

Modelos de IA a invadirem empresas vão ser mais comuns no futuro, sublinha CEO da Visionware

 jornaleconomico.sapo.pt/noticias/modelos-de-ia-a-invadirem-empresas-vao-ser-mais-comuns-no-futuro-sublinha-ceo-da-visionware

Depois da OpenAI e da Anthropic terem revelado que os seus modelos de IA invadiram sistemas informáticos de empresas, sem que estas se apercebessem, esta ferramenta esteve, novamente, no centro do debate. Bruno Castro avisa: estas situações irão tornar-se cada vez mais comuns.

Em agosto, vieram a público notícias sobre algo impensável ainda há pouco tempo: modelos de inteligência artificial (IA), da OpenAI e da Anthropic, invadiram organizações durante a realização de testes de cibersegurança. Estes desenvolvimentos geraram sérias preocupações na sociedade sobre os impactos que esta tecnologia tem nas empresas.

Bruno Castro, fundador & CEO da VisionWare, explicou ao Jornal Económico que é necessário as empresas apostarem em cibersegurança, e não darem um maior acesso aos dados do que o necessário. Outro fator importante para as empresas perceberem se sofreram algum tipo de ataque é ter capacidade própria para perceber comportamentos anómalos nos seus sistemas.

Como devemos classificar estas situações? Falhas de segurança?

Eu diria que são, antes de mais, falhas de segurança nos ambientes de teste, agravadas pelo facto de estarmos perante modelos com capacidades que já conseguem explorar essas falhas de forma autónoma.

É importante separar duas coisas. O modelo não foi necessariamente “mandado atacar” uma empresa, mas também não o fez por “vontade própria”. Em vários destes casos, os testes foram concebidos para avaliar as capacidades ofensivas dos modelos, com algumas das salvaguardas habituais reduzidas. O problema surge quando o modelo encontra uma forma de ultrapassar os limites que lhe foram impostos e continua a executar a tarefa fora do âmbito previsto.

No caso da OpenAI e da Hugging Face, por exemplo, os modelos conseguiram explorar uma vulnerabilidade zero-day num componente utilizado no ambiente de avaliação para obter acesso à Internet e, posteriormente, alcançar sistemas da Hugging Face.

Portanto, não classificaria isto simplesmente como “uma IA que decidiu atacar”. O que aconteceu foi que um sistema recebeu um objetivo, encontrou um caminho não previsto para o cumprir e as barreiras à sua volta não foram suficientes para o impedir.

Como é que podemos perceber que nos aconteceu a nós, uma vez que as “donas” dos modelos demoraram a perceber o que se passou?

Esta é talvez uma das perguntas mais importantes. A resposta mais imediata é que temos de assumir que não podemos depender exclusivamente do fornecedor do modelo para nos dizer que alguma coisa aconteceu.

Uma empresa deve ter capacidade própria para perceber comportamentos anómalos nos seus sistemas: acessos fora do padrão, utilização inesperada de contas, alterações de permissões, criação de recursos, transferência de informação ou atividade proveniente de sistemas que normalmente não comunicam entre si – um SOC (Security Operations Center) ajuda bastante neste processo.

E há uma mudança de mentalidade importante. Até agora, muitas organizações procuram indicadores de ataque associados a humanos ou a malware. Com agentes de IA, teremos de aprender a reconhecer também comportamentos automatizados que parecem legítimos individualmente, mas que, vistos em conjunto, contam uma história diferente.

É precisamente esse um dos problemas dos agentes autónomos: cada ação pode parecer razoável, mas a sequência completa pode não ser.

Qual o risco para as empresas “atacadas”? E para as empresas que fizeram o modelo?

Para a empresa atingida, o risco é semelhante ao de qualquer outro compromisso de sistemas, isto é, exposição de informação, roubo de credenciais, adulteração de dados, acesso a sistemas internos ou utilização da infraestrutura como ponto de partida para outros ataques.

