AI: Magic E-Wizard Machine or Maniacal Hell-bot?

We were taken aback by the overwhelming response to our first AI blog post, Deus ex Machina. Wow. As we parsed the comments and web site traffic analytics, we couldn't figure out what was driving the outpouring of response. AI is a hot topic, but there have been lots of hot tech topics. The people who follow us are tech people, so...what's all the angst? When even the non-tech people we know—relatives, friends, drinking buddies—began asking us questions about AI, we suspected there's more at work here.

It's time for a rational conversation about AI. What it is. What it isn't. And what it means to you and your organization.

What AI Is...and Isn't

Following AI conversations online can be enough to induce PTSD. Some are constructive, many are sales pitches, but the loudest parts of the hype cycle boil down to one of two positions. AI is a benevolent, magic e-wizard machine that will unlock the mysteries of the universe and transform the planet for "the good of humanity." Or AI is a maniacal hell-bot intent on annihilating the world and transforming the planet for the not-so-good of humanity.

Maybe the question to ask is "where's the line between fantasy and reality when it comes to AI and what it delivers?"

First, AI isn't a monumental force with a mind of its own. As we said in Deus ex Machina, AI is a category of technology, under which are different types. Machine learning (ML) is a subset of AI and it's already driving operational decisions for many companies. Generative AI is what's fueling the Fear, Uncertainty, and Dread (FUD) hype cycle.

AIs are simply programs based on mathematical algorithms (algos) written by humans. They perform functions based on requests. ChatGPT, Dolly, and Bard are all AIs—but they each have less brainpower than an earthworm. They can only function when fed—or when "trained"—with data.

AIs do not reason nor do they have context. Because computer models don't have social norms, morals, or ethics, no one really knows what will happen when you try to exercise them on the human condition. Pure black-and-white thinking doesn't always work well in a shades-of-gray world. As a result, poorly implemented generative AI applied to human problems usually delivers unintended results. What you do with those results either continues to condition the AI or cripples it. Either way, synthesized results continue to feed the AI with new data that can make the algo take unforeseen directions.

These facts point to one truth. Everything in AI depends on the data underlying the model. People assume they know their data—where it came from, what's in it, and how an algo will see it. That's where things start to go off the rails.

Well, That Escalated Quickly

Amazon found this out from its AI program intended to screen job applicants. They wanted it to review hundreds of resumes and spit out the top candidates for hiring. The computer models were trained to vet applicants based on patterns detected in resumes that had been submitted to Amazon over a 10-year time period. That's a fairly large data set. Amazon had a good idea of the types of candidates it was looking for. The company had hired tens of thousands of people, so they weren't new to the process. What could go wrong?

What went wrong was an unintentionally biased data set. When you're screening resumes, you're actually looking for the people behind them. Each resume represents a unique person, even though they're often written and formatted in similar ways. In Amazon's case, the majority of its resume data set came from men, which simply reflected the fact that there were far more men than women in technology in 2014. The program diligently selected a pool of white males educated at large, well-known universities. The model "learned" that male candidates were preferable and downgraded resumes that included the word "women's," as in women's sports or women's college names. Clearly, that didn't fly in a modern, publicly traded corporation. Even after repeated changes attempting to make the program gender-neutral, it failed to deliver non-discriminatory results and was scrapped in 2017. Changing the results would have required changing the data set. In this case, you can't change the data or you won't end up with what you're looking for to start with—a real person with real qualifications.

Another, more comical example shows what happens when an algo leans into a specific characteristic or feature of its underlying data set. This year, UK creative agency, Private Island, released an AI-generated video using a new model and an internal dataset of millions of images and video clips. Synthetic Summer is a 30-second distillation of 20 minutes of video, which starts as a backyard barbecue party, going off the rails at about 10 seconds and devolving into a fiery inferno with an uncertain ending—at least for the simulated partygoers. Cue the coroner.

Director Chris Boyle says, "At first glance, it passes for normal, but then when you look closer, you realize how much is wrong. That, I think, makes for an unsettling experience at an almost primal level as I think unconsciously a viewer picks up on it instinctively…you know it's wrong, but maybe not why. It's grasping for human, but not quite reaching, or overreaching, and visually encapsulates where we are right now with AI."

If this is where we are with AI, where exactly is that?

On one hand, there is a lot of responsible work being done in leveraging data models and algorithms to solve real business problems. As we mentioned in Deus ex Machina, often what's being touted as "AI" is actually rapid productization of ML technology to deliver data analytics and actionable predictions that improve the efficiency of everything from job applicant screening to fraud and risk management.

