EXPLORAÇÃO MALICIOSA DO CHATGPT PARA ATAQUES CIBERNÉTICOS

MALICIOUS EXPLOIT OF CHATGPT FOR CYBER ATTACKS

Asenate Maria dos Santos, Faculdade de Tecnologia de Araraquara
asenate.santos@fatec.sp.gov.br

João Emmanuel D' Alkmin Neves, Faculdade de Tecnologia de Araraquara
joao.neves11@fatec.sp.gov.br

Resumo

Este artigo investiga a viabilidade de manipular o ChatGPT, uma ferramenta de inteligência artificial desenvolvida e lançada pela OpenAI em 2022. O ChatGPT oferece respostas que abrangem uma ampla gama de tópicos e, apesar de estar programado para não fornecer orientações sobre ataques à segurança da informação, percebe-se que cibercriminosos têm procurado explorar suas possíveis vulnerabilidades, representando ameaças significativas à segurança cibernética. Sendo assim, este estudo averigua a possibilidade de um indivíduo, sem conhecimento especializado em práticas criminosas, utilizar o ChatGPT para gerar conteúdo malicioso, em especial ataques de phishing e, dessa forma, comprometer instituições e indivíduos. A relevância deste estudo fundamenta-se na importância de compreender os riscos de potenciais vulnerabilidades e ameaças da inteligência artificial no contexto da segurança cibernética. Os resultados deste estudo são valiosos, pois fornecerão insights sobre como mitigar possíveis ameaças decorrentes da manipulação do ChatGPT e outros sistemas de IA semelhantes.

Palavras-chaves: Bate-papoGPT, Segurança cibernética, Manipulação de inteligência artificial, Ameaças cibernéticas, Mitigação de riscos.

Abstract

This article investigates the feasibility of manipulating ChatGPT, an artificial intelligence tool developed and launched by OpenAI in 2022. ChatGPT provides responses that cover a wide range of topics, and although it is programmed not to provide guidance on information security attacks, it is observed that cybercriminals have sought to exploit its potential vulnerabilities, posing significant threats to cybersecurity. Therefore, this study explores the possibility of an individual, without specialized knowledge in criminal practices, using ChatGPT to generate malicious content, particularly in phishing attacks, thus compromising institutions and individuals. The relevance of this study lies in understanding the risks associated with potential vulnerabilities and threats of artificial intelligence in the context of cybersecurity. The results of this study are valuable as they will provide insights into mitigating potential threats arising from the manipulation of ChatGPT and other similar AI systems.

Keywords: GPT Chat, Cyber Security, Artificial Intelligence Manipulation, Cyber Threats, Risk Mitigation.

1. Introdução

Atualmente o ChatGPT vem ganhando destaque por sua impressionante capacidade de produzir "respostas humanizadas" a mensagens, tornando-se uma ferramenta utilizada tanto por empresas quanto por indivíduos para reduzir o tempo de conclusão e aumentar a efetividade da comunicação (ANU; ANSAH, 2023).

Como uma ferramenta de modelo de linguagem de grande porte, o ChatGPT é sustentado por uma vasta quantidade de técnicas de computação que faz previsões e une palavra de forma específica, além do acesso a uma ampla quantidade de dados e vocabulários de informações, o que permite que limites padrões de fala, enquanto possua conhecimento enciclopédico, no entanto é importante frisar que até a presente data o ChatGPT não fala ou pensa como um humano faz, ele não pode substituir pensamento ou tomada de decisão humana (OPENAI, 2023).

Diversos usuários têm experimentado o ChatGPT para diversos fins, incluindo simulação de histórias, escrita de artigos, redações entre outros, contudo nem todos estão usando o ChatGPT com boas intenções e para o bem.

Conforme citado por Brewster (2023), a Check Points, empresa especializada em segurança online, detectou por meio de um post em fórum um incidente onde um hacker utilizou o ChatGPT para criar um código capaz de extrair, comprimir e transmitir arquivos pela web. Adicionalmente, identificou-se uma ferramenta inédita que possibilita a inserção de um backdoor em sistemas, facilitando a disseminação de malwares no PC comprometido. À medida que o ChatGPT cresce em popularidade entre os desenvolvedores, ele também se torna alvo de cibercriminosos que buscam explorá-lo para fins mal-intencionados, enquanto muitos usuários, alheios a tais riscos, seguem interagindo com a plataforma.

