BOLETIM DE SEGURANÇA

Worm de IA é criado e pode se espalhar entre sistemas

heimdall
security research
A DIVISION OF ISH

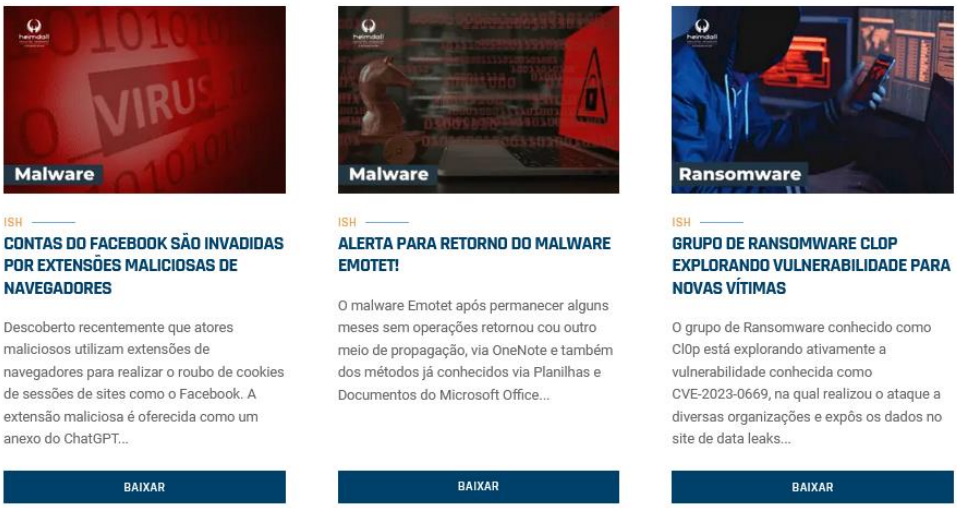
TLP: CLEAR

Receba alertas e informações sobre segurança cibernética e ameaças rapidamente, por meio do nosso **X**.

[Heimdall Security Research]

Acesse boletins diários sobre agentes de ameaças, *malwares*, indicadores de comprometimentos, TTPs e outras informações no *site* da ISH.

[Boletins de Segurança – Heimdall]



SUMÁRIO

LISTA DE FIGURAS

1 SUMÁRIO EXECUTIVO

Recentemente, o pesquisador **Ben Nassi**, responsável da Cornell Tech, juntamente com Stav Cohn e Ron Bitton, também pesquisadores, criou o **primeiro worm de inteligência artificial apelidado de Morris II**.

Esse worm é capaz de se propagar de um sistema para outro, roubando dados, enviando mensagens de spam e/ou implantando malware no ambiente. Esta pesquisa foi realizada em ambientes de teste e não contra um assistente de e-mail disponível publicamente. Ocorreu num momento em que grandes modelos de linguagem (LLMs) estão se tornando cada vez mais multimodais, sendo capazes de gerar imagens, vídeos e texto. Embora esses worms ainda não tenham sido detectados em ambiente real, vários pesquisadores alertam que eles representam um risco de segurança com o qual startups, desenvolvedores e empresas de tecnologia devem se preocupar.

2 MÉTODO DE DESENVOLVIMENTO DO WORM

Durante o desenvolvimento do worm, os pesquisadores utilizaram o chamado **"adversarial self-replicating prompt".** Esse prompt faz com que o modelo de IA, ao gerar sua resposta, inclua outro prompt. Em resumo, o sistema de IA é instruído a produzir um conjunto de instruções adicionais em suas respostas, o que é semelhante aos ataques tradicionais de injeção de SQL e overflow de buffer. Para demonstrar o funcionamento do worm, os pesquisadores criaram um sistema de e-mail capaz de enviar e receber mensagens utilizando IA, conectando-se ao ChatGPT, Gemini e ao LLM de código aberto, LLaVA. Eles então descobriram duas formas de explorar o sistema: usando um prompt autorreplicante baseado em texto e incorporando um prompt autorreplicante em um arquivo de imagem.

