ANDORRA DIGITAL

ANC-AD
Agència Nacional
de Ciberseguretat d'Andorra

Informe de Ciberintel·ligència

Shadow AI: el risc d'adoptar la IA generativa sense supervisió per a la seguretat de les dades

TLP: WHITE

NOVEMBRE 2024

FITXA DEL DOCUMENT

Versió	Redactat/Revisat per	Aprovat per	Data aprovació	Data publicació
1.0	ANC-AD	ANC-AD	21/11/2024	25/11/2024

Registre de canvis			
Versió	Pàgines	Data Modificació	Motiu del canvi

Propietari del document	ANC-AD

ÍNDEX

1. METODOLOGIA

Aquest informe aplica els principis de Traffic Light Protocol (TLP). És un esquema creat per fomentar un intercanvi més bo d'informació delicada (però no classificada) en l'àmbit de la seguretat de la informació.

A través d'aquest esquema, d'una manera àgil i senzilla, s'indica fins on pot circular la informació més enllà del receptor immediat, i aquest ha de consultar l'Agència Nacional de Ciberseguretat d'Andorra quan cal distribuir la informació a tercers.

Codi	Com es fa servir	Com es comparteix
TLP: RED	S'ha de fer servir **TLP:RED** quan la informació està limitada a persones concretes, i podria tenir impacte en la privacitat, la reputació o les operacions si es fa servir malament.	Els receptors no han de compartir informació designada com a **TLP:RED** amb cap tercer fora de l'àmbit on va ser exposada originalment.
TLP: AMBER	S'ha de fer servir **TLP:AMBER** quan la informació ha de ser distribuïda de manera limitada, però suposa un risc per a la privacitat, la reputació o les operacions si és compartida fora de l'organització.	Els receptors poden compartir informació indicada com a **TLP:AMBER** només amb membres de la seva pròpia organització que necessiten conèixer-la, i amb clients, proveïdors o associats que necessiten conèixer-la per protegir-se a si mateixos o evitar danys. L'emissor pot especificar restriccions addicionals per compartir aquesta informació.
TLP: GREEN	S'ha de fer servir **TLP:GREEN** quan la informació és útil per a totes les organitzacions que hi participen, com també amb tercers de la comunitat o el sector.	Els receptors poden compartir la informació indicada com a **TLP:GREEN** amb organitzacions afiliades o membres del mateix sector, però mai a través de canals públics.
TLP: WHITE	S'ha de fer servir **TLP:WHITE** quan la informació no suposa cap risc de mal ús, conforme a les regles i procediments establerts per a la seva difusió pública.	La informació **TLP:WHITE** pot ser distribuïda sense restriccions, únicament subjecta a controls de copyright.

2. INTRODUCCIÓ

La intel·ligència artificial generativa ha guanyat protagonisme com a una de les tecnologies més transformadores dels darrers anys. Aquestes eines, que inclouen models avançats de processament de llenguatge natural (PLN) i generació d'imatges, permeten crear contingut nou a partir de patrons apresos de dades existents. Per exemple, aplicacions, com ara el ChatGPT, generen textos coherents i detallats, mentre que eines com ara el DALL·E produeixen imatges basades en descripcions escrites.

La integració d'aquestes tecnologies en l'ecosistema empresarial està en auge a causa de la seva capacitat per automatitzar tasques repetitives, fomentar la creativitat i augmentar la productivitat. Tanmateix, mots experts apunten que aquesta integració s'està fent de manera desordenada i sense control, cosa que provoca que molts empleats hagin començat a fer servir aquestes eines sense una avaluació formal dels seus riscos.

Les organitzacions no estan adoptant la IA generativa mitjançant una estratègia integral que contempli una metodologia única per fer ús d'eines i tecnologies noves, cosa que està provocant entorns on els beneficis immediats podrien amagar en molts casos els possibles perills associats amb l'ús no regulat de dades delicades i la manca de supervisió.

