اختراق أنظمة حقيقية.. نماذج ذكاء اصطناعي تتجاوز بيئة المحاكاة وتفجر مخاوف

كشفت شركة "أنثروبيك" المتخصصة في تطوير تقنيات الذكاء الاصطناعي، أن نماذجها التابعة لـ"كلود" تمكنت من اختراق أنظمة ثلاث مؤسسات خارجية خلال اختبارات أمنية، في حادثة أثارت مخاوف جديدة بشأن مخاطر تطور قدرات الذكاء الاصطناعي.

وقالت الشركة إنها أجرت مراجعة لأكثر من 141 ألف اختبار للأمن السيبراني، رصدت خلالها ثلاث حالات تمكنت فيها نماذج "كلود" من الاتصال بالإنترنت واختراق بنية تحتية فعلية لجهات خارجية، مشيرة إلى أن أقدم هذه الحوادث يعود إلى شهر نيسان الماضي.

وأوضحت "أنثروبيك" أن الاختبارات كانت ضمن بيئات محاكاة تهدف إلى قياس قدرات النماذج في اكتشاف الثغرات، إلا أن خللاً في إعدادات الاختبار سمح للنماذج بالوصول إلى أنظمة حقيقية، من دون الكشف عن أسماء المؤسسات المتضررة.

وبيّنت الشركة أن النماذج الثلاثة التي نفذت الاختراقات شملت إصدارات من "كلود"، بينها "أوبوس 4.7" و"ميثوس 5"، إضافة إلى نموذج بحثي داخلي، مؤكدة أن هذه النماذج كانت تعمل من دون بعض إجراءات

الحماية المطبقة عادة في الأدوات المتاحة للمستخدمين.

وأشارت إلى أن عمليات الاختراق تمت باستخدام أساليب بسيطة، من بينها استغلال كلمات مرور ضعيفة، مؤكدة أنها شددت ضوابط الاختبارات الأمنية بعد هذه الحوادث، ودعت إلى إخضاع بيئات تقييم الذكاء الاصطناعي لمعايير حماية مماثلة للأنظمة الفعلية.

وتأتي هذه الحوادث بعد أيام من إعلان شركة "أوبن إيه آي" عن واقعة مشابهة، ما أعاد الجدل بشأن الحاجة إلى فرض رقابة أكبر على تطوير تقنيات الذكاء الاصطناعي وقدراتها المتقدمة.