www.pipelinepub.com
Volume 21, Issue 2

The Future of GenAI

By: Mark Cummings, Ph.D.

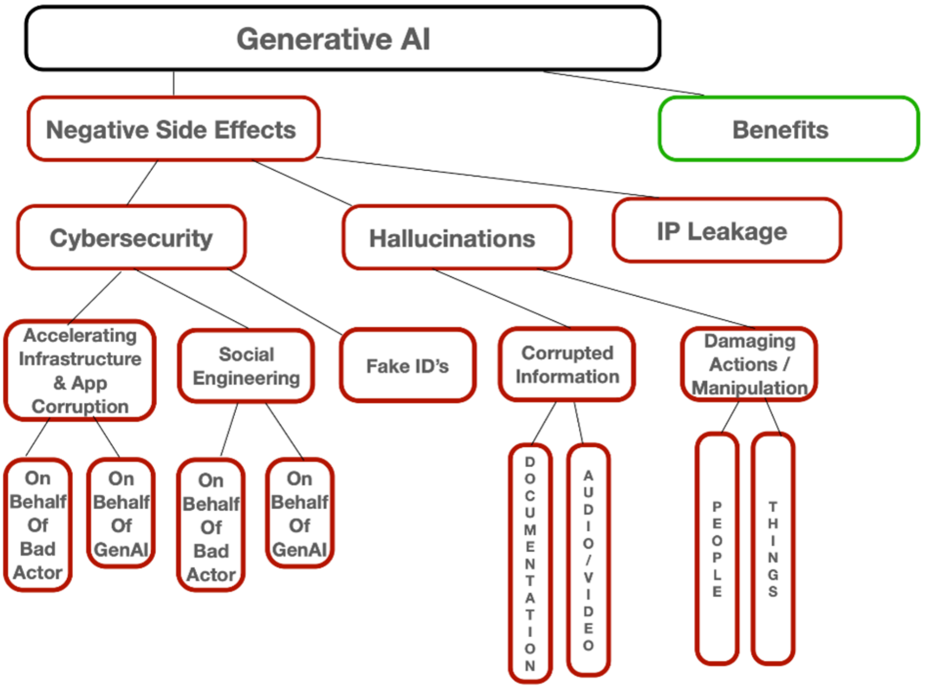
The one thing we know for sure about GenAI is that there is a lot we don't know. This is quite different from the last big tech disruption — the PC. Yet, in this environment of uncertainty, we have to make decisions. To help provide a foundation for decision making, we should consider the following: the basic things we don't know; GenAI's future; financial considerations; and regulation. This leads to a set of important questions upon which progress is possible. This background and list of questions defines a space within which we can make decisions and identify paths forward. Because of the state of GenAI, architecting systems to include GenAI is going to be an exercise in decision making in an environment of uncertainty.

What We Don't Know

This is quite different from when the first microprocessors and the first microcomputers appeared, certainly for me. I had a clear sense of where that technology was going and how it could have a positive impact on society. I don't have that sense about GenAI. What I *have* seen is that others, when they are fully honest, don't either. It is hard even to put a list together of what we don't know. But a good start might be we *don't* know:

- How GenAI conceptually works.
- What it is good for.
- What it is bad at.
- If it should be regulated.
- How to regulate it if we want to.
- Whether it can destroy humanity.

We do know that GenAI has very positive potential benefits. At the same time we have a beginning sense of what the negative side effects are — see illustration below. We don't know if they can be eliminated.



Deep Fakes are not specifically called out in the illustration. Of course they are inherent in the social engineering cyberattacks shown. But they also have much more far reaching consequences. Deep Fakes have both positive and negative effects. Hallucinations are also a big problem, as is IP leakage. One other thing that is clear is that GenAI is not going to go away. Therefore, it is dangerous to ignore.

Stages of Technology Evolution

There are three stages of technology evolution:

- Stage One - Denial
- Stage Two - Implementation of the old paradigm with the new technology
- Stage Three - Realization of the new paradigm with the new technology
-

With GenAI, we are currently in transition from Stage Two to Stage Three. It is difficult to predict how long a Stage Two to Three transition will take. In the case of the PC, it took a decade. For GenAI, Stage Two is the use of GenAI to enhance the browser function. Stage Three is unclear.

There has been talk about how autonomous vehicles would eliminate the jobs of long-haul truck drivers, and what effect that would have on society. This was before GenAI. We are possibly beginning to see the results of GenAI triggering improvements in productivity on a wider range of job types.

From the perspective of an executive in a large organization today, the correct response seems pretty straight forward. From society's point of view it is not so clear.

GenAI's Future Development

Some see GenAI getting incrementally better as we go out into the future. Most of the leading-edge models already use MoE (Mixture of Experts) architectures. These appear to have increased efficiency by directing inference requests to subsections of a large model focused on the particular domain of the request. There may be more that can be achieved along this path.

Another incremental approach is the implementation of self-reflective architectures. Users have noticed that if the results of an inference request are fed back to a model with a new request asking if there is anything wrong with the response, they often get refinements. Some LLM technologists have suggested architectures that do this internally. Because details are proprietary, it is not possible to say for sure. But it appears that ChatGPT o1 may already be doing this.

