knowbe4

Schutz der hybriden Belegschaft aus Menschen und KI-Agenten in einem neuen Zeitalter

WWW.KNOWBE4.DE

KnowBe4 Germany | Rheinstr. 45/46 | 12161 Berlin – Deutschland
KnowBe4.de | +49 30 34 64 64 60 | kontakt@KnowBe4.com

CR-0663 0326

Einführung

Die Arbeitswelt hat sich verändert.

Ihre Mitarbeitenden arbeiten nicht mehr allein, sondern nehmen zum Programmieren, Zusammenfassen, Kommunizieren und Analysieren immer häufiger die Hilfe von KI-Assistenten und autonom agierenden Agenten in Anspruch. Die anfänglich zur Förderung der Produktivität eingesetzten KI-Tools bilden heute eine digitale Belegschaft, die in Unternehmensabläufe integriert ist.

Dieser Wandel schreitet rapide voran. Laut Prognose von Goldman Sachs machen KI-Agenten bis 2030 voraussichtlich einen Anteil von 60 % am Wert des Softwaremarkts aus. Dies signalisiert einen drastischen Wandel in der Art, wie Arbeit erledigt wird. Die moderne Belegschaft umfasst nicht mehr nur Menschen. Sie haben ein hybrides Team aus Menschen und KI-Agenten, die zusammenarbeiten und Entscheidungen treffen.

Dieser Wandel führt zu neuen Herausforderungen, auf die viele Organisationen nicht vorbereitet sind.

Jahrzehntelang haben sich Cybersicherheitsteams auf den Schutz von Systemen, Netzwerken und Endpunkten konzentriert und versucht, menschliche Fehler durch Security Awareness Training und den Aufbau einer Sicherheitskultur zu reduzieren. Dieser Ansatz bot Schutz vor Phishing, Social Engineering und anderen Formen der kognitiven Manipulation.

Die Angriffsfläche ist jedoch größer geworden. Die Verarbeitung sensibler Daten, Content-Generierung oder Entscheidungsfindung wird nicht mehr allein von Menschen übernommen. Immer häufiger erledigen KI-Agenten solche Aufgaben – und zwar schnell und „selbstbewusst“, jedoch ohne Risikotoleranz oder Kenntnis von Organisationsrichtlinien und Kontext.

KI-Agenten lassen sich im Gegensatz zu Ihren Mitarbeitenden nicht durch herkömmliches Training oder Security-Awareness-Programme verbessern. Organisationen können nicht alle Prompts prüfen oder kontrollieren. Trotzdem fließen die Ausgaben von KI-Agenten bereits in die Arbeit ein und beeinflussen somit die Geschäftsergebnisse.

IT-Sicherheitsteams müssen daher menschliche Interaktionen mit der eingesetzten Technologie regeln. Im Zeitalter der KI verschiebt sich das Risiko zunehmend auf die Schnittstelle zwischen Menschen und intelligenten Systemen.

Stellen Sie sich einen KI-Agenten wie einen durchaus fähigen, jedoch äußerst beeinflussbaren Praktikanten vor. Er kann große Mengen an Informationen aufnehmen und Anweisungen schnell und zuverlässig ausführen. Er ist sehr hilfsbereit, ihm fehlt es jedoch an Urteilsvermögen, Kontext und Verständnis für die spezifischen Risiken Ihrer Organisation.

Ohne entsprechende Vorgaben kann sich selbst der fähigste Assistent unerwartet verhalten.

Die wachsende Mensch-KI-Angriffsfläche

KI-Agenten sind unglaublich leistungsstark und von Natur aus verwundbar. Daher ist es nicht verwunderlich, dass Angreifende es auf die Mensch-KI-Schnittstelle abgesehen haben. Es wird meist nicht ein Ziel (Mensch oder KI) angegriffen, sondern das bestehende Vertrauensverhältnis ausgenutzt.

