knowbe4

Sécuriser la main-d'œuvre hybride : protéger les humains et les agents d'IA dans une nouvelle ère

WWW.KNOWBE4.COM/FR

KnowBe4 France | 132 Rue Bossuet, Lyon, France, 69006
+31 (0)30 7996074 | www.KnowBe4.com/fr | Sales@KnowBe4.com

CR-0663 0326

Introduction

Le monde du travail a changé.

Les employés ne travaillent plus seuls. Ils collaborent de plus en plus avec des copilotes d'IA, des assistants et des agents autonomes qui les aident à écrire du code, à résumer des incidents, à rédiger des communications et à analyser des données ainsi qu'à répondre aux questions de leurs clients. Ce qui a commencé comme de simples outils de productivité a rapidement évolué vers une nouvelle catégorie de collègues numériques intégrés aux flux de travail des entreprises.

Cette évolution s'accélère rapidement. Goldman Sachs estime que l'IA agentique pourrait représenter environ 60 % de la valeur du marché des logiciels d'ici 2030, ce qui indique un changement fondamental dans la façon de travailler. La main-d'œuvre moderne n'est plus uniquement humaine, mais une équipe hybride composée de personnes et d'agents d'IA travaillant ensemble pour accroître la productivité et faciliter la prise de décision.

Cette transformation introduit un nouveau défi de sécurité auquel de nombreuses organisations ne sont pas préparées.

Pendant des décennies, les équipes de cybersécurité se sont concentrées sur la protection des systèmes, des réseaux et des terminaux, tout en réduisant les risques humains grâce à la sensibilisation à la sécurité, à la culture de sécurité et à la formation. Ces efforts visaient à se défendre contre l'hameçonnage, l'ingénierie sociale et d'autres formes de manipulation cognitive.

Aujourd'hui, cette surface de risque s'est agrandie. Les humains ne sont plus les seuls acteurs à interagir avec des données sensibles, à générer des contenus ou à prendre des décisions au sein de l'entreprise. Les agents d'IA réalisent désormais nombre de ces tâches, souvent avec rapidité et assurance, mais sans compréhension complète des politiques de l'organisation, du contexte ou de la tolérance au risque.

Contrairement aux employés, les agents d'IA ne peuvent pas être améliorés par des formations traditionnelles ni par des programmes de sensibilisation. Les organisations ne peuvent pas entièrement inspecter ni contrôler les invites qui façonnent les réponses, alors même que ces agents influencent déjà les résultats commerciaux.

Cela crée un défi familier sous une nouvelle forme. Les équipes de sécurité doivent gérer la façon dont les humains interagissent avec la technologie qu'ils utilisent. À l'ère de l'IA, le risque se situe de plus en plus dans la couche d'interaction entre les humains et les systèmes intelligents.

On peut comparer un agent d'IA à un stagiaire très compétent, mais extrêmement influençable. Il traite de vastes quantités d'informations et exécute les instructions rapidement et avec assurance. Il est désireux d'aider, mais il manque de jugement, de contexte et d'une compréhension intrinsèque du risque organisationnel.

Sans les bons garde-fous, même l'assistant le plus performant peut faire des choses inattendues.

La nouvelle surface d'attaque humains-IA

Les agents d'IA sont extrêmement puissants et intrinsèquement vulnérables. Les cybercriminels apprennent déjà à exploiter la dynamique entre les humains et l'IA. Dans de nombreux cas, l'attaque ne vise ni l'humain ni l'IA isolément. Les pirates exploitent plutôt la relation de confiance qui existe entre eux.

L'introduction de l'IA dans les flux de travail quotidiens a créé une nouvelle catégorie de risques de sécurité. Les pirates informatiques utilisent l'IA pour faire évoluer des attaques connues, à savoir générer des e-mails d'hameçonnage plus convaincants, produire des contenus de deepfake pour l'ingénierie sociale et automatiser la reconnaissance. Dans le même temps, ils développent des techniques spécifiquement conçues pour manipuler les systèmes d'IA :

- **Attaques par injections d'invites**
  Entrées malveillantes conçues pour influencer le comportement d'un agent d'IA, pouvant potentiellement l'amener à contourner des contrôles, à divulguer des données sensibles ou à exécuter des actions non intentionnelles.
- **Usurpation d'identité d'un agent d'IA**
  Agents malveillants qui imitent des outils d'entreprise légitimes afin de collecter des identifiants, des informations sensibles ou des données de flux de travail auprès d'employés peu méfiants.
- **Ingénierie sociale humains-IA**
  Attaques qui exploitent la confiance accordée aux résultats générés par l'IA, transformant potentiellement des agents compromis en menaces internes.

Dans de nombreux cas, aucun logiciel malveillant n'est déployé et aucun identifiant n'est volé. L'employé donne simplement une instruction, et l'agent d'IA l'exécute normalement. L'humain comme l'IA se comportent tous deux comme prévu, et c'est précisément là que réside la vulnérabilité.

