Communiqués de presse

IA contre fraude humaine : décoder la nouvelle ère des tactiques d’hameçonnage

Par Stephanie Carruthers, le 26 octobre 2023 : les attaquants semblent innover presque aussi vite que la technologie évolue. Jour après jour, la technologie et les menaces progressent. Aujourd'hui, alors que nous entrons dans l'ère de l'IA, les machines ne se contentent pas d'imiter le comportement humain, elles s'immiscent dans presque tous les aspects de notre vie. Pourtant, malgré l'inquiétude croissante suscitée par les implications de l'IA, l'ampleur de sa potentielle utilisation abusive par les attaquants reste largement méconnue.

Pour mieux comprendre comment les attaquants peuvent tirer parti de l'IA générative, nous avons mené un projet de recherche qui met en lumière une question essentielle : **Les modèles actuels d'IA générative ont-ils les mêmes capacités de tromperie que l'esprit humain ?**

Imaginez un scénario dans lequel l'IA affronte les humains dans une bataille d’hameçonnage. L'objectif ? Déterminer quel concurrent peut obtenir un taux de clics plus élevé lors d'une simulation d’hameçonnage contre des organisations. En tant qu'auteure professionnelle d’e-mails d’hameçonnage, j'étais impatiente de connaître la réponse.

Avec seulement cinq requêtes (« prompts ») simples, nous avons réussi à manipuler un modèle d'IA générative pour qu'il développe des e-mails d’hameçonnage très convaincants **en seulement 5 minutes - le même temps qu'il me faut pour préparer une tasse de café**. Il faut généralement environ *16 heures à mon équipe pour créer un e-mail d’hameçonnage*, sans compter la configuration de l'infrastructure. Les attaquants peuvent donc potentiellement **économiser près de deux jours de travail en utilisant des modèles d'IA générative**. L’hameçonnage généré par l'IA était si convaincant qu'il a presque battu celui élaboré par des ingénieurs sociaux expérimentés. Le fait qu'il soit à un niveau de performance si comparable est déjà une évolution importante.

Dans ce blog, nous expliquons en détail comment les requêtes d’IA ont été créées, comment le test a été mené et ce que cela signifie pour les attaques d'ingénierie sociale d'aujourd'hui et de demain.

Premier Round : l'essor des machines

Dans un coin, nous avions des e-mails d’hameçonnage générés par l'IA avec des récits très astucieux et convaincants.

Création des requêtes. Grâce à un processus systématique d'expérimentation et de perfectionnement, une série de seulement cinq requêtes a été créée pour demander à ChatGPT de générer des e-mails d’hameçonnage adaptés à des secteurs d'activité spécifiques.

Pour commencer, nous avons demandé à ChatGPT de détailler les principaux domaines de préoccupation des employés de ces secteurs. Après avoir donné la priorité au secteur d’activité et aux préoccupations des employés, nous avons demandé à ChatGPT de faire des sélections stratégiques sur l’utilisation des techniques d’ingénierie sociale et de marketing dans l’e-mail. Ces choix visaient à optimiser la probabilité qu’un plus grand nombre d’employés cliquent sur un lien dans l’e-mail lui-même. Ensuite, une requête demandait à ChatGPT qui devait être l’expéditeur (par exemple, une personne interne à l’entreprise, un fournisseur, une organisation externe, etc.). Enfin, nous avons demandé à ChatGPT d'ajouter les compléments suivants pour créer l'e-mail d’hameçonnage :

- **Principaux sujets de préoccupation des employés du secteur de la santé** : évolution de carrière, stabilité de l'emploi, travail épanouissant, etc.
- **Techniques d'ingénierie sociale à utiliser** : confiance, autorité, preuve sociale.
- **Techniques de marketing à utiliser** : personnalisation, optimisation pour le téléphone mobile, appel à l'action.
- **Personne ou entreprise dont il faut usurper l'identité** : responsable des ressources humaines internes.
- **Génération de l'e-mail** : compte tenu de toutes les informations énumérées ci-dessus, ChatGPT a généré l'e-mail expurgé ci-dessous, qui a ensuite été envoyé par mon équipe à plus de 800 employés.

J'ai près d'une décennie d'expérience en ingénierie sociale, j'ai créé des centaines d’e-mails d’hameçonnage et j'ai trouvé les e-mails d’hameçonnage générés par l'IA assez convaincants. Intéressant à noter également : parmi les trois organisations qui avaient initialement accepté de participer à ce projet de recherche, deux se sont complètement retirées après avoir examiné les deux e-mails d’hameçonnage car elles s'attendaient à un taux de réussite élevé. Comme le montrent les requêtes, l'organisation qui a participé à cette étude appartenait au secteur de la santé, qui est actuellement l'un des secteurs les plus ciblés.

Gains de productivité pour les attaquants. Alors qu'il faut généralement environ 16 heures à mon équipe pour créer un e-mail d’hameçonnage, l’e-mail d’hameçonnage généré par l’IA l’a été en seulement **cinq minutes, avec uniquement cinq requêtes simples.**

Deuxième Round : la touche humaine

Dans l'autre coin, nous avions des ingénieurs sociaux chevronnés d’X-Force Red.