Die Integration von KI-Tools in alltägliche Arbeitsabläufe hat neue Sicherheitsrisiken hervorgebracht. Bedrohungsakteurinnen und Bedrohungsakteure generieren mithilfe von KI-Tools überzeugendere Phishing-E-Mails in großem Maßstab sowie Deepfake-Inhalte für Social-Engineering-Angriffe und automatisieren ihre Recherche. Gleichzeitig werden Techniken zur Manipulation von KI-Systemen entwickelt:

- **Prompt-Injection-Angriffe**
  Schädliche Eingaben beeinflussen das Verhalten von KI-Tools, z. B. um Kontrollen zu umgehen, sensible Daten auszugeben oder unzulässige Aktionen auszuführen.
- **KI-Identitätsdiebstahl**
  Gefährliche KI-Tools, die legitime Unternehmenstools nachahmen, werden erstellt, um Anmeldedaten, sensible Daten oder Workflow-Daten von arglosen Mitarbeitenden zu entwenden.
- **Social Engineering an der Mensch-KI-Schnittstelle**
  Bei Angriffen wird das Vertrauen in KI-generierte Ausgaben ausgenutzt. Kompromittierte KI-Agenten können zur Insider-Bedrohung werden.

In den meisten Fällen erfolgt weder eine Malware-Infektion noch der Diebstahl von Anmeldedaten. Die Mitarbeitenden übermitteln einfach eine Anweisung, die der KI-Agent wie vorgesehen ausführt. Mensch und KI verhalten sich erwartungsgemäß. Doch genau hier liegt die Schwachstelle.

Die Vertrauensfrage

Mitarbeitende vertrauen Tools, die ihnen weiterhelfen. Ob Zusammenfassungen, Empfehlungen oder Verbesserungen – Nutzerinnen und Nutzer schenken den Ausgaben von KI-Tools schnell Vertrauen. Es entsteht ein Automation Bias: die Neigung, automatisierten Entscheidungen zu vertrauen, selbst wenn diese möglicherweise falsch sind oder manipuliert wurden.

Angreifende versuchen, dieses Vertrauen auszunutzen. Durch einen in einem Dokument, in einem Datensatz oder auf einer Webseite versteckten schädlichen Prompt können KI-generierte Ausgaben unbemerkt manipuliert werden. Die KI-Ausgabe selbst scheint überzeugend und zuverlässig. Den Mitarbeitenden fällt nicht auf, dass die scheinbar unbedenkliche KI-Ausgabe böswillig manipuliert wurde.

Herkömmliche Sicherheitsarchitekturen schützen Systeme, Netzwerke und Endpunkte. Solche Kontrollmechanismen sind weiterhin wichtig, reichen für die Risiken an der Mensch-KI-Schnittstelle jedoch nicht mehr aus. Die größte Schwachstelle bildet die Interaktionsebene – die Mitarbeitenden vertrauen bei der Interpretation von Informationen, der Content-Generierung und der Entscheidungsfindung auf KI-Tools.

Beim Schutz dieser Umgebung müssen beide Seiten der modernen Belegschaft berücksichtigt werden. Organisationen müssen dafür sorgen, dass die Mitarbeitenden Manipulationsversuche erkennen und KI-Tools sicher einsetzen. Gleichzeitig muss dafür gesorgt werden, dass KI-Tools resilient gegen Prompt-Injection, Datenexposition und unbefugte Aktionen sind.

Doppelte Abwehr von Sicherheitsbedrohungen

In der Cybersicherheit verschwinden die Grenzen zwischen Menschen und KI-Tools. KI-Tools werden zunehmend in tägliche Arbeitsabläufe integriert. Organisationen müssen ihre Sicherheitsstrategien an diese neue Realität anpassen. Bei Cybersicherheit geht es nicht mehr nur um den Schutz von Systemen vor Menschen oder den Schutz von Menschen vor Systemen. Es geht um sichere Mensch-KI-Zusammenarbeit. Denn dort liegen die größten Schwachstellen, aber auch große Chancen.

Zur Stärkung sowohl der Mitarbeitenden wie auch der KI-Systeme müssen Organisationen eine doppelte Verteidigungsstrategie verfolgen.

1. Stärkung der Mitarbeitenden

In einem KI-gestützten Arbeitsumfeld wandelt sich die Rolle der Mitarbeitenden von der reinen Aufgabenausführung hin zu Kontrolle und Einschätzung. Ihr Security Awareness Training muss sich entsprechend weiterentwickeln. Es reicht nicht mehr aus, den Nutzerinnen und Nutzern beizubringen, wie sie eine Phishing-E-Mail erkennen. Die Mitarbeitenden müssen digitale Achtsamkeit lernen – eine Kombination aus gesunder Skepsis, Kontextbewusstsein und der Kompetenz, KI-generierte Ausgaben kritisch zu hinterfragen.