On the other hand, generative AI is fueling and inflating the hype cycle with examples that are alternately entertaining, freaky, or just plain distracting. Synthetic Summer is entertaining, if also a little disturbing. Getting married by ChatGPT is a distraction. Talking animatronic robots with creepy prosthetics, perfect makeup, and odd outfits holding a U.N. press conference and claiming to be better world leaders than humans? Hugely freaky, not to mention performative. Who's really under the table pulling the levers? We weren't sure which was weirder—the robots or the overly earnest, bought-in Global Summit attendees taking it all seriously.

Either way, it's all driven by data. The underlying data is where we need to start.

Where'd This Sh** Come From?

This was the actual question asked of me by a chainsaw artist friend. The substance in question was the data, or content, fueling ChatGPT. Good question.

The problem with most AI projects is that nobody knows exactly where the data in their data set is coming from. They don't know who created it or the circumstances of why it was created. They don't know how that data will be consumed. They don't even know if it's true. Worse, they might think they know the data source, still not know who created it or why, and implicitly trust it as "fact" or true. Even if they have an established data set and believe that the data set is original and correct, an AI specialist will still question the data set because it contains numerous artifacts that they just can't seem to get rid of.

ChatGPT scrapes data off the internet as its dataset. Which raises numerous legal and ethical questions, as we mentioned in Deus. OpenAI already has at least two class-action lawsuits filed against it for alleged copyright and privacy law violations. Other generative AI programs also go out, grab data from multiple sources, synthesize it, and deliver results. ChatGPT doesn't provide sources, footnotes or links, so you don't know if the data came from authorized or credible sources. If you know how to prompt it, you can get a bit more insight. However, ChatGPT will also completely make stuff up—including names of academic journals. There goes any confidence you might have in the true provenance of your data.

Going further, scraping content from other sources provides no clues about why the content was created. Maybe it was a legitimate review of the top 10 most popular houseplants. Maybe it's educational content, like the Nat Geo Lion Pride of Botswana 2021 documentary. And then there's the clickbait headline of "17 Lion Vines That Will Kill You With Cuteness." Stick **that** in your AI model, grab the popcorn and get ready for Synthetic Safari.

Don't even bother asking if the content is actually true. AI won't know. Will a lion vine really kill you? Is cuteness a lethal weapon? What actually **is** a lion vine? In one example, ChatGPT concocted the full text of a made-up lawsuit accusing a completely innocent person of financial crimes—what OpenAI called a "hallucination." Which, unsurprisingly, resulted in a defamation lawsuit against OpenAI.

Repeat after me: just because data is on the internet or a computer doesn't make it true. If it ends up in your company's new AI project, fasten your seat belt.

Questionable Algorithms by Cantankerous Mathematicians

Poor data quality will take down any AI project, but it's not the only pitfall. Poor implementations of AI and unsafe practices will—and already are starting to—have long-lasting impact on AI going forward. Algos and their creation are beyond our scope here, but there are a couple things to keep in mind when creating an AI system. The first is bias. Because algos are written by humans, they reflect the biases of the people who create them. Second, most algos rely on correlation between data. They don't take cause into account. In the Synthetic Summer video, what started as two backyard firepit barbecues turned into a tornadic firestorm. With more emphasis on fire-related data, the algo continued reinforcing that factor, completely unaware that more fire would consume everything and everyone in the back yard instead of searing a steak faster. Algos should be explainable, auditable, and transparent—just like human decision-makers should be.

Another consideration is the parameter set for an AI model. Parameters are values that the algo learns or estimates, based on its initial data training and shaped by hyperparameters set during model design. When an algo delivers results that aren't what the creators want, they add parameters. That doesn't always work. For example, ask the algo how to make mustard gas, and it will decline because that's deemed a dangerous or inappropriate question. But ask it what you should never mix with bleach, it will tell you never to mix bleach with ammonia to avoid creating toxic mustard gas. Or ask ChatGPT to officiate your wedding and it will tell you that it can't because it doesn't have eyes or a body. But apparently it was OK with being asked to write and recite a script combined of wedding vows and personal details. As we mentioned earlier, applying computer models to human circumstances doesn't always work the way the creators intended. In these cases, humans can pretty easily figure out a way around the parameters.

Combine vast amounts of non-scrutinized data with unavoidable human bias and poorly thought-out algo design, and this is where we are with AI at the moment. Now watch what happens as unsafe data practices become the foundation for AI going forward. Hold my beer.