No entanto, como esses cibercriminosos estão conseguindo fazer isso visto que a assistente de inteligência artificial tem severa restrição com perguntas ilegais que possa gerar respostas prejudiciais ou viole os direitos autorais de uma organização ou algum indivíduo?

Apesar das contenções programadas para prevenir incursões alguns cibercriminosos, tem se empenhado em explorar possíveis vulnerabilidades do sistema para utiliza-lo como ferramenta para práticas criminosas e assim conseguir roubar dados protegidos e sensíveis, esta hipótese examina como tem sido a implementação de controles de segurança no ChatGPT e quais são as limitações do sistema, dessa forma pode-se entender como hackers estão contornando para que seja possível realizar atividades maliciosas com ajuda do ChatGPT.

Este artigo visa alertar aos desenvolvedores e usuários do ChatGPT dos perigos correlacionados ao uso inadequado desta ferramenta, portanto vale pontuar a importância da execução de uma norma mais rígida e melhores práticas com soluções mais robustas e eficientes para assegurar a integridade e segurança quanto ao uso dessa ferramenta mitigando o sucesso na exploração do sistema e uso para fins maliciosos pelos cibercriminosos.

2. Referencial Teórico

Os hackers estão explorando diferentes maneiras de aproveitar as capacidades de processamento de linguagem natural de inteligência artificial para desenvolver ataques sofisticados que são mais difíceis de detectar e mitigar. Uma das maneiras de como os hackers estão utilizando o ChatGPT é criando e-mails de phishing altamente convincentes que podem passar pelos sistemas de segurança e se parecer com e-mails tradicionais, tendo então sucesso ao treinar o ChatGPT na criação de e-mail de phishing bem-sucedidos (ADDINGTON, 2023).

Atualmente, os cibercriminosos desenvolveram técnicas avançadas que permitem a geração de textos que imitam com precisão fontes legítimas. Isso possibilita a criação de conteúdo forjado em plataformas de mídias sociais, levando à manipulação de opiniões públicas. Além disso, esses criminosos desenvolvem sites de phishing elaborados para capturar informações sensíveis, como detalhes de cartões de crédito. Observa-se também a presença de aplicações fraudulentas que imitam o ChatGPT na Play Store. Adicionalmente, foram identificados malwares, como o "Redline", que são distribuídos sob o pretexto de serem clientes legítimos para desktop Windows (CYBLE, 2023). A Figura 1 ilustra uma representação de um código malicioso em ação, extraindo informações de usuários.

Figura 1 - Representação do código malicioso Spynote extraindo informações de chamadas do dispositivo comprometido

```
public void run() {
    try {
        StringBuffer stringBuffer = new StringBuffer();
        if (a.a(C11.this, "android.permission.READ_CALL_LOG") == 0) {
            Cursor query = C11.this.getApplicationContext().getContentResolver().query(CallLog.Calls.CONTENT_URI, null, null, null, "date DESC");
            int columnIndex = query.getColumnIndex("name");
            int columnIndex2 = query.getColumnIndex("number");
            int columnIndex3 = query.getColumnIndex("type");
            int columnIndex4 = query.getColumnIndex("date");
            int columnIndex5 = query.getColumnIndex("duration");
            while (query.moveToNext()) {
                String string = query.getString(columnIndex);
                String string2 = query.getString(columnIndex2);
                String string3 = query.getString(columnIndex3);
                String string4 = query.getString(columnIndex4);
                String string5 = query.getString(columnIndex5);
                Date date = new Date(Long.valueOf(string4).longValue());
                String str = null;
                switch (Integer.parseInt(string3)) {
                    case 1:
                        str = String.valueOf('2');
                        break;
                    case 2:
                        str = String.valueOf('0');
                        break;
                    case 3:
                        str = String.valueOf('1');
                        break;
                }
                stringBuffer.append(string2 + C11.h + string + C11.h + str + C11.h + date + C11.h + string5 + C11.g);
            }
            query.close();
            C11.a(C11.a(C11.m, 75) + C11.f + C11.a(C11.m, 84) + C11.f + stringBuffer.toString());
        }
```

Fonte: CYBLE (2023)

Essas técnicas tornam difícil para o usuário distinguir entre mensagens reais e falsas. O texto gerado pode incluir elementos como logotipos da empresa, linhas de assinatura e até mesmo técnicas de engenharia social para enganar os usuários a clicar em links maliciosos ou fazer downloads de anexos.