Das beginnt mit dem richtigen Training. Die Mitarbeitenden müssen sowohl die Fähigkeiten wie auch die Grenzen von KI-Agenten verstehen und wissen, wie diese manipuliert werden können. Sie müssen lernen, ungewöhnliches oder unerwartetes KI-Verhalten zu erkennen, beispielsweise nicht konforme oder kontextuell falsche Ausgaben. Organisationen benötigen einfache und effektive Verifizierungsverfahren, insbesondere für KI-gestützte Aktionen mit hohem Risiko. Ist etwas dringlich, ungewöhnlich oder liegt außerhalb des Kompetenzbereichs, müssen die Mitarbeitenden innehalten, prüfen und ggf. eskalieren.

2. Stärkung der KI-Agenten

Gleichzeitig erfordern KI-Systeme als neue Unternehmenstools Governance, Überwachung und Kontrolle. Damit das Verhalten von KI-Agenten nicht durch schädliche Prompts oder Datenmanipulation beeinflusst werden kann, ist eine starke Eingabevalidierung erforderlich. KI-Ausgaben sollten regelmäßig analysiert werden, um Anomalien, Richtlinienverstöße oder Anzeichen von Manipulation aufzudecken.

Darüber hinaus müssen klare Grenzen für KI-Agenten definiert werden. Systeme dürfen nur in einem festgelegten Bereich agieren und müssen nicht autorisierte Aufgaben ablehnen. Zusätzlich ist ratsam, über eine KI-Nutzungsrichtlinie zu regeln, wie KI-Agenten mit Daten, Systemen sowie Nutzerinnen und Nutzern interagieren.

Kritische Kompetenzen

Der Umgang mit Risiken in einer hybriden Umgebung erfordert Transparenz bezüglich der Funktionsweise von KI-Agenten, deren Interaktion mit Nutzerinnen und Nutzern sowie deren Zugriff auf Daten und Systeme. Die komplette Abdeckung von der Entdeckung bis zur Abwehr in Bezug auf die Sicherheit von KI-Agenten umfasst vier wichtige Säulen:

Ermitteln

Automatische Identifizierung und Katalogisierung aller in Ihrer Umgebung eingesetzten KI-Agenten – dazu gehören KI-Agenten von zulässigen Plattformen und KI-Agenten, die ohne Genehmigung eingesetzt werden. Durch eine regelmäßige Ermittlung der eingesetzten KI-Agenten lassen sich Schwachstellen beseitigen. Außerdem wird sichergestellt, dass IT-Sicherheitsteams die gesamte KI-Nutzung im Blick haben.

Überwachen

Erfassen Sie alle Abfragen durch KI-Tools, Prompt-Interaktionen und KI-Tool-Aufrufe von Mitarbeitenden in einem ausführlichen, durchsuchbaren Audit-Protokoll. Anhand von umfassenden Telemetriedaten, einschließlich Metadaten von KI-Assistenten wie Microsoft Copilot, können Teams nachvollziehen, wie KI-Agenten genutzt werden und wo potenzielle Risiken liegen.

Erkennen

Identifizieren Sie riskantes Verhalten in Echtzeit mithilfe fortschrittlicher Erkennungsfunktionen. Dazu gehören Erkennung von Prompt-Injection-Angriffen, Eskalation von Privilegien, Offenlegung sensibler Daten sowie ungewöhnliche KI-Agenten-Aktivitäten über verschiedene Sprachen und Anwendungen hinweg. Mithilfe einer Verhaltensanalyse lassen sich bekannte und neue Bedrohungen aufdecken, die herkömmliche Kontrollmechanismen möglicherweise übersehen.

Schützen

Unternehmen müssen imstande sein, die gewünschte Sicherheitslage zu definieren und durchzusetzen. Teams können Aktivitäten passiv überwachen, Warnmeldungen generieren oder Vorgänge mit hohem Risiko aktiv in Echtzeit abfangen und verhindern. Dank flexibler Durchsetzungsoptionen können IT-Sicherheitsteams im Zuge der vermehrten Nutzung von KI-Tools ein Gleichgewicht zwischen Produktivität und Schutz herstellen.