Armés de créativité et d'un soupçon de psychologie, ces ingénieurs sociaux ont créé des e-mails d’hameçonnage qui ont trouvé un écho personnel auprès de leurs cibles. L'élément humain a ajouté un air d'authenticité qu'il est souvent difficile de reproduire.

Étape 1 : OSINT. Notre approche de l’hameçonnage commence invariablement par la phase initiale d'acquisition de renseignements de sources ouvertes (OSINT : Open-Source Intelligence). L'OSINT consiste à récupérer des informations accessibles au public, qui font ensuite l'objet d'une analyse rigoureuse et servent de ressource fondamentale dans la formulation des campagnes d'ingénierie sociale. Les principales sources de données pour nos efforts d'OSINT comprennent des plateformes telles que LinkedIn, le blog officiel de l'organisation, Glassdoor et une pléthore d'autres sources.

Au cours de nos activités d’OSINT, nous avons réussi à découvrir un article de blog détaillant le lancement récent d'un programme de bien-être pour les employés, coïncidant avec l'achèvement de plusieurs projets importants. Il était encourageant de constater que ce programme avait fait l'objet de témoignages favorables de la part des employés sur Glassdoor, attestant de son efficacité et de la satisfaction des employés. En outre, nous avons identifié une personne responsable de la gestion du programme via LinkedIn.

Étape 2 : Élaboration de l'e-mail. En utilisant les données recueillies au cours de notre phase d’OSINT, nous avons entamé le processus de construction méticuleuse de notre e-mail d’hameçonnage. Comme étape fondamentale, il était impératif que nous nous fassions passer pour une personne ayant l'autorité nécessaire pour aborder le sujet de manière efficace. Pour renforcer l'aura d'authenticité et de familiarité, nous avons incorporé un lien vers un site web légitime concernant un projet récemment achevé.

Pour renforcer l'impact persuasif, nous avons stratégiquement intégré des éléments d'urgence perçue en introduisant des « contraintes de temps artificielles ». Nous avons indiqué aux destinataires que l'enquête en question ne comportait que « 5 questions brèves », et nous leur avons assuré que sa réalisation ne nécessiterait que « quelques minutes » de leur précieux temps et que la date limite était fixée à « ce vendredi ». Cette formulation délibérée a permis de souligner le peu d'impact sur leur emploi du temps, renforçant ainsi la nature non intrusive de notre approche.

L'utilisation d'une enquête comme prétexte à l’hameçonnage est généralement risquée, car elle est souvent considérée comme un signal d'alarme ou simplement ignorée. Toutefois, au vu des données que nous avons découvertes, nous avons décidé que les avantages potentiels l'emportaient sur les risques associés.

L'e-mail d’hameçonnage expurgé suivant a été envoyé à plus de 800 employés d'une organisation mondiale de

soins de santé :

Le Champion : les humains triomphent, mais de justesse !

Après une série intensive de tests A/B, les résultats ont été clairs : les humains sont sortis vainqueurs, mais avec une marge très faible.

Bien que les e-mails d’hameçonnage conçus par des humains aient réussi à surpasser l'IA, la lutte a été très serrée. Voici pourquoi :

- **L’intelligence émotionnelle** : les humains comprennent les émotions d'une manière encore inaccessible à l'IA. Nous pouvons tisser des récits qui touchent la corde sensible et semblent plus réalistes, ce qui rend les destinataires plus enclins à cliquer sur un lien malveillant. Par exemple, les humains ont choisi un exemple réel au sein de l'organisation, tandis que l'IA a choisi un sujet général, ce qui a rendu l’hameçonnage généré par les humains plus crédible.
- **Personnalisation** : en plus d'incorporer le nom du destinataire dans l'introduction de l’e-mail, nous avons également fourni une référence à une organisation réelle, offrant ainsi des avantages tangibles à son personnel.
- **Ligne d'objet courte et succincte :** l’objet du mail d’hameçonnage généré par un humain était court et précis (« Enquête sur le bien-être des employés ») alors que l’hameçonnage généré par l'IA avait un objet très long (« Changez votre avenir : opportunités d’évolution limitées dans l'entreprise X »), ce qui a pu éveiller les soupçons avant même que les employés n'ouvrent l'e-mail.

Non seulement l’hameçonnage généré par l'IA a perdu face aux humains, mais il a également été **signalé comme suspect à un taux plus élevé.**

Ce qu’il faut retenir : un aperçu de l'avenir

Bien qu’X-Force n'ait pas encore été témoin de l'utilisation à grande échelle de l'IA générative dans les campagnes actuelles, des outils tels que WormGPT, qui ont été conçus pour être des modèles de langage de grande taille (LLM) non restreints ou semi-restreints, ont été mis en vente sur divers forums annonçant des capacités d’hameçonnage. Cela montre que les attaquants testent l'utilisation de l'IA dans les campagnes d’hameçonnage. Bien que les versions restreintes des modèles d'IA générative puissent déjà être utilisées pour faire de l’hameçonnage à l'aide de simples requêtes, ces versions non restreintes pourraient offrir des moyens plus efficaces aux attaquants pour mettre à l'échelle des e-mails d’hameçonnage sophistiqués à l'avenir.