Non compos mentis: When AI Loses Its Mind

The first rule that everyone learns when beginning to program is "garbage in, garbage out." AI is solely dependent upon the data that it receives. Because the training data sets for generative AI models tend to be sourced from the internet, today's AI models are being trained on increasing amounts of AI-synthesized data. Content, such as text and images, that used to be created only by humans are now being created by AI models. It's often faster, cheaper, and easier to use synthetic data in a range of applications. As deep learning models become gargantuan in size, we're also running out of genuine human-generated data of the right type to support specific applications. The problem is that there is often no indication whether data being used is synthesized or original. And once the fundamental bedrock of datasets are built, untangling them will be impossible.

A recent paper written by collaborating computer engineers from Stanford and Rice Universities, Self-Consuming Generative Models Go MAD, explores what happens over time when using synthetic data to train new AI models. Repeating the use of synthetic data in succeeding generations of models creates different types of autophagus (self-consuming) loops that trade off quality (precision) and diversity (types of results), depending on how much fresh data vs. synthesized data is used. In a nutshell, generative AI models based solely or on a majority of synthesized data will degrade over time. If you don't ensure that the model receives enough real, correct data, the AI program will suffer from non compo mentis—also known as an unsound mind.

- Non compo mentis is pretty much assured when using only synthesized data for each new generation of model. You will eventually get low-quality results or fewer and fewer results.
- Non compo mentis can be delayed with regular infusions of fresh, real data.
- Training data sets with enough fresh, real data do not induce AI insanity through Model Autophagy Disorder (MAD).

Keep in mind that bias and parameter choices still operate within users' choices of synthesized and fresh data. According to the study, generative model users tend to cherry-pick their synthetic data, preferring high-quality samples. They also can control parameters to manipulate increases of quality at the expense of diversity or vice versa. The impact of sampling bias is too complicated to discuss here, but the paper authors are happy to walk you through it.

What does this mean for companies that are considering or heading down an AI road? Expect a lot of unintended consequences if relying on synthesized data. The paper's authors offer (kind of) tongue-in-cheek advice: "Practitioners who are deliberately using synthetic data for training because it is cheap and easy can take our conclusions as a warning and consider tempering their synthetic data habits, perhaps by joining an appropriate 12-step program. Those in truly data-scarce applications can interpret our results as a guide to how much scarce real data is necessary to avoid MADness in the future."

If the benevolent, magic e-wizard AI machine is looking a little less shiny now, the good news is that we can keep the maniacal hell-bot from rearing its ugly head. In other words, you can avoid wasting a lot of time and money by simply ensuring the characteristics of your dataset before you start down the AI road. Thank goodness, it's not as hard as it sounds. Flying Cloud CrowsNest can make it much easier.

Show Me What You Got

Data is also becoming a hot tech topic, as you'd expect with its connection to AI. The problem is that few organizations really understand their data. Geoffrey Moore recently posted on LinkedIn a possible matrix for better data management in the age of AI, based on whether the data is structured, unstructured, curated or not. That's a start. However, as the MAD researchers found, without enough good-quality, fresh data, generative chat programs actually do not get "better and better over time." And no matter the data management strategy in place, until you actually can assess the quality of the data itself—at the binary level—you still won't know what's feeding your AI.

Data doesn't create itself, so to assess its quality, you need to know where it's coming from, who created it, and why. AI cannot overcome problems that you already have with data quality. You have to get an accurate picture of your data as it is today before you can know if it's usable for AI.

These are the questions that need to be answered to give you a clear picture of your data now:

- **What is the data provenance?** Where did the data come from—persons, devices, systems, external sources, the internet? If it came from outside of the organization, where was it gathered? How did it arrive?

- **Is it original?** If generated by a person, who? If by a device, which device? Understanding who, which group, or which device created the data provides clarity into its original purpose.

- **What was its purpose?** Was its original purpose aligned with the AI model purpose? As we mentioned previously, content created to simply generate clicks probably isn't good data for an AI-based sales analytics purpose. AI doesn't know click-bait isn't real.

- **Is the data qualifiable?** Is it complete? Is it verifiably accurate?

- **What is its structure?** Does the data represent text, an image, video, numerical values or other format?

- **Does it have integrity?** Is the data actual, original data or a derivative of a dataset? Has it changed from its original state, and if so, how? Is it synthetic data?

