AI and Its Impact on Cybersecurity

Part 3: Your Inbox's Invisible Guardian

Now that we have looked at how AI, or Large Language Models work, let's look at some of the first places you may have encountered AI whether you knew it or not. While modern uses of AI seem new, they have been in use for years, just working behind the scenes. In this section we will look at how AI protects your email inbox by filtering spam.

AI and Your Inbox

Here is an area where AI has been protecting you for years without you even knowing it: your email spam filter. Have you ever wondered how your email provider knows what emails to send to spam and which ones to send to your inbox? Spam filters didn't start out using what we know as AI, it was more of an evolution of adapting technology as it got smarter. Let's look at why those emails from a Nigerian Prince asking for your help with his vast fortune and other junk emails get sent to spam, mostly!

In the 1990s and into the early 2000s, the use of email was becoming more readily accessible and working its way into not just the corporate world, but everyday life. This gave the bad guys another way to play out their cons. No longer did they need to be face to face with their victim, or even on a phone call. Now they could work their cons from a computer anywhere in the world and no longer was it one at a time, now they could reach hundreds if not thousands at a time with their scams. Email providers had to find a way to protect their users to keep confidence up and people using their services.

Keyword Blocking: The First Defense

Email service providers started with keyword blocking. This was simply having a list of "bad" keywords and phrases most used in spam emails. This list would be developed by the system administrator. Some examples would be "free money", "click here to collect your prize", "congratulations you won", or "I am a Nigerian prince and need your help." If the email filter saw any of the words or phrases on the naughty list, it would send the email to the spam folder. Think of it as the TSA no fly list! If you are on the list, you do not get through! This method was chosen because it was simple, fast, and could be implemented easily given the limited computing power at the time.

As is usually the case, the bad guys figured out how to bypass the naughty list. They simply misspelled the words. Instead of "free money", it would be "fr33 m0n3y" or use some creative spacing "F R E E M O N E Y", all to bypass being detected by the filter when compared to the administrators list. This cat and mouse game between spammers and filters has been going on for decades. Another issue was if the administrator had words or phrases that could be part of a

legitimate email, like “prescription” from your doctor, or “account verification” from your bank, those emails would be flagged and sent to spam. The keyword approach was not enough.

Bayesian Filtering: The Probability Approach

Computer scientists realized a better way than keyword filtering needed to be developed. They wanted filters that thought like humans do, able to weigh the evidence and decide rather than following rigid rules. In the early 2000s, a better way was developed. It was based on theory developed in the 18th century by a mathematician by the name of Thomas Bayes whose work on probability was the foundation for this new method called Bayesian filtering, of course. What Bayesian filtering did that keyword filtering did not was instead of just comparing words to a naughty list, it looked at the words in the email and gave it a probability that the email was spam, based on patterns it had seen before. Does this sound familiar? Go back to part 2 and see our discussion on patterns.

Think of Bayesian filtering like your favorite TV detective solving a crime. Instead of just checking if a suspect's name appears on a most wanted list, the detective weighs multiple clues and analyzes evidence to calculate the probability of guilt. Bayesian filters do the same thing with emails. Here is a simple example. Let’s say the spam filter analyzed thousands of emails marked as spam and thousands of emails marked as legitimate and it noticed the following:

- The word “Nigerian” appeared in 98% of spam emails, but only 0.5% of legitimate emails
- The word “meeting” appeared in 60% of legitimate emails, but only 2% of spam emails
- The word “free” appeared in 80% of spam emails, but also in 30% of legitimate emails

Now when the filter looks at an email and sees “free” it no longer automatically sends it to spam but will now look at all the other words in the email and calculate a probability of whether it is spam or not. “Join us for a free lunch meeting” would most likely be listed as legitimate, while “free money” would be calculated as spam.

The Bayesian filter was a significant step towards AI because it introduced two concepts:

- Pattern Recognition. The filter was no longer following fixed rules. The filter was now looking at the whole email for patterns and deciding based on those patterns.
- Learning from the Data. The more emails that were analyzed, the better the filter got in calculating probabilities. It was learning from experience. This method was still not as good as the AI we use today, but it was a huge step in that direction.

From Bayesian to AI

Today's spam filters have evolved far beyond Bayesian filtering into sophisticated machine learning systems that would make those early filters look primitive. Modern spam filters don't just look at individual words and calculate probabilities. They analyze dozens of factors all at once:

- Email headers and routing: Where did this email really come from? Is it pretending to be from your bank but originated in another country?

- Sender reputation: Does this sender have a history of sending spam? Have other users marked emails from this source as spam?
- Link destinations: Where do these links lead? Is the link text saying "yourbank.com" but pointing to "y0urbank-secure.xyz"?
- Image content: What's hidden in those images? Spammers often hide text and code in images to bypass filters.
- User behavior: Does this person usually open emails from this sender? Do they regularly interact with this type of email?
- Context and language patterns: Does this sound like a real human wrote it, or does it have the telltale signs of a mass-produced spam email?
- Similarities to known spam campaigns: Is this part of a larger spam campaign that the filter has seen patterns from across millions of other users?

These modern filters use the same principles we discussed in Part 2: analyzing massive amounts of data, recognizing subtle patterns across multiple variables, and making predictions. They're learning systems that adapt as spammers develop new tricks. When a new type of spam emerges, the AI filter sees it across its user base, learns the pattern, and starts catching similar emails, all without anyone having to manually add new keywords to a list.

The AI Advantage

Think back to the four comparisons we made in Part 2 about why AI does pattern recognition on major steroids. Modern spam filters demonstrate every one of these advantages:

- Scale: They analyze billions of emails across millions of users, learning from every single one. Your spam filter isn't just learning from your email, it's learning from patterns across its entire user base.
- Scope: They look at hundreds of variables at once, not just a few keywords. Everything from sender reputation to link destinations to writing style gets analyzed simultaneously.
- Speed: They can identify new spam patterns in hours, not weeks or months. When a new spam campaign launches, modern AI filters spot it and adapt almost immediately.
- Subtlety: They catch sophisticated spam that looks almost legitimate because they understand context. That email that looks like it's from your CEO asking you to buy gift cards, AI knows that's not how your CEO communicates.

The Evolution Complete

The evolution from keyword blocking to Bayesian filtering to modern AI spam filters perfectly illustrates the journey from simple algorithms to true artificial intelligence. What started as a bouncer with a simple list became a sophisticated security system that learns, adapts, and gets smarter every day.

And the best part? It's all happening invisibly. You probably haven't thought about your spam filter in years, and that's exactly how it should be. When technology works well, it disappears into the background. That's AI working behind the scenes, using pattern recognition to protect you from the estimated 14.5 billion spam emails sent every single day worldwide.

This same pattern of evolution—from simple algorithms to sophisticated AI—happened in other areas of the internet too. In Part 4, we'll look at how search engines made this same journey, transforming from simple keyword counters into systems that can practically read your mind. We'll explore how Google went from just counting words on a page to understanding what you're really looking for, even when you can't quite put it into words yourself.

Until then, take a moment to appreciate that spam filter working quietly in your inbox. It's been one of AI's longest-running success stories, protecting billions of people every day. And please consider heading over to our website, knowphishing.com, and signing up for our free weekly newsletter. Each week our goal is to inform everyday users of technology on what to look out for, the most recent dangers, and how to make sure all your technology is as safe as possible to prevent you from becoming a statistic in a world filled with phishing, malware, and ransomware! Thank you for your "attention"!