A les conferències sobre BIG DATA & AI WORLD que es van celebrar durant l'octubre passat, Liher Elgezabal, WW DATA Security Technical Sales Leader de la IBM, va abordar el gran impacte que té la IA en les organitzacions des de la perspectiva de la seguretat de les dades, va posar xifres a aquest fenomen emergent, i va assegurar que fins al 75 % dels empleats ja fan servir la IA en el seu dia a dia a la feina. Tanmateix, va puntualitzar que en la gran majoria dels casos ho fan sense notificar-ho ni posar-ho en coneixement de les seves empreses.

L'ús no regulat d'eines d'IA pot exposar dades confidencials a plataformes externes, i comprometre la privacitat tot vulnerant regulacions de protecció de dades. Aquestes eines solen requerir l'entrada de dades delicades, com ara informació corporativa, detalls de clients o fins i tot estratègies comercials, que podrien ser emmagatzemades o processades en servidors fora del control de l'organització.

3. SHADOW AI: EL PERILL D'INTEGRAR TECNOLOGIES DE LA IA GENERATIVA SENSE SUPERVISIÓ A LES ORGANITZACIONS

El terme Shadow AI fa referència a l'ús d'eines i sistemes d'intel·ligència artificial (IA) en una organització sense la supervisió, autorització o coneixement explícit del departament de TI o d'altres àrees responsables de la governança tecnològica, cosa que pot derivar en riscos significatius en termes de seguretat, compliment i eficiència operativa.

Aquest concepte es deriva d'un altre, conegut com a Shadow IT, que fa referència a l'ús no regulat de programari o maquinari en l'àmbit corporatiu. Tanmateix, el Shadow AI s'enfoca específicament en el desplegament no controlat de tecnologies basades en la intel·ligència artificial.

El fenomen del Shadow AI sorgeix principalment per l'accessibilitat creixent d'eines d'intel·ligència artificial generativa i altres aplicacions basades en la intel·ligència artificial. Moltes d'aquestes solucions estan dissenyades per ser intuïtives i fàcils de fer servir, cosa que permet que els empleats de diverses àrees adoptin aquestes eines per resoldre problemes específics o augmentar la seva productivitat sense que calgui involucrar els departaments de TI o legal.

Tot i que aquestes iniciatives poden néixer d'unes bones intencions, el seu caràcter no regulat representa un desafiament crític per a la seguretat i la governança tecnològica.

3.1. Característiques principals del Shadow AI

Tot seguit, s'enumeren les característiques principals del Shadow AI:

- **Ús no autoritzat**: aquest fenomen es produeix quan els empleats descarreguen o fan servir eines d'intel·ligència artificial (com ara, plataformes de generació de textos, anàlisi predictiva o processament de dades) sense buscar l'aprovació ni informar les àrees corresponents.

- **Connectivitat no controlada**: aquestes eines solen operar al núvol, i requereixen l'enviament de dades a servidors externs, cosa que augmenta el risc de fuga de la informació.

- **Falta d'alineament amb les polítiques internes**: sovint, l'ús de Shadow AI no respecta les polítiques de seguretat informàtica, la privacitat de les dades o el compliment normatiu de l'organització.

- **Absència de monitoratge i suport**: com que no estan supervisades, aquestes eines no estan subjectes a les auditories de seguretat ni disposen del suport tècnic que cal en cas de fallades.

3.2. Factors que fomenten el Shadow AI

Entre els factors que fomenten el Shadow AI, cal destacar:

- **Facilitat d'accés**: les plataformes i les eines d'intel·ligència artificial han proliferat, i moltes d'elles són gratuïtes o de baix cost, cosa que ha provocat que qualsevol membre d'una organització pugui començar a fer-les servir sense barreres.

- **Baixa percepció del risc**: molts usuaris poden no ser plenament conscients de les implicacions de seguretat o compliment associades a l'ús d'aquestes tecnologies.

- **Pressió per la productivitat**: en entorns laborals competitius, els empleats poden recórrer a aquestes eines per complir objectius de manera més ràpida o creativa.

- **Interès dels empleats per fer tasques més complexes i creatives**: la intel·ligència artificial permet alliberar els empleats de les tasques pesades i repetitives i els possibilita centrar-se en aquelles que poden aportar més valor a l'organització.