Auch Skalierbarkeit ist wichtig. Risikomanagement erfordert angesichts der steigenden Zahl menschlicher Nutzerinnen und Nutzer sowie KI-Agenten klare Transparenz und operative Effizienz. Wichtige Funktionen:

Transparenz und Auditing

Zentrale Dashboards bieten einen klaren Überblick über Mensch-KI-Interaktionen und KI-System-Interaktionen. Detaillierte Audit-Protokolle ermöglichen es IT-Sicherheitsteams, auffälliges Verhalten zu erkennen, Vorfälle zu analysieren und Aktivitäten zu identifizieren, die außerhalb der festgelegten Richtlinien oder der vorgesehenen Anwendung liegen.

Dynamische Richtliniendurchsetzung

IT-Sicherheitsteams können mithilfe von APIs risikobasierte Richtlinien festlegen und die KI-Nutzung in allen Umgebungen einheitlich steuern. Richtlinien können auch auf neue Anwendungsfälle angepasst werden, um sicherzustellen, dass neue KI-Tools stets gemäß den Sicherheits- und Compliance-Anforderungen der Organisation agieren.

Integrierte Risikobewertung

Interaktionsdaten fließen in das übergeordnete Framework für Human Risk Management ein, sodass Organisationen die Risiken von menschlichem Verhalten und Verhalten der KI-Tools bewerten können. Durch die Beachtung von Mensch-KI-Interaktionen in Risikobewertungsmodellen erhalten Führungskräfte in der IT-Sicherheit ein umfassenderes Verständnis von den Risiken und können effektiv Maßnahmen zur Risikominderung ergreifen.

Fazit

Menschen und KI-Agenten arbeiten heute als Team. Organisationen benötigen einen umfassenden Ansatz, um den Risiken zu begegnen, die sich aus dieser neuen Arbeitssituation ergeben. Wenn Sie Ihr Human Risk Management auf KI-Agenten ausweiten, können Sie Innovationen fördern sowie Risiken minimieren, ohne den Fortschritt zu bremsen. Gleichzeitig behalten Sie die Kontrolle.

Kostenloser Phishing Security Test

Wie anfällig sind Ihre Mitarbeitenden gegenüber Phishing? Unser kostenloser Phishing Security Test verrät es Ihnen.

Kostenloses Automated Security Awareness Program

Erstellen Sie ein auf Ihre Organisation abgestimmtes Security Awareness Program.

Kostenloser Phish Alert Button

Mit diesem Tool können Mitarbeitende ab sofort Phishing-Angriffe mit nur einem Klick sicher melden.

Kostenloser Email Exposure Check

Welche Ihrer E-Mail-Anmeldedaten wurden bereits offengelegt? Werden Sie aktiv, bevor es Kriminelle tun.

Kostenloser Domain Spoof Test

Finden Sie heraus, ob Hackerinnen und Hacker die E-Mail-Adressen Ihrer Domain spoofen können.

Über KnowBe4

KnowBe4 versetzt die moderne Belegschaft in die Lage, jeden Tag intelligentere Sicherheitsentscheidungen zu treffen. Mehr als 70.000 Organisationen setzen bereits auf KnowBe4 – den Pionier für die Sicherheit für die digitale Belegschaft und den Schutz von KI-Agenten und Menschen. Die KnowBe4-Plattform bietet Angriffssimulationen und Training, Sicherheit bei der Zusammenarbeit und Sicherheit auf Ebene der KI-Agenten durch AIDA (Artificial Intelligence Defense Agents) und einen proprietären Risk Score. Mithilfe von Verhaltensdaten aus über 15 Jahren bekämpft die Plattform hochentwickelte Bedrohungen wie Social Engineering, Prompt-Injection und Schatten-KI. Durch den Schutz von Menschen und KI-Agenten nimmt KnowBe4 eine führende Rolle in der Branche bei Vertrauen und Schutz der Belegschaft ein.

Weitere Informationen auf knowbe4.de. KnowBe4 auf LinkedIn und X folgen.

Follow us on Linkedin

Follow us on X

knowbe4

KnowBe4 Germany | Rheinstr. 45/46, 12161 Berlin – Deutschland
+49 30 34 64 64 60 | KnowBe4.de | kontakt@knowbe4.com