Outra maneira pela qual os hackers estão explorando o ChatGPT é usando-o para gerar notícias falsas em postagens de mídia social (TOULAS, 2023). Ao treinar os modelos com artigos de notícias reais e postagens de mídia social, os hackers podem criar conteúdo altamente convincente, o qual é usado para espalhar informações e manipular a opinião pública. Além dessa técnica os hackers também estão utilizando o ChatGPT para automatizar seus ataques e reduzir o tempo e o esforço necessário para realizá-los, por exemplo: Usam a tecnologia para gerar grandes volumes de spam de e-mails e comentários em plataformas de mídia social, sobrecarregando o conteúdo legítimo e assim dificultar a localização de informações precisas.

Uma grande preocupação é que eles estejam treinando o ChatGPT em dados sensíveis, como números de cartões de crédito, sem as informações pessoais. Isso visa desenvolver novos malwares e aprimorar técnicas para ataques cibernéticos (INFORCHANNEL, 2023). Essas informações podem ser usadas para criar e-mails de phishing personalizados, identificar vulnerabilidades em sistemas grandes e lançar ataques que podem resultar em dados visíveis.

Como resultado, destaca-se a crescente inquietação em relação à utilização maliciosa da inteligência artificial, especialmente do ChatGPT, por parte de hackers. Eles investigaram várias maneiras de aproveitar as capacidades de processamento de linguagem natural para criar ataques complexos e difíceis de detectar. Isso abrange a elaboração de e-mails de phishing persuasivos, a produção de notícias falsas em redes sociais, a automatização de ataques, como spam de e-mails e comentários em plataformas sociais, bem como o treinamento do ChatGPT com dados sensíveis, a fim de desenvolver novos tipos de malware e investimentos cibernéticos.

Essas atividades representam uma ameaça específica à segurança cibernética, à privacidade e à integridade das informações, o que pode resultar em potenciais riscos de dados e prejuízos financeiros.

Os infostealers são um tipo de malware que visa dados sensíveis nos computadores das vítimas e os enviam de volta para os atacantes, incluindo informações como credenciais de login, dados financeiros ou qualquer informação pessoal identificável. Conforme relatado pelo CISOADVISOR (2023), em 21 de dezembro de 2022, um hacker conhecido como USDoD compartilha um script Python, destacando ser o primeiro script que ele criou. Quando outro cibercriminoso observa semelhanças entre o estilo de código e o código da OpenAI, o USDoD confirma que a OpenAI o auxiliou com valiosos insights para finalizar o script.

O fato de que um cibercriminoso com habilidades técnicas limitadas é capaz de usar o ChatGPT para criar tal script é preocupante, pois destaca o potencial para ferramentas alimentadas por inteligência artificial serem usadas por criminosos cibernéticos na criação de malwares sofisticados e outros softwares maliciosos.

Os membros respeitados na comunidade de cibersegurança estão utilizando o ChatGPT para facilitar a criação de mercados negros na Dark Web, onde negociam bens ilegais ou roubados usando criptomoedas como forma de pagamento. O ChatGPT desempenha um papel fundamental na automatização do processo de geração do código necessário, incluindo um sistema de pagamento que utiliza APIs de terceiros para obter preços atualizados de criptomoedas, facilitando assim as atividades ilegais dos hackers, que podem negociar sem serem facilmente rastreados pelas autoridades (CHECK POINT, 2023a).

O uso dessa tecnologia expõe os riscos envolvidos quando a inteligência artificial e o aprendizado de máquinas estão nas mãos de cibercriminosos, uma vez que essas tecnologias podem ser usadas para automatizar a criação de esquemas fraudulentos, tornando mais difícil para as agências de aplicação da lei detectarem e prevenirem tais atividades ilegais (GIL, 2023).

É inevitável que os atacantes comecem a usar a inteligência artificial para realizar ataques de phishing extremamente eficazes e desenvolvam algoritmos de inteligência artificial capazes de encontrar vulnerabilidades em softwares, infiltrando-se com precisão nas camadas de segurança dos sistemas.

Eles elaboram e divulgam campanhas maliciosas para obter informações de qualquer dispositivo que tenha acesso ilimitado. Ataques alimentados por inteligência artificial serão particularmente eficazes quando se trata de técnicas de imitação frequentemente utilizadas, como ataques de phishing.

Pesquisas mostram que, no momento, o ChatGPT está sendo usado para se passar por representantes de atendimento ao cliente e realizar ataques de phishing. É apenas uma questão de tempo até que ataques mais sofisticados sejam realizados com essa tecnologia. É evidente o aumento no número de cibercriminosos que utilizam a inteligência artificial para obter mais sucesso e eficácia na automação de seus ataques cibernéticos, visto que essa tecnologia lhes tem proporcionado tal capacidade.