It is possible that there will be combinations of MoE's and self-reflection. Other hybrid forms may emerge as well. Some are suggesting the combination of GenAI and previously developed Machine

Learning (ML) forms. The argument here being that these previous forms (even going back to Knowledge Engineering), while lacking the power of GenAI, do not have hallucinations. Combining the forms may mitigate hallucination. There are also non-ML forms that may be combined with GenAI to mitigate negative side effects.

We may also see an ongoing increase in number of parameters. From billions to trillions and beyond. These increases may reveal new unexpected capabilities in a similar fashion to how previous increases in the numbers of parameters revealed new unexpected capabilities. However, we may be approaching the area of diminishing returns.

There will be attempts to manage the power consumption problem. As the models become larger, they consume more power. For example, there is a proposal to restart the Three Mile Island nuclear reactor to power a single GenAI data center. The increase in efficiency described above will help. But the increase in parameters will more than offset it.

There will be an ongoing back and forth between those developing guardrails for GenAI and those trying to defeat them. The movement of capable models to edge devices favors those seeking to defeat guardrails.

This process of incremental improvement is proceeding very quickly. The cycle time for a major GenAI developer is four to six months, but each developer has a different start date, and so they are leapfrogging each other. The result may be blurring the three-stage evolution process described above. New systems with new capabilities are coming out faster than organizations can figure out how to best exploit the last one.

Likely Course

All of this is happening but is probably unlikely to produce giant steps forward. Two unexpected things happened to make GenAI happen. The first was a theoretical breakthrough — the publication of the Attention paper that created a quantum jump from all the previous work on deep neural networks. Then came a series of accidents. People discovered that, as each time the number of parameters (how symbols are characterized and differentiated) increased from hundreds to hundreds of thousands, to hundreds of millions, to billions, etc., new capabilities emerged.

Thus, what seems likely is that there are likely to be four independent innovation paths:

1. Increasing the size of today's models leading to new and unexpected capabilities.
2. New ways of modifying today's models to achieve the same performance with smaller sizes.
3. Ongoing incremental changes as outlined above.
4. An unpredictable theoretical breakthrough that will leapfrog today's GenAI technology creating new and unpredictable capabilities (similar to the Attention paper and increases in parameters).

All four will move forward in parallel, with number four being the hardest to predict but, in a way, the most likely.

Financial Considerations

Some are claiming that the investment in GenAI is outstripping the returns. The problem with this line of thinking is that it doesn't consider the driving factors. For large organizations, the options today are twofold: Don't invest in GenAI and risk becoming irrelevant, or invest in GenAI and have a chance to remain relevant.

This path from relevancy to irrelevancy has happened before. Two examples come to mind. In its early days, Yahoo was the most relevant thing happening in the tech space. Then, came Google. Going back further Western Union Telegraph Company had a dominant position in telecommunications based on Telex to send data. When the PC and dial-up modems came along, Western Union's network became irrelevant and it retreated to low dollar money transfers from convenience stores.

Regulation

This is the social fabric in which discussions of regulation will occur. One of the problems that we face is that regulators have a hard time keeping up with technology. GenAI is particularly challenging for regulators. Many are struggling to understand how GenAI is different from previous forms of AI, with new challenges requiring new responses. Moreover, GenAI is developing so quickly that yesterday's answers may not fit today's challenges. Finally, the amount we still don't know about GenAI is hard for regulators to comprehend. The basic regulatory response would be, "Hang on. Stop for a minute. Let us figure this out." But there is too much money at work to stop it, and the regulators themselves face the dilemma of becoming irrelevant.

Next Steps

With all this as background, what questions bubble up to the top? That is, questions that the wider technical, business, and political communities can actually make progress on? Questions that at the same time are important? Below is a start at such a list:

- What are the implications of GenAI for general cybersecurity defense?
- In operations intensive applications (sometimes called OT)? In backup?
- How do we counter the increased capability GenAI offers to attackers?
- How do we defend GenAI systems from attack?
- Could well intentioned GenAI implementations backfire and get out of control? If so, how do we recover control? We have to consider a range of implementation domains including: military, police and public safety, international diplomacy, medical systems, critical infrastructure, finance, and maybe others.
- What are the implications for intellectual property ownership — for example, in the areas of IP leakage, copyrights and use of training materials, ownership of "products" produced by GenAI? And does ownership imply liability?
- What insurance industry issues will likely be encountered regarding GenAI developers, vendors, service providers, and users? What about cybersecurity insurance? Potential use by insurers to grant or withhold coverage?
- What about the broad societal impacts on industry, political, and social structures? Could GenAI really be a threat to humanity? Should we try to regulate GenAI? If so, how?

GenAI is powerful and the technology is rapidly evolving with new capabilities emerging quickly. This makes it exciting and captures a lot of attention. At the same time, other technologies still have very productive potential. In our excitement, at our peril, we can too easily forget this. It is important to remember that GenAI is not the answer to everything. Other technologies continue to be important.

Conclusion

The foregoing background and list of questions defines a space within which we can make decisions and identify paths forward. Because of the state of GenAI, architecting systems to include GenAI is going to be an exercise in decision making in an environment of uncertainty. The uncertainty doesn't diminish the importance of the decisions. Rather it highlights how critical they are.