Para quem desenvolve o modelo, o problema é diferente. Existe um risco reputacional, jurídico e de responsabilidade. Se um modelo demonstrar capacidade para ultrapassar as fronteiras que lhe foram impostas, torna-se legítimo perguntar se o fornecedor avaliou suficientemente esse comportamento antes de lhe dar acesso a determinadas ferramentas ou ambientes.

E há uma questão ainda mais incómoda: quem é responsável quando o modelo executa uma cadeia de ações que ninguém lhe pediu explicitamente para executar?

Como nos podemos preparar para uma situação destas?

A primeira regra é simples, e começa por nós (organização) em não dar a um agente autónomo mais acesso do que aquele de que realmente precisa.

Se uma IA apenas precisa de consultar uma base de dados, não deve poder modificar essa base. Se precisa de executar código, isso deve acontecer num ambiente isolado. Se

precisa de aceder à Internet, esse acesso deve ser limitado e monitorizado.

Adicionalmente, importa limitar o impacto caso um agente consiga explorar uma vulnerabilidade. Tal significa manter os ambientes de teste rigorosamente separados dos sistemas de produção, reduzir ao mínimo as permissões das contas utilizadas, e garantir que uma credencial comprometida não permite saltar para outros sistemas. É também fundamental monitorizar comportamentos anómalos e deter mecanismos para bloquear, rapidamente, os devidos acessos, revogar credenciais e isolar sistemas.

Se não conseguimos necessariamente impedir que um modelo tente explorar uma vulnerabilidade que encontrou, podemos, pelo menos, garantir que essa vulnerabilidade não lhe dê a chave de toda a organização.

Como é que olha para estas situações?

Naturalmente, olho para estas situações com preocupação, mas sem alarmismo. O mais relevante não é atribuir intenção ao modelo, mas antes reconhecer que pode encontrar formas de cumprir um objetivo que os seus criadores não anteciparam.

Para a cibersegurança, esta é uma mudança importante. Obriga-nos a pensar não apenas em proteger os sistemas contra alguns ataques conhecidos, mas também contra comportamentos inesperados de sistemas que, até há pouco tempo, considerávamos ferramentas.

Quando damos mais autonomia a um modelo, temos de estar preparados para comportamentos que não conseguimos prever à partida.

O desafio não será impedir a IA de fazer apenas aquilo que sabemos que pode fazer. É garantir que, quando faz algo inesperado, existem barreiras suficientes para impedir que isso se transforme num incidente sério e com proporções indesejadas.

Acha que esta vai ser uma situação comum?

Acredito que vamos assistir a mais situações deste género, sobretudo à medida que os modelos ganham maior autonomia e acesso a ferramentas e sistemas externos.

Tal não significa que todos os modelos vão comportar-se desta forma, nem que estamos perante uma inevitabilidade. Significa sim, que comportamentos imprevistos vão tornar-se um risco que as organizações terão de considerar nos seus modelos de segurança.

Quanto mais capacidade dermos a estes sistemas para agir por conta própria, maior será a necessidade de testar não apenas aquilo que esperamos que façam, como também aquilo que podem fazer quando encontram um caminho que ninguém antecipou.

Qual o impacto que estas situações podem ter na confiança da IA?

Podem ter um impacto negativo na confiança, sobretudo, se forem apresentadas como casos em que a IA “escapou ao controlo”. Mas também podem ter o efeito contrário.

Estes incidentes permitem-nos descobrir comportamentos que não conhecíamos e perceber melhor onde estão os limites dos modelos e que mecanismos precisamos de desenvolver para os tornar mais previsíveis e controláveis. É preferível descobrir estas situações num ambiente de testes do que quando um agente já estiver integrado num sistema crítico.

Por estes motivos, não vejo estes casos somente como falhas. São também uma oportunidade para aprender como limitar melhor a autonomia dos agentes, definir fronteiras mais eficazes e construir sistemas capazes de detetar e travar comportamentos inesperados.