Le problème de la confiance

Les humains font naturellement confiance aux outils qui les aident à réussir. Lorsque les systèmes d'IA produisent des résumés clairs, des recommandations utiles ou des contenus soignés, les utilisateurs accordent rapidement leur confiance dans leurs résultats. Avec le temps, cela peut engendrer un biais d'automatisation, à savoir la tendance à faire confiance à des décisions automatisées, même lorsqu'elles peuvent être erronées ou manipulées.

Les cybercriminels peuvent exploiter cette confiance. Une invite malveillante dissimulée dans un document, un jeu de données ou une page Web peut influencer une réponse générée par l'IA et diffuser des informations manipulées sur un ton confiant et assuré. Pour l'employé, le résultat semble crédible, même s'il a été façonné par un attaquant.

Les architectures de sécurité traditionnelles ont été conçues pour protéger les systèmes, les réseaux et les terminaux. Bien que ces contrôles restent essentiels, ils ne couvrent pas entièrement les risques créés par la collaboration entre l'humain et l'IA. La principale lacune se situe désormais au niveau de la couche d'interaction, où les employés s'appuient sur l'IA pour interpréter des informations, générer des contenus et soutenir la prise de décision.

Sécuriser cet environnement nécessite de protéger les deux parties de la main-d'œuvre moderne. Les organisations doivent donner aux employés les moyens de reconnaître les tentatives de manipulation et d'utiliser les outils d'IA en toute sécurité, tout en veillant à ce que les agents d'IA soient résilients face aux injections d'invites, aux expositions de données et aux actions non autorisées.

L'avenir de la sécurité réside dans un modèle de défense hybride

En cybersécurité, la frontière entre l'humain et l'IA disparaît. À mesure que les agents d'IA s'intègrent dans les flux de travail quotidiens, seules les organisations capables d'adapter leurs stratégies de sécurité à cette réalité sauront faire preuve de résilience. Le constat est clair : la cybersécurité ne consiste plus seulement à protéger les systèmes contre les humains ou les humains contre les systèmes. Il s'agit de garantir l'interaction entre eux, car c'est au cœur de cette interaction que résident à la fois la plus grande vulnérabilité et la meilleure défense.

Pour relever ce défi, les organisations doivent adopter une stratégie de défense double qui renforce à la fois l'élément humain et les systèmes d'IA eux-mêmes.

1. Renforcer la résilience humaine

Dans un environnement de travail augmenté par l'IA, le rôle de l'employé évolue, passant de l'exécution de tâches à la supervision et à la prise de décision. La formation sur la sensibilisation à la sécurité doit évoluer en conséquence. Il ne suffit plus de former les utilisateurs à reconnaître un e-mail d'hameçonnage. Les employés doivent développer une vigilance numérique, à savoir une combinaison de scepticisme sain, de conscience contextuelle et de capacité à évaluer de manière critique les contenus générés par l'IA.

Ceci commence par l'éducation. Les employés doivent connaître à la fois les capacités et les limites des agents d'IA, et notamment la manière dont ils peuvent être manipulés. Ils doivent être formés à reconnaître les comportements anormaux ou inattendus de l'IA, tels que des résultats qui entrent en conflit avec les politiques ou le contexte. De même, les organisations doivent mettre en place des protocoles de vérification simples mais efficaces pour les actions à risque élevé, en particulier celles initiées ou assistées par l'IA. Lorsqu'une situation paraît urgente, inhabituelle ou hors du cadre normal, les employés doivent savoir comment faire une pause, vérifier et faire remonter l'information.

2. Sécuriser l'agent d'IA

En parallèle, les systèmes d'IA doivent être considérés comme une nouvelle classe de ressources de l'entreprise nécessitant gouvernance, surveillance et contrôle. Cela commence par la mise en place d'une validation rigoureuse des entrées afin d'empêcher que des invites malveillantes ou des données empoisonnées n'influencent le comportement des agents. Les équipes doivent également analyser les sorties de l'IA en continu afin de détecter des anomalies, des violations de politiques ou des signes de manipulation.

Il est tout aussi important de définir clairement les limites des rôles des agents d'IA. Les systèmes doivent être conçus pour fonctionner dans des cadres stricts et refuser les requêtes qui sortent de ces cadres. Ceci doit être renforcé par des politiques d'utilisation de l'IA clairement définies, qui encadrent la manière dont les agents interagissent avec les données, les systèmes et les utilisateurs.

Capacités critiques

La gestion des risques dans cet environnement hybride nécessite une visibilité sur la manière dont les agents d'IA fonctionnent, interagissent avec les utilisateurs et accèdent aux données et aux systèmes. Pour assurer une couverture complète, de la découverte à la défense, les solutions efficaces de sécurité des agents d'IA doivent s'articuler autour de quatre piliers fondamentaux :

Découvrir
Identifier et cataloguer automatiquement chaque agent d'IA fonctionnant dans votre environnement, ce qui inclut aussi bien les agents introduits par des plateformes approuvées que ceux déployés sans validation officielle. La découverte continue permet d'éliminer les angles morts et garantit que les équipes de sécurité connaissent le cadre complet de l'utilisation de l'IA.