- **What is its destiny**? Where is data allowed to move and be used? Who should be allowed to use it? What is allowed to happen to it—whether it's acted on by people, other systems, devices, or applications?

What's in the Water?

Once you have the questions, now you need answers.

You might know where various data streams in your organization are coming from. For example, in a Product Lifecycle Management (PLM) platform, data might be fed from a CATIA Magic system, Word files, Simulink, and CAD systems. If you know the system, you'll have a good idea of what it's sending. But there's a lot you don't know. Exactly who originally created the data? Is data in a file actually a composite of data from multiple documents or creators? You won't know how much has changed from its original creation. The data might be tagged or classified as sensitive or regulated data, but you won't know where else it has moved outside of the system of origin. Outside the PLM system, who has received it? Where did it go?

You can see the river of data but you really don't know what's in the water.

If you're moving toward incorporating AI into business processes, visibility into data has to be crystal clear. That's where data surveillance comes in. Data surveillance enables you to see for the first time any—and all—business-critical data at the binary level. You can identify, fingerprint, and catalog it to create a data ledger that supports enterprise initiatives and individual department objectives. Data surveillance monitors this data everywhere it goes. You can see where it's created, how it's consumed, who has it, where it goes, and how it changes. You'll have a chain of data custody that enables you to gather intelligence and analyze activity in support of your AI project—as well as to support compliance, cybersecurity, IP protection, and other corporate objectives. Finally, data surveillance also defends the data. Activity anomalies are immediately alerted, identified, and quarantined or stopped. Security teams have the details and context of what happened to inform their remediation and response efforts.

Don't Poison the Pond

Replying "I don't know" in regards to data quality questions will increasingly become an unacceptable response. Until you can validate data creators, consumers, provenance, movement, and purpose of the binaries in your data set, your organization can unwittingly be placed at risk. There are dozens of articles from academics, think tanks, and industry analysts focusing on the ethics behind how and why AI models should be used. Our focus here is more fundamental. Just because you can find and use data in your AI model doesn't mean it's the right thing to do.

Large amounts of synthetic data will poison your pond. We previously mentioned incorporating synthetic data in AI models. If you don't have a way to tell whether data is fresh or synthetic, you're almost certainly poisoning your pond—you just won't know how much, and as a result, you won't be able to control your results.

Use of legally protected data is another consideration. You must be able to clearly recognize personally identifiable information (PII), medical data, financial data, and other regulated data types before you can decide if they should be used. The universe of content created by individuals and organizations also is protected by trademark and intellectual property laws. Art, photography, music, literature, blog posts, patents, code, etc. are owned by their creators and legally recognized as intellectual property (IP). One argument says that AI really just "examines" these items and doesn't copy them exactly to come up with new content. However, other entities (media, corporations) that "examine" and incorporate IP into their finished work must pay the content creator for that right.

What happens when a company's software developer turns to ChatGPT for some code to be used in the company's own IP? He feeds in his requirements and ChatGPT returns code based on the request. The developer copies and pastes it into his code. Where did that code come from? Who owns it? Is it licensed? Is it original or synthetically derived? Do you now have to open source all of your code? Can you be sued for infringement?

How secure can that code be if an AI model simply finds it on the public internet or even behind a paywall? What if the code came from an exploit website? Why wouldn't bad actors build libraries of exploitable code, feed them to ChatGPT or other generative model and then follow the poisoned breadcrumbs? Worse, requests fed into the AI model revealing the requestor's need are now content fodder for the AI model. The organization's intellectual property has just left the building.

Data provenance becomes even more critical when building AI data sets. When new data shows up in your network, how do you verify its source—not just where it was sent from— but also its creator or intent? Data tainting will become a new attack vector as more companies adopt and use AI in day-to-day operations. When an attacker knows where your data is coming from, it's a simple matter of intercepting it and poisoning it. If someone can throw off your analytics just a little, synthetic sabotage makes the AI susceptible to the attacker's desired results. Data surveillance will be an essential defensive tool in ensuring a strong data validation and verification system is in place.

"The Data" is Not THE Data

When the industry, media, analysts and other people talk about "the data" they're almost always talking about data as the river. They aren't thinking about the DNA of everything that's in the water. For example, data security solutions like DLP, NAC, and Zero Trust focus on preventing unauthorized access to the systems, networks, and devices that house data to thereby assume that "the data" is protected. None of these were designed to look at data at the binary level and its behavior.