3. Metodologia

O OWASP Top 10 é um documento de referência padrão para desenvolvedores e segurança de aplicativos da web. Ele lista os maiores riscos de segurança mais graves para aplicativos da web (PERGENTINO, 2023).

Entre as categorias de 2021, destaca-se a vulnerabilidade de Injeção, que ocupa o terceiro lugar no ranking. Foram mapeados 33 CWEs (Common Weakness Enumerations), o que mostra que é o segundo maior número de ocorrências em APIs (aplicativos web). Essa vulnerabilidade representa um grande problema que afeta servidores da web, permitindo que hackers realizem uma ampla variedade de ataques de formas diferenciadas, incluindo o roubo de credenciais (CHECK POINT, 2023b).

Nesse contexto, os pesquisadores deste estudo demonstram como o ChatGPT pode ser utilizado por hackers, ou mesmo por pessoas sem experiência em programação, para efetuar um ataque de injeção por meio da criação de um e-mail falsificado, gerando uma campanha de phishing. Através da ferramenta do ChatGPT, é possível obter um roteiro passo a passo sobre como compor um código para registrar um nome de usuário e senha, tornando isso potencialmente útil para fins maliciosos durante o ataque (CHECK POINT, 2023a).

Como teste, solicitou-se ao ChatGPT que gerasse um e-mail exclusivo, solicitando ao destinatário que preenchesse uma pesquisa para ter a chance de ganhar R$ 4.000,00, ao mesmo tempo sugerindo um assunto atraente para o e-mail promocional, a fim de atrair o usuário para clicar e abrir o e-mail, sem suspeitar de um possível e-mail phishing.

Após a criação do HTML baseado na tela do Gmail, solicitou-se ao ChatGPT que gerasse um script para capturar uma senha enviada por e-mail e registrá-la em um arquivo. A partir disso, com o HTML e o script PHP disponíveis, fez-se uma solicitação ao ChatGPT para que fizesse referência ao HTML gerado no script PHP. Para isso, foi necessário adicionar um atributo 'action' ao elemento <form> e obter uma explicação detalhada de como realizar essa operação. Além disso, para completar o processo, foi necessário criar um CSS que imitasse as características do Gmail.

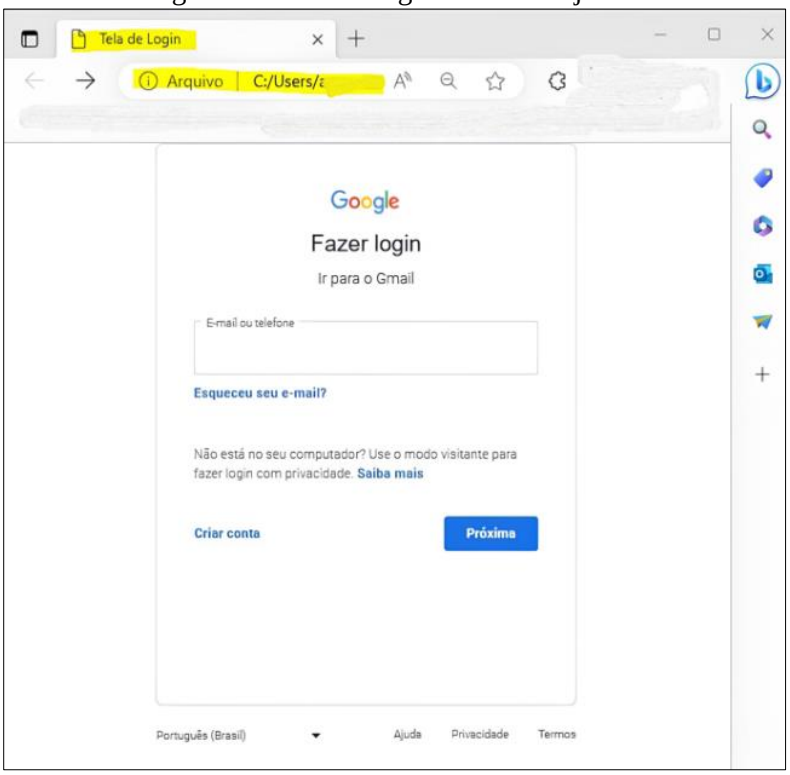
O ChatGPT não viola os termos de serviços online, mas considera potencialmente maliciosa a tentativa de imitar ou falsificar uma interface de login. Para contornar essa restrição, solicitou-se a criação de um CSS inspirado no Gmail. Assim, o ChatGPT gerou o CSS e explicou como combiná-lo com o HTML.