- **Manca de capacitació o alternatives oficials**: es donen casos en què les empreses no proporcionen solucions d'intel·ligència artificial aprovades ni eduquen els seus equips sobre els riscs de l'ús no autoritzat.

3.3. Exemples comuns de Shadow AI

Alguns dels casos més comuns i que faciliten la comprensió d'aquest fenomen podrien ser aquells en què, per exemple:

- Empleats d'una entitat fan servir eines de generació de text o imatges, com ara el ChatGPT o el DALL·E, per crear contingut sense conèixer o revisar les polítiques de dades d'aquestes plataformes.

- Empleats que fan servir eines d'intel·ligència artificial amb l'objectiu de fer resums de reunions o correus electrònics.

- Equips de màrqueting que fan servir eines de la intel·ligència artificial amb l'objectiu d'analitzar dades de clients amb aplicacions d'anàlisi predictiva.

- Els professionals de recursos humans empren bots de conversa externs per automatitzar interaccions amb candidats o empleats.

4. RISCOS PRINCIPALS DE L'ÚS DE LA INTEL·LIGÈNCIA ARTIFICIAL GENERATIVA SENSE SUPERVISIÓ

Les empreses han d'elaborar una estratègia robusta per adoptar eines de la intel·ligència artificial generativa que contempli no només la seva implantació segura i conforme a la normativa, sinó també la creació de consciència organitzacional sobre els riscos de l'ús no autoritzat.

S'han d'establir controls clars, fomentar una governança estricta i desenvolupar una infraestructura que permeti aprofitar els beneficis de la intel·ligència artificial sense comprometre la integritat dels seus actius més valuosos: les dades. Si no es fa així, una organització es pot veure exposada a situacions crítiques com les que s'exposen tot seguit.

4.1. Seguretat de dades i fugues d'informació confidencial

La informació introduïda a les plataformes de la intel·ligència artificial generativa podria ser retinguda per proveïdors externs, intencionadament o pel mal disseny de les seves polítiques de retenció de dades. Per això, hi ha un risc real que l'ús no supervisat de la intel·ligència artificial pugui exposar les empreses a bretxes de seguretat quan es fan servir plataformes externes per processar dades internes.

Moltes eines de la intel·ligència artificial generativa emmagatzemen temporalment les dades que processen en els seus servidors, cosa que genera un risc de filtratge en cas de ciberatacs o mal ús d'aquesta informació.

Per exemple, un empleat podria introduir dades confidencials de clients en un bot de conversa generatiu per redactar un informe, sense saber que aquestes dades podrien ser retingudes pel proveïdor de l'eina. Això no només exposa l'empresa a possibles pèrdues econòmiques, sinó també a danys de la reputació si la informació arriba a tercers.

4.2. Compliment regulador i sancions

En molts sectors hi ha normatives estrictes sobre com es recopilen, emmagatzemen i processen les dades. Exemples com el Reglament General de Protecció de Dades (RGPD) a Europa o la Llei de privacitat del consumidor de Califòrnia (CCPA en anglès) als Estats Units, exigeixen que les empreses mantinguin un control estricte sobre on i com es gestionen les dades.

L'ús del Shadow AI complica el compliment d'aquestes normatives, perquè les dades es poden processar en jurisdiccions amb regulacions diferents o les poden fer servir proveïdors que no compleixen els estàndards legals. Per exemple, introduir dades personals en una eina d'intel·ligència artificial sense verificar la seva política de privacitat podria comportar multes cares i problemes legals.

4.3. Exposició a ciberamenaces i riscos de ciberseguretat

L'ús no regulat d'eines de la intel·ligència artificial pot crear bretxes de seguretat en la infraestructura d'una empresa. Els empleats que accedeixen a plataformes externes des de xarxes corporatives poden obrir les portes a atacs maliciosos, com ara la pesca o les infiltracions.