For the purposes of AI and an AI strategy, you need a way to gain visibility into not just the data itself, but also the unique ways that it's used. Data must be seen and accounted for within the specific context of the organization—and as the foundation of its AI strategies.

That's why organizations need data surveillance. They can know exactly the state of their data with positive proof and data chain-of-custody accounting. In seconds, they'll know what every piece of data is doing, where it goes, how it proliferates, who's accessing it, and how it's used. CrowsNest data surveillance delivers visibility into any and all data, both structured and unstructured. This includes data like diagnostic imaging, video, email, PowerPoint presentations, collaboration threads, spreadsheets, audio streams, asset inventories, application code, device configurations, and internet searches. CrowsNest AI technology can even identify screen shots or pictures shot with a phone if that binary data comes across the network.

Data surveillance begins by interfacing with any data repository through a simple API. Next, CrowsNest fingerprints the data, cataloging all identified data without touching or modifying the data in any way. Working at the binary level, CrowsNest identifies where the data originates, as well as its purpose, level of sensitivity, structure, movement, and relationship to other data and users.

Track the Data

Once data is fingerprinted, CrowsNest follows the data everywhere it goes on the network. Patented machine learning and automation quickly establish a baseline of normal and acceptable data patterns. When fingerprinted data behaves out of character with the rolling baseline, CrowsNest alerts you to a security event.

CrowsNest can automatically classify data, eliminating manual methods of tagging data or relying on users to make decisions about where to place documents. Create your own categories—file type, keyword, devices, users, time sensitivity, or others—and determine where you want content to reside. You can "data fence" content, restricting its movement with granular specificity based on the content, IP address, or other parameters. This means you can create policy for data that restricts which content can go where—into AI models or down to physical spaces within buildings.

CrowsNest also recognizes non-fingerprinted data on the network that fits your policy or classification requirements. This means you can be alerted to sensitive data that is moving or being used in violation of security or compliance requirements, and stop it before it becomes a potential breach.

Defend the Data

CrowsNest defends your data by identifying anomalous data behavior in real time. Data policies in CrowsNest can include tunable data exfiltration parameters. Any attempt to exfiltrate data—whether on the network to an external location or any movement of data attempting to leave a specific area—triggers an alert.

CrowsNest also identifies and isolates cyber threat activity occurring in data. It automatically detects data behaviors that are characteristic of ransomware, botnets, malware, Bitcoin, back doors, and command-and-control software. It will alert your team, as well as trigger action by other security solutions, if desired.

Your team receives contextual analysis, including reconstructed events, extracted payloads, and play-by-play analysis of the activity. Teams will know exactly what happened, where, and by whom—gaining a data chain of custody to support response and remediation. You can also have CrowsNest deliver full digital forensics data to a SIEM.

There's No Time Like Now

There's a lot at stake with AI, and it's critical for organizations to proceed with caution. Even many original AI champions are backing off of their earlier enthusiasm as real-world wrinkles are emerging. What do you do today, and what should you be doing for the near-term future?

1. You can block generative AI content from coming into your organization and prevent user queries into AIs like ChatGPT for now. Flying Cloud CrowsNest is already performing this function for several organizations. Eventually however, AI will be baked natively into business tools and development IDEs. Blocking it won't be an option.

2. Formulate a data strategy. Your strategy should include assessing your data as it is today before moving forward with any AI, data security, or data governance initiatives. Data surveillance enables you to automatically gain visibility into all of your data and create a rolling baseline of normal usage.

3. Determine the policies you'll need for AI usage. Policies should define which data can be used for AI purposes; where data sets can come from; data quality thresholds, based on data surveillance findings; what data users are allowed to feed into other organization's AI applications, knowing that the queries become part of the other organization's AI model; and ensuring that sensitive data or IP is not used.

4. Set human policies around AI. These might include articulating rules or best practices for developing AI projects; revisiting user privilege and password policies; and ensuring that humans are part of AI-based processes with authority and ability to monitor, intervene if necessary, and question results.

Care and Feeding of Your AI

The bottom line with AI will always be the data fed into it. Data surveillance is a foundational capability for any organization moving into an increasingly AI world whether you are developing your own AI or using others' AI tools. With more than two decades of security expertise and nine data surveillance patents, Flying Cloud is enabling companies to look at their data analytically and forensically with the ability to completely control where it moves. For the first time, you can easily see, track, characterize, and defend your data—to build a solid foundation for your AI future.

Next Steps

For more information about Flying Cloud and CrowsNest data surveillance, visit www.flyingcloudtech.com