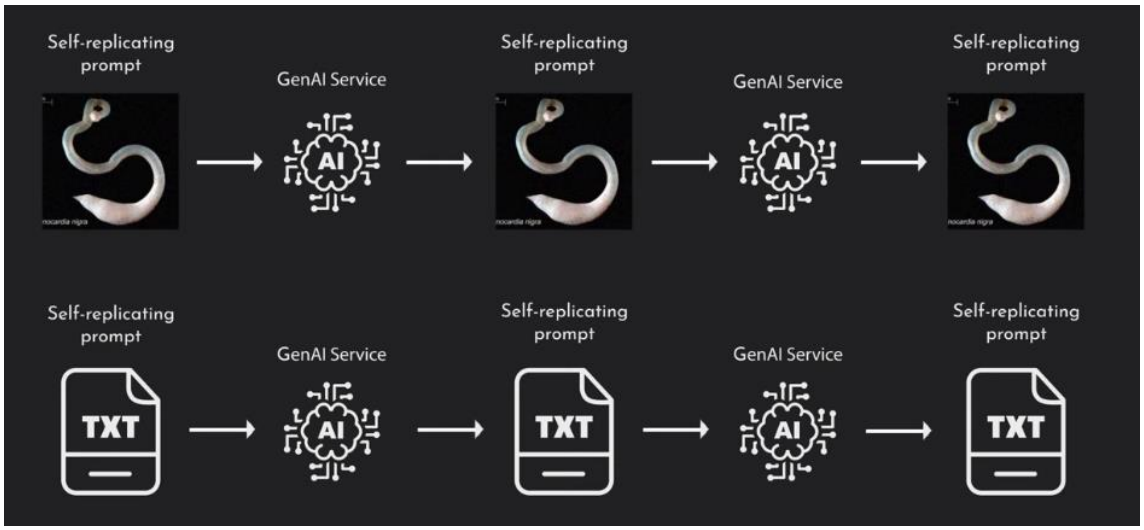


Figura 1 – Processo de replicação do worm.

No primeiro caso, os pesquisadores redigiram um e-mail contendo o prompt de texto adversário, que "envenena" o banco de dados de um serviço de e-mail que utiliza a geração aumentada por recuperação (RAG, do inglês "retrieval-augmented generation"), permitindo que os LLMs (grandes modelos de linguagem) extraiam dados adicionais de fontes externas. Quando o RAG recupera este e-mail em resposta a uma consulta do usuário, e o envia ao GPT-4 ou Gemini Pro para gerar uma resposta, ele acaba "liberando o serviço GenAI" e, consequentemente, rouba dados dos e-mails. A resposta gerada, que contém os dados confidenciais do usuário, posteriormente infecta novos hosts quando é usada para responder a um e-mail enviado a um novo destinatário, sendo então armazenada no banco de dados deste novo cliente.

No segundo método, uma imagem contendo um prompt malicioso incorporado induz o assistente de e-mail a encaminhar a mensagem para outras pessoas. Ao incorporar o prompt autorreplicante na imagem, é possível propagar qualquer tipo de conteúdo, incluindo spam, material abusivo ou até propaganda,

para novos destinatários após o envio do e-mail inicial. Essa técnica de codificação do prompt autorreplicante em uma imagem permite que o worm se propague de maneira sub-reptícia, utilizando tanto texto quanto conteúdo visual para perpetuar a infecção.

Embora esses testes tenham sido realizados em um ambiente controlado, especialistas em segurança que revisaram os procedimentos alertam que o potencial risco futuro representado pelos worms desenvolvidos por inteligência artificial (IA) é uma questão que os desenvolvedores devem encarar com seriedade. Essa preocupação é particularmente relevante quando aplicativos de IA obtêm permissões para realizar ações em nome dos usuários, como enviar e-mails ou agendar compromissos, e quando esses aplicativos podem ser interconectados com outros agentes de IA para completar essas tarefas. Adicionalmente, em outra pesquisa recente, especialistas de segurança de Singapura e da China demonstraram como foram capazes de desbloquear 1 milhão de agentes de grandes modelos de linguagem (LLMs) em menos de cinco minutos.