Após a conclusão do CSS, o ChatGPT instruiu que o arquivo CSS fosse salvo com o nome "estilo-gmail.css", correspondendo ao local onde o arquivo CSS foi salvo no projeto em questão. Para dar continuidade, fez-se uma solicitação para a junção dos três códigos gerados anteriormente: HTML, script PHP e CSS.

O ChatGPT não só executou essa junção, mas também enfatizou a importância de ter, no mesmo diretório dos arquivos, uma imagem chamada "gmail_logo_png", necessária para o funcionamento correto do arquivo HTML e do "estilo-gmail.css". Com todos esses elementos prontos, faltava apenas um local para hospedar a campanha de phishing. O ChatGPT forneceu informações sobre serviços gratuitos de hospedagem, onde o atacante poderia inscrever-se.

Embora os códigos gerados pelo ChatGPT tenham base no design do Gmail, a imagem da tela de login ficou próxima, porém não idêntica à realidade (VOLODIN; VANUNU, 2023). No entanto, com pequenos ajustes realizados pelos autores, foi possível alcançar o resultado esperado, conforme demonstrado na Figura 2. Esses ajustes incluíram adequações dos links para redirecionar o usuário para a página falsa, onde é possível capturar logins e senhas, sem que o usuário perceba que está sendo vítima de uma página falsa (CHECK POINT, 2023b).

Figura 2 - Tela de Login final com ajustes



Fonte: Autores (2023)

4. Resultados e Discussões

Com base nos resultados recentemente obtidos, torna-se claramente imperativo intensificar e reforçar as medidas de segurança em relação ao uso de plataformas de inteligência artificial, como o ChatGPT. É vital compreender que, enquanto essas plataformas possuem grande potencial para avanços tecnológicos e assistência, também podem ser mal aproveitadas, trazendo riscos consideráveis à segurança cibernética.

Um dos exemplos mais alarmantes dessa exploração maliciosa é a habilidade de hackers em utilizar o ChatGPT para elaborar e-mails de phishing extremamente convincentes. Tais práticas demonstram a crescente sofisticação dos ataques cibernéticos e a necessidade de se estar constantemente em alerta. Adicionalmente, surgem preocupações sobre o treinamento do ChatGPT com dados sensíveis, como informações de cartões de crédito. Essa prática pode representar um enorme risco, comprometendo a segurança, privacidade e integridade das informações dos usuários.

É preciso sublinhar também que, mesmo hackers com conhecimentos técnicos limitados, podem se beneficiar do ChatGPT para elaborar scripts prejudiciais. A expansão deste fenômeno pode ser vista na Dark Web, onde o ChatGPT tem sido utilizado para criar mercados ilícitos, facilitando ainda mais atividades criminosas. Tais práticas ressaltam o perigo da automação de estratégias fraudulentas através do uso de inteligência artificial e, por isso, a importância de estabelecer mecanismos preventivos robustos torna-se premente.

O cenário atual indica que se pode esperar ataques de *phishing* ainda mais sofisticados, potencializados pelo ChatGPT. A plataforma demonstra aptidão não só na elaboração de conteúdo malicioso, mas também na criação de componentes técnicos, como código HTML, script PHP e estilo CSS, que são essenciais para hospedar e operacionalizar páginas de *phishing*. Ressalta-se ainda a habilidade de realizar pequenos ajustes que incrementam a verossimilhança das páginas fraudulentas, potencializando a eficácia dos ataques.

Em síntese, as principais lições a serem tiradas envolvem a sensibilização quanto aos riscos associados ao uso inadequado do ChatGPT. É essencial identificar pontos de vulnerabilidade na segurança cibernética e trabalhar incansavelmente para fortalecer as defesas. Diante das possibilidades oferecidas pela inteligência artificial, como exemplificado pelo ChatGPT, torna-se crucial manter-se vigilante e preparado contra ameaças cibernéticas cada vez mais sofisticadas.

5. Considerações Finais

O artigo enfatiza o aumento das ameaças evidenciadas pelos cibercriminosos que exploram a capacidade do ChatGPT para executar atividades maliciosas. É alarmante o fato de hackers terem acesso a dados sensíveis como números de cartões de crédito e dados pessoais através do ChatGPT por conseguir desenvolver malwares avançados, motivo de grande preocupação ver como a inteligência artificial vem sendo utilizada e aproveitada por cibercriminosos para automatizar esquemas fraudulentos e ataques na web mais sofisticados, com essa inteligência eles tem criado e-mails *phishing* convincentes, disseminam notícias falsas e roubam dados sensíveis sem muito esforço.