Les humains ont peut-être remporté ce match de justesse, mais l'IA ne cesse de s'améliorer. Au fur et à mesure que la technologie progresse, nous pouvons nous attendre à ce que l'IA devienne de plus en plus sophistiquée et qu'elle puisse même un jour surpasser l'homme. Comme nous le savons, les attaquants s'adaptent et

innovent en permanence. Cette année encore, nous avons vu des escrocs utiliser de plus en plus souvent des clones vocaux générés par l'IA pour inciter les gens à envoyer de l'argent, des cartes-cadeaux ou à divulguer des informations sensibles.

Si les humains ont encore la main lorsqu'il s'agit de manipuler les émotions et de rédiger des e-mails persuasifs, l'émergence de l'IA dans le domaine de l'hameçonnage marque un tournant dans les attaques d'ingénierie sociale. Voici cinq recommandations clés à l'intention des entreprises et des consommateurs pour les aider à rester préparés :

- **En cas de doute, appelez l'expéditeur :** si vous vous demandez si un e-mail est légitime, décrochez le téléphone et vérifiez. Pensez également à choisir un mot de sécurité avec vos amis proches et les membres de votre famille, que vous pourrez utiliser en cas de « vishing » (hameçonnage vocal par téléphone) ou d'escroquerie téléphonique générée par l'IA.
- **Abandonnez le stéréotype de la grammaire** : réfutez le mythe selon lequel les e-mails d'*hameçonnage* sont truffés de fautes de grammaire et d'orthographe. Les tentatives d'hameçonnage pilotées par l'IA sont de plus en plus sophistiquées et font souvent preuve d'exactitude grammaticale. C'est pourquoi il est impératif de rééduquer nos employés et d'insister sur le fait que les fautes de grammaire ne sont plus le principal signal d'alarme. Nous devrions plutôt les former à être vigilants quant à la longueur et à la complexité du contenu des e-mails. Les e-mails plus longs, qui sont souvent caractéristiques d'un texte généré par l'IA, peuvent être un signe d'avertissement.
- **Réorganiser les programmes d'ingénierie sociale :** il s'agit notamment d'intégrer des techniques telles que le vishing dans les programmes de formation. Cette technique est simple à mettre en œuvre et souvent très efficace. Un rapport d'X-Force a révélé que les campagnes d'hameçonnage ciblées qui ajoutent des appels téléphoniques sont trois fois plus efficaces que celles qui ne le font pas.
- **Renforcer les contrôles de gestion des identités et des accès** : les systèmes avancés de gestion des identités et des accès peuvent aider à vérifier qui accède à quelles données, si ces personnes disposent des droits appropriés et si elles sont bien celles qu'elles prétendent être. Un exemple est le MFA (Multi-Factor Authentication) résistant à l'hameçonnage, qui ajoute une deuxième couche de protection en plus des mots de passe, va au-delà des codes d'accès ou des SMS et peut inclure des données biométriques, telles que la lecture des empreintes digitales, etc.
- **S'adapter et innover en permanence** : l'évolution rapide de l'IA signifie que les cybercriminels continueront à affiner leurs tactiques. Nous devons adopter ce même état d'esprit d'adaptation et d'innovation permanentes. Il est essentiel de mettre régulièrement à jour les TTPs* (tactiques, techniques et procédures) internes, les systèmes de détection des menaces et les supports de formation des employés pour garder une longueur d'avance sur les acteurs malveillants.

L'émergence de l'IA dans les attaques d'*hameçonnage* nous pousse à réévaluer nos approches en matière de cybersécurité. En adoptant ces recommandations et en restant vigilants face à l'évolution des menaces, nous pouvons renforcer nos défenses, protéger nos entreprises et assurer la sécurité de nos données et de nos collaborateurs dans l'ère numérique dynamique d'aujourd'hui.

Pour en savoir plus sur les recherches d'X-Force en matière de sécurité, de renseignements sur les menaces et d'informations fournies par les pirates, rendez-vous sur le **X-Force Research Hub**.

Pour en savoir plus sur la façon dont IBM peut aider les entreprises à accélérer leur parcours en matière d'IA en toute sécurité, rendez-vous **ici**.

** Les TTPs (tactiques, techniques et procédures) analysent le fonctionnement d'un acteur malveillant, elles décrivent comment les cyberattaquants orchestrent, exécutent et gèrent les attaques opérationnelles. Les TTPs contextualisent une menace.*

Contacts Presse :

IBM
Gaëlle Dussutour
Tél. : + 33 (0)6 74 98 26 92
dusga@fr.ibm.com

Weber Shandwick pour IBM
Louise Weber
Tél. : + 33 (0)6 89 59 12 54
ibmfrance@webershandwick.com