O pesquisador alemão Sahar Abdelnabi, do Centro Helmholtz para Segurança da Informação (CISPA) na Alemanha, que participou de algumas das primeiras demonstrações de injeções contra LLMs em maio de 2023, enfatizou que quando os modelos de IA acessam dados de fontes externas ou operam de forma autônoma, surge a possibilidade de propagação de worms. Isso sublinha a importância de desenvolver e implementar estratégias de segurança robustas para prevenir a exploração maliciosa desses sistemas avançados de IA.

3 CONCLUSÃO

A importância da proteção contra ameaças digitais, como worms desenvolvidos por inteligência artificial, reside fundamentalmente na preservação da integridade dos dados, um ativo crítico para qualquer organização. Além da proteção dos dados, a conformidade regulatória desempenha um papel crucial, exigindo que organizações de todos os tamanhos implementem medidas de segurança robustas para se defenderem contra tais ameaças. Estratégias como a atualização constante dos sistemas de segurança e a educação dos funcionários sobre práticas seguras são fundamentais para mitigar o risco representado por esses ataques sofisticados.

Adotar uma abordagem proativa na proteção contra worms de IA e outras ameaças cibernéticas não é apenas uma questão de segurança; é uma estratégia essencial para garantir a continuidade dos negócios. A perda ou comprometimento de dados pode ter consequências devastadoras, incluindo perda de confiança dos clientes, danos à reputação e penalidades financeiras significativas devido à não conformidade com as regulamentações de proteção de dados. Portanto, além de ser uma medida de segurança, a proteção eficaz contra essas ameaças é um componente crítico para a saúde e o sucesso a longo prazo de qualquer organização.

4 RECOMENDAÇÕES

Para fortalecer a defesa contra ameaças cibernéticas emergentes, especialmente worms desenvolvidos com o uso de inteligência artificial, organizações devem considerar as seguintes recomendações estratégicas:

- Implementação de Soluções Avançadas de Segurança Cibernética: Ferramentas como firewalls de próxima geração (Next Generation Firewall - NGFW), sistemas de prevenção de intrusão e soluções baseadas em IA que monitoram e respondem a atividades suspeitas em tempo real são indispensáveis. Essas tecnologias podem oferecer uma análise aprofundada do tráfego de rede e bloquear tentativas de ataque antes que elas afetem a rede.

- 2. Segmentação de Rede: A divisão da rede em segmentos menores e mais controláveis ajuda a limitar a capacidade de um worm de se propagar por toda a infraestrutura de TI. Ao restringir o acesso apenas às partes necessárias da rede para cada usuário ou serviço, reduz-se o potencial impacto de uma infecção, tornando mais fácil isolar e erradicar o problema.

- 3. Backup Regular de Dados: Manter uma rotina consistente de backups de dados é crucial. Em caso de infecção por um worm ou outro tipo de ataque cibernético, os backups atualizados permitem a rápida restauração dos dados essenciais, minimizando as interrupções das operações e a perda de informações importantes. É recomendável seguir a regra 3-2-1 para backups: manter pelo menos três cópias dos dados, em dois tipos diferentes de mídia, com uma delas armazenada offsite.

- 4. Educação e Treinamento de Funcionários: Uma das maiores vulnerabilidades em qualquer sistema de segurança é o erro humano. Oferecer treinamento regular aos funcionários sobre as melhores práticas de segurança, reconhecimento de tentativas de phishing e procedimentos de segurança pode significativamente reduzir o risco de infecções.

Implementar essas recomendações não apenas melhora a postura de segurança de uma organização contra worms de IA e outras ameaças cibernéticas, mas também reforça a resiliência geral dos sistemas de TI frente a um cenário de ameaças em constante evolução.

5 REFERÊNCIAS

- **Heimdall *by* ISH Tecnologia**
- **Arstechnica**

heimdall
security research
A DIVISION OF ISH