Surveiller
Enregistrer chaque invocation d'agent, interaction d'invite et appel d'outil dans une piste d'audit complète et consultable. Une télémétrie riche, incluant des métadonnées provenant d'assistants numériques d'IA tels que Microsoft Copilot, permet aux équipes de comprendre comment les agents sont utilisés et où des risques potentiels peuvent apparaître.

Détecter
Appliquer des capacités de détection avancées pour identifier les comportements à risque en temps réel. Cela inclut l'identification de tentatives d'injections d'invites, d'élévation de privilèges, d'expositions de données sensibles et d'activités anormales des agents, quels que soient les langues et les cas d'utilisation. L'analyse comportementale permet de mettre en évidence à la fois les menaces connues et émergentes que les contrôles traditionnels peuvent ne pas détecter.

Protéger
Permettre aux organisations de définir et d'appliquer la stratégie de sécurité de leur choix. Les équipes peuvent choisir de surveiller l'activité de manière passive, de générer des alertes, ou d'intercepter et d'empêcher activement les opérations à risque élevé en temps réel. Une mise en application flexible permet aux équipes de sécurité de trouver un équilibre entre productivité et protection à mesure que l'adoption de l'IA évolue.

Enfin, pour les professionnels de l'informatique et de la sécurité, l'évolutivité est essentielle. Gérer les risques à travers une population croissante d'utilisateurs humains et d'agents d'IA nécessite une visibilité centralisée et une grande efficacité opérationnelle. Les capacités clés doivent inclure :

Visibilité et audit
Des tableaux de bord centralisés offrent une vue claire des interactions entre humains et agents, ainsi qu'entre agents et systèmes. Des journaux d'audit détaillés permettent aux équipes de sécurité de détecter les comportements anormaux, d'enquêter sur les incidents et d'identifier les activités qui s'écartent des politiques définies ou des cas d'utilisation prévus.

Application dynamique des politiques
Les équipes de sécurité peuvent définir des politiques fondées sur des risques par le biais de contrôles pilotés par API, ce qui permet une gouvernance cohérente de l'utilisation de l'IA dans différents environnements. Les politiques peuvent être ajustées à mesure que de nouveaux cas d'utilisation émergent, afin de garantir que l'adoption de l'IA reste alignée sur les exigences de sécurité et de conformité de l'organisation.

Score de risque intégré
Les données d'interaction alimentent le cadre global de gestion du risque humain, ce qui permet aux organisations d'évaluer les risques liés à la fois aux comportements humains et à ceux des agents d'IA. En intégrant les interactions entre humains et agents dans les modèles de score de risque, les responsables de la sécurité obtiennent une meilleure compréhension de l'exposition et peuvent prioriser plus efficacement les efforts d'atténuation.

Conclusion

Les humains et les agents d'IA travaillent désormais côte à côte, les organisations ont besoin d'une approche unifiée pour gérer les risques qui en découlent. En élargissant les principes existants de gestion du risque humain aux agents d'IA, les équipes de sécurité peuvent favoriser l'innovation tout en conservant le contrôle, afin de réduire l'exposition sans ralentir les progrès.

En savoir plus

Test de sécurité gratuit relatif à l'hameçonnage

Découvrez le pourcentage de Phish-Prone (pourcentage de vulnérabilité à l'hameçonnage) de vos employés, en profitant de votre test de sécurité gratuit relatif à l'hameçonnage.

Programme automatisé de sensibilisation à la sécurité gratuit

Créez un programme de sensibilisation à la sécurité personnalisé pour votre organisation.

Outil Phish Alert Button gratuit

Un seul clic suffit désormais à vos employés pour signaler les attaques par hameçonnage de manière sécurisée.

Outil Email Exposure Check (EEC) gratuit

Identifiez avant les pirates les adresses e-mail à risque de vos utilisateurs.

Outil Domain Spoof Test gratuit

Déterminez si les pirates peuvent usurper une adresse e-mail de votre domaine.

À propos de KnowBe4

KnowBe4 offre aux équipes modernes les moyens de prendre au quotidien des décisions plus éclairées en matière de sécurité. Fort de la confiance de plus de 70 000 organisations dans le monde, KnowBe4 est le pionnier de la sécurité de la main-d'œuvre numérique, protégeant à la fois les agents d'IA et les humains. La Plateforme KnowBe4 propose des simulations d'attaques et des formations, la sécurité de la collaboration, ainsi que la sécurité des agents, grâce à AIDA (Artificial Intelligence Defense Agents) et à un score de risque propriétaire. Elle s'appuie sur 15 années de données comportementales pour lutter contre les menaces avancées, notamment l'ingénierie sociale, l'injection d'invites et l'IA fantôme. Sécurisant à la fois les humains et les agents, KnowBe4 est un leader du secteur en matière de confiance et de protection des équipes.

Obtenez plus d'informations à l'adresse knowbe4.com/fr. Suivez KnowBe4 sur LinkedIn et X.

Follow us on Linkedin

Follow us on X

knowbe4

KnowBe4 France | 132 Rue Bossuet, Lyon, France, 69006
+31 (0)30 7996074 | www.KnowBe4.com/fr | Sales@KnowBe4.com