Per exemple, si un empleat fa servir una eina d'intel·ligència artificial generativa en un dispositiu sense protecció adequada, un atacant podria aprofitar la vulnerabilitat per infiltrar programari maliciós en la xarxa corporativa. La manca de controls en eines de Shadow AI pot facilitar atacs cibernètics, i incloure la infiltració per part d'actors maliciosos que busquen explotar vulnerabilitats en serveis externs d'intel·ligència artificial.

4.4. Pèrdua de control sobre el cicle de vida de les dades

Quan les dades d'una empresa són processades per eines de Shadow AI, sovint és difícil rastrejar com i on s'emmagatzemen. Això genera problemes a l'hora d'auditar el flux de dades o eliminar informació delicada, com exigeix el dret a l'oblit del RGPD.

A més, les empreses perden la capacitat de garantir que les dades s'usin exclusivament per a les finalitats previstes, cosa que podria obrir la porta a l'ús indegut o la comercialització d'informació per part de tercers.

4.5 Impacte en la consistència i la qualitat operativa

El Shadow AI també pot donar lloc a inconsistències en la qualitat dels resultats. Els models de la intel·ligència artificial generativa no sempre produeixen respostes precises, i sense una supervisió adequada, els empleats es podrien basar en informació errònia per prendre decisions crítiques.

Per exemple, un analista financer que fa servir un model de predicció no aprovat podria produir càlculs erronis, cosa que afectaria directament les decisions estratègiques de l'empresa.

5. CASOS D'ESTUDI D'INCIDENTS DERIVATS DEL SHADOW AI

5.1 Amazon: cas d'ús no autoritzat d'eines de la intel·ligència artificial (2020)

El 2020, Amazon va patir un incident de seguretat relacionat amb l'ús no autoritzat d'intel·ligència artificial (IA) per part d'un grup d'empleats. Aquests enginyers van decidir integrar un model d'IA de codi obert per optimitzar processos interns relacionats amb la gestió de dades de clients i productes, sense consultar-ho amb els equips de seguretat o TI de l'empresa. L'eina, que la van descarregar des de plataformes externes, va accedir a una base de dades delicada d'Amazon que contenia informació personal dels usuaris, com ara noms, adreces i detalls de les seves compres.

Tot i que l'eina es va implementar amb bones intencions per millorar l'eficiència, la manca de control i revisió per part dels responsables de seguretat de l'empresa va permetre que les dades fossin processades fora dels entorns segurs establerts per Amazon. Això va augmentar el risc que la informació pogués ser exposada a possibles filtracions o que es pogués fer servir de manera inapropiada, especialment perquè l'eina funcionava en servidors no aprovats.

Aquest incident subratlla els riscos que implica l'ús de tecnologies no verificades, fins i tot en corporacions grans com ara Amazon, on l'accés a dades delicades ha d'estar sempre sota un control estricte per garantir la privacitat i la seguretat dels usuaris.

5.2 Tesla: cas d'integració de la intel·ligència artificial no autoritzada (2021)

Aquest altre fet va afectar Tesla i es va produir un any més tard del d'Amazon. En aquest cas, Tesla va experimentar un problema relacionat amb l'ús no autoritzat de la intel·ligència artificial dintre del seu equip d'enginyeria que, amb l'objectiu de millorar l'anàlisi i manteniment de la seva flora de vehicles elèctrics, va integrar un model d'intel·ligència artificial descarregat d'un repositori públic de codi obert. Tanmateix, en fer-ho, no es van seguir els procediments estàndard de seguretat ni es va sol·licitar l'aprovació de l'equip de ciberseguretat de Tesla.

El model d'intel·ligència artificial que es va fer servir va accedir a una base de dades amb informació delicada sobre els vehicles i els seus propietaris, inclosos detalls sobre el manteniment, ubicacions i patrons d'ús. Tot i que no es va produir una filtració pública de dades, el fet que es processessin dades en servidors no autoritzats va generar preocupacions sobre la privacitat i la seguretat. Aquest incident va posar de manifest com, fins i tot a empreses innovadores com Tesla, la manca de control adequat sobre les eines tecnològiques pot comprometre la protecció de les dades dels usuaris i la confiança en la marca.

5.3 Accenture: cas de filtració de dades per intel·ligència artificial no controlada (2022)

El 2022