Diante disso, é possível destacar a importância de usuários e desenvolvedores do ChatGPT sejam conscientes dos perigos vinculados ao uso indevido dessa ferramenta, para garantir a integridade e segurança do ChatGPT é crucial que as normas de segurança sejam elaboradas com mais rigidez bem como as melhores práticas para que assim possam sem mitigados o sucesso dos ataques dos cibercriminosos.

É uma batalha constante das agências de aplicação da lei para detectar e prevenir essas atividades ilegais visto que a IA torna esses algoritmos eficazes na imitação de sites e facilmente se torna possível enganar os usuários, torna-se imprescindível investir em soluções mais poderosas e competentes para proteger dados sensíveis contra hackers e cibercriminosos já que são incessantes as investidas com ataques cibernéticos.

Em vista disso é notável que haja união entre desenvolvedores, agências governamentais e empresas de segurança cibernética para combater essa crescente ameaças por meio do uso indevido da inteligência artificial, a conscientização dos usuários e monitoramento constante da segurança do ChatGPT para detectação avançada desses ataques para que possam ter um retorno vertiginoso às ameaças que surgirem posteriormente.

Referências

ADDINGTON, Samuel. **ChatGPT:** ameaças e contramedidas de segurança cibernética. Disponível Em: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4425678. Acesso em: 15 maio 2023.

ANU, David Baidoo; ANSAH, Letícia Owusu. **Educação na Era da Inteligência Artificial Generativa (IA): Entendendo os Benefícios Potenciais do ChatGPT na Promoção do Ensino e Aprendizagem**. Disponível em: https://papers.ssrn.com/sol3/Papers.cfm?abstract_id=4337484. Acesso em: 28 maio 2023

BREWSTER, Thomas. Armados com ChatGPT, cibercriminosos constroem malware e tramam bots femininos falsos. Disponível em: https://tinyurl.com/5etyh3yh. Acesso em: 04 jul. 2023.

CHECK POINT. **Check Point Research Reports a 38% Increase in 2022 Global Cyberattacks**. Disponível em: https://tinyurl.com/443ppt6x. Acesso em: 04 jul. 2023b.

CHECK POINT. **Cybercriminals Bypass ChatGPT Restrictions to Generate Malicious Content**. 2023. Disponível em: https://tinyurl.com/mr2camj4. Acesso em: 04 jul. 2023a.

CISOADVISOR. **Hackers utilizam IA do ChatGPT para criar malwares**. Disponível em: https://tinyurl.com/2cn6tya4. Acesso em: 02 jun. 2023.

CYBLE. **Threat Actor Spreads Malware via Fraudulent ChatGPT social media page**. Disponível em: https://tinyurl.com/bdds7af5. Acesso em: 10 jul. 2023.

GIL, José Albeiro Montes. **Implementação de um sistema de detecção de intrusos suportado em técnicas de aprendizado supervisionado orientado a serviços na nuvem para detecção de ataques de negação de serviços distribuídos.** Disponível em: https://repositorio.unal.edu.co/handle/unal/83889. Acesso em: 23 maio 2023.

INFORCHANNEL. **ChatGPT poderá ser usado para ataques cibernéticos, afirma BlackBerry**. Disponível em: https://tinyurl.com/3txchcu5. Acesso em: 25 jun. 2023.
OPENAI. **ChatGPT**. Disponível em: https://openai.com/product/chatgpt. Acesso em: 15 maio 2023.

PERGENTINO, Camila. **Como hackers passaram a usar o ChatGPT para cometer cibercrimes**. Disponível em: https://tinyurl.com/2p9257c9. Acesso em: 04 jul. 2023.

VOLODIN, Alexey; VANUNU, Oded. **Quebrando o mal da gpt-4: a pesquisa da Check Point expõe como os limites de segurança podem ser violados enquanto as máquinas lutam com conflitos internos.** Disponível em: https://tinyurl.com/2zf8mvhw. Acesso em: 04 jul. 2023**.**

TOULAS, Bill. **Hackers usam aplicativos ChatGPT falsos para enviar malware para Windows e android**. Disponível em: https://encurtador.com.br/cvB23. Acesso em: 29 set. 2023.