www.pipelinepub.com
Volume 20, Issue 6

The Dark Side of Generative AI: A Taxonomy of Negative Possibilities

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

Much has been written in the last year about the negative side effects of generative AI. This article provides a good general understanding, including a general background and taxonomy, of the negative side effects, as a way to organize thinking about them.

The development of generative AI (GenAI) started in approximately 2017 and its power is rising exponentially. This exponential growth is likely to produce an expanding palette of benefits as well as negative side effects.

GenAI technology is based on Large Language Models (LLMs). LLMs use symbols. Symbols can include alphanumerics, phonemes, images, brush strokes, musical notes, etc. LLMs consider all symbols as a part of a language model. LLM-based systems operate on the basis of probability. An oversimplified explanation of LLMs is that they start with a symbol (letter, word, ideograph, image, etc.) and determine what symbol is most likely to follow. Based on this process they can create new content. This ability to create new content is a fundamental differentiator from previous forms of AI resulting in a quantum leap in capability. To achieve this, these models are trained on massive data sets. For example, data sets crafted using internet/web crawlers. That is, a capture of all the information that is publicly available on the Internet. Some models are also trained on private information. In conjunction with this data, the models use a huge number of parameters—tens of billions or more. How these, often non-transparent and black-box, large probabilistic systems produce their results with potential negative side effects is still not well understood. Yet, the resulting benefits are so attractive that they are being implemented widely, despite the potential for negative consequences.

The negative side effects of GenAI that have been documented to date are organized into the taxonomy presented in Figure 1.

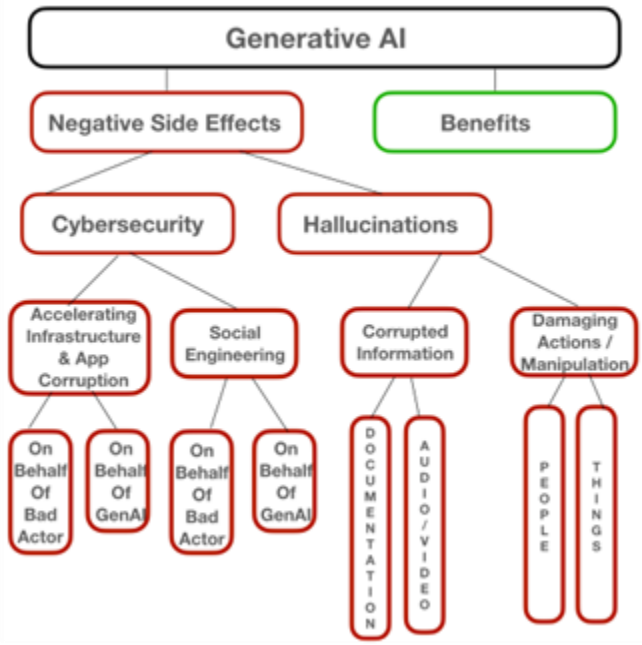

Figure 1. Taxonomy of GenAI Negative Side Effects

The first two major categories under GenAI are negative side effects and benefits. There are likely to be more negative side effects that have not yet been documented, and more are likely to emerge. However, it is probable that many of them can fit into this taxonomy.

To differentiate between positive and negative effects, financial gain/loss and quality of life (QoL) factors are used.

The first major categories of negative side effects are Cybersecurity and Hallucinations. Cybersecurity involves the unauthorized use of communications and computing resources. Hallucinations is a term that is emerging in common usage forthe range of negative side effects that involve GenAI producing what looks like convincing output, but which is actually false or unreal. The hallucinations discussion is focused on the typical types that have been observed.

Cybersecurity

Attempts are being made to control GenAI so that it's not used for nefarious purposes. But there are also well documented ways of bypassing those controls. In addition, GenAI SaaS systems have appeared on the Dark Web tailored to create cyber-attack scenarios in return for cryptocurrency. Rogue states that participate in cyber-crime have likely developed fit for purpose GenAI attack engines. As a result, the number and frequency of Zero-day attacks has been increasing.

From a financial perspective, the costs associated with cyber crime are massive. Moreover, these costs act like a tax on all goods and services in the global economy. This form of 'taxation' is

regressive. That is, it hits people with less economic resources harder than those with more. Its negative effects disproportionately impact the quality of life of those most at risk.

In addition to the economic impacts on quality of life, cyber crime threatens our modern infrastructures. Impairment of these infrastructure systems by cyber crime has already resulted in short-term negative impacts on quality of life.

There are two main sub areas of cybersecurity negative side effects: 1) application and infrastructure corruption attacks; and 2) social engineering attacks.

Application and Infrastructure Corruption

Application and infrastructure attackers seek unauthorized access to data, or unauthorized power to make cyber systems perform actions. They do this in a variety of ways that include unauthorized change of configurations, introduction of unauthorized code, etc. GenAI is exponentially increasing the capabilities of cybersecurity attackers to do these things and the current widely deployed defending technology is challenged to protect against it. It is as though attackers were acquiring bombs, while the defenders are still limited to knives.

This vulnerability stems from the fact that today's defenses are primarily static. That is, they use pre-determined (static) patterns to identify attacks and scripts to apply responses (often called remediation). Because of their static nature, they are denoted as S2 (static attack recognition, and static remediation) systems. They work well against classes of attacks that are employed repeatedly and change relatively slowly. Such pattern recognition defenses act as specialized sieves that identify and filter data for specific threats. Each sieve resembles a guardian at the digital gate watching for and detecting known attack shapes. Experts then analyze the attacks and follow step-by-step guides to counteract. The process relies on a sequence of actions, much like a cooking recipe.

The effectiveness of sieve and recipe defenses depends on the pace an attack pattern changes. Rapid attack changes leave insufficient time to prepare new sieves or appropriate remediations.

GenAI can rapidly create (generate) a very large number of new attack types. The cost of each attack launch is relatively low, meaning that not every attack has to be successful. As a result, the number and variability of attacks can accelerate dramatically. The consequences are large numbers of attacks that change very rapidly—too rapidly for patterns to be identified and installed using current defensive tools. It is extremely difficult for such systems to defend against these GenAI attacks.

These types of GenAI-created attacks can be characterized as dynamic. Since the attacks are ever-changing, an effective response cannot be easily anticipated and scripted. A different approach is needed that can respond to the dynamic nature of the attacks. Those attempting to develop dynamic defenses have to deal with two sets of problems: 1) latency; and 2) reliance on scripted remediation. Current defense systems are primarily central site based. Central site systems struggle with latency. They face the dilemma of quantity of data. The more data you have the more likely you are to be able to find an attack. However, the more data you have the longer it

takes to process it. Attack latency is falling dramatically. Central site systems struggle to keep pace. On the remediation side, these defense systems either use pre-canned remediation scripts or rely on manual intervention. In the world of rapidly changing GenAI attacks, the remediation scripts struggle to respond effectively, and manual responses can't keep up with attack cycle times.

A new approach is needed.

Social Engineering

Social engineering attacks involve manipulation to get innocent people to do things and act in ways they otherwise would not. We have seen that GenAI is capable of creating very convincing video, audio and text that appear to come from trusted sources asking for money, seeking to obtain credentials, attempting to sway public opinion, etc. GenAI is becoming an increasingly formidable weapon for social engineering attacks.

Hallucinations

Hallucination is a term that has come into common usage as a label for a particular set of GenAI negative side effects. In humans, hallucinations are perceptions of sensory data (images, sounds, tastes, smells, etc.) that do not actually exist, yet seem real.

GenAI is capable of creating outputs of increasingly "high fidelity," so high that they can be very convincing and reliable, but which upon examination may be false, erroneous, misrepresentative, etc. But whereas the use of GenAI in social engineering attacks involves intentional deception, hallucinations can be unintended outcomes caused by a variety of model- and training-related factors. Their potential negative impacts, however, are no less serious. Moreover, when a GenAI system is confronted with a hallucination, the GenAI system generally continues to maintain that it is true.

This is a critical and widespread problem. Recent studies show that for simple hallucinations depending on the particular public GenAI systems deployed, the range of occurrence of hallucinations ranges from 3% to 27% and may be even higher. For complex questions in an area such as law, rates range from 69% to 88%.

As shown in the taxonomy above, there are two general types of hallucinations: those concerning corrupted information, and others concerning damaging actions. Within corrupted information, there are two subtypes: documentation; and audio/video.

Negative impacts of corrupted documentation examples include, e.g., citing non-existent cases in legal briefs, or falsely attributing information that negatively affects reputations, etc. Hallucinations resulting from corrupted audio/video purported to be of real events that never happened. These can impact political, business, social, etc. domains.

Within examples of damaging actions are two subcategories: actions on people; and actions on things. Examples concerning damaging actions negatively affecting people include medical

systems such as medical GenAI diagnostics, lab tests, treatments, etc. Examples concerning damaging actions negatively affecting both people and things include military systems killing the wrong people, designating wrong targets, etc. Similar examples exist for police, legal, political systems, etc., and a broad range of infrastructure control systems.

Conclusion

With the foundation understanding of this taxonomy and accompanying discussion, future consideration can turn to mitigating these negative side effects.