos that go into the context window. In both types of prompt injection attacks, corrupting information is introduced.

There are many other ways that attack information can be delivered to the AI. The attack information can be presented as a normal inference request. That is, language that acts similarly to how a scammer gets a well-intentioned human to assist in an attack. It can also be presented in ways that only the AI system can perceive. Ways that are invisible to humans. For example, through the use of invisible fonts/colors in text. Or characters embedded in photos that go into the context window. Corrupting information can also be introduced in different languages.

Currently, contact center automation with AI intelligent agents is receiving a lot of attention.

Prompt injection attacks can be used against intelligent agent contact centers. Over time, AI will be used in a broad range of online systems, exposing both B2B and B2C systems to vulnerabilities. They will also be used in control systems for factories, homes, and infrastructure. Thus, exposing a broad range of society to vulnerability.

As attackers get more experience with AI systems, additional security vulnerabilities are likely to emerge. Keeping developers and defenders up to date on these developments and how to update defenses will also be important.

AI Agent Development Challenges

AI chatbots have proven to be very helpful and popular. They will continue to be so. With that said, the impact of intelligent agents based on GenAI may have a greater impact on productivity measured broadly across society. It is similar to the change we went through with PC's. One where we went from computers that were easy to use to computers that did things for us.

As a result, there is a lot of work underway in creating AI agent applications. Such work is going on in large corporations, government, and academia. Unfortunately, it is not currently producing results that equal its promise. For example, a recent MIT report says that 95% of GenAI pilots at companies are failing. This indicates that, as an industry, we are very low on the learning curve. Not surprising given how new and different GenAI is. Moving up the learning curve more quickly is important to achieve GenAI's full promise and to achieve it in a cost-effective fashion.

The Need for Sharing

AI implementers, each working in their own organizations, are trying to implement intelligent agents. In the process, each is discovering pieces of new security vulnerabilities, ways of mitigating security vulnerabilities, and new ways of creating effective intelligent agents. At

the same time, vendors are making guesses about what implementers want and need. The isolation of these efforts is a contributing factor to the low level of application success.

To achieve the full promise of the technology, we have to move the industry up the learning curve. The best way to do that is to provide a way for people to come together and share their experiences - both successes and failures.

Such a cooperative group needs to encompass all different types of AI systems. In past steps of technology evolution, vendors have created vendor-specific user groups such as the San Francisco Apple Core or IBM's SHARE. With AI intelligent agents, it is not unusual for an application developer to use more than one LLM and choose them from different vendors. Today, these LLM choices are based on considerations of best fit for functionality, hallucination mitigation, local resources available, latency, etc. In the future, LLM security will also be a decision criterion. LLM evolution is also increasing at a rapid rate. This can further complicate LLM suite choices. For these reasons, an organization that brings knowledge and experience with all types of AI systems is important. It will also be important to include people from both the LLM development communities and the application development communities.

In previous generations of technology, this was done in groups that operated on a principle of coop-etition. That is cooperation on fundamentals that creates a foundation that each participant can build on to compete. This kind of cooperation will lead to more effective, secure applications delivering the productivity benefits being sought.

Such a cooperative organization could develop tutorials and best practices in a wide ranging set of areas including: effective defenses against prompt injection attacks; architectural structures that prevent or minimize damage from cybersecurity attacks; selecting best functional areas for AI automation; orchestrating suites of LLM(s); handling hallucinations, dealing with end user acceptance, deployment, maintenance, etc. problems; meeting up-time / reliability requirements; creating requirements for future LLM creation, and technology projection to help in life cycle management of both applications and LLM's. Developing and distributing this type of knowledge will be one of the valuable functions that a sharing organization can perform.

There has been previous work done on the possibility of a similar sharing organization dealing with the societal effects of GenAI. As can be seen in the illustration below, the types of expertise required for sharing security / application expertise and for sharing societal adaptation expertise have a great area of overlap.

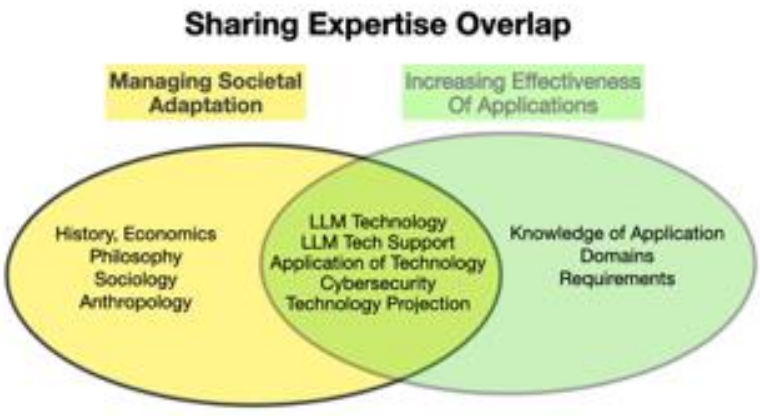

Figure 1: The Overlap Between Societal Adoption and Effectiveness of Applications
click to enlarge

Because of the overlapping areas of expertise and potential memberships, combining the two will produce efficiency increases. Maybe more importantly, work in the two areas - speeding up deployment of quality applications and speeding up societal adaptation - will catalyze both groups. Progress in each area will stimulate the other. Creating better and faster results.

Therefore, it makes sense to combine these two work areas within one organization. They both share the objective of finding ways of maximizing the benefits from AI while minimizing the downsides.

Conclusion

GenAI and intelligent agents using it have the potential for significant productivity improvements. To achieve the full productivity benefits, we need to learn how to secure, develop, and deploy these technologies. Currently, it appears that we are very low on the learning curve. As an industry, we are just beginning to understand the security vulnerabilities and how to develop effective applications of the technologies. The best way to quickly move up the learning curve and capture the full productivity benefits is to create a way that implementers can share experience and develop best practices. Combining that with similar work on societal adaptation to the technology will turbocharge both. The Bace Cybersecurity Institute (BCI) is exploring the creation of such a sharing organization. To learn more or become part of the discussion, go to https://www.bacesecurity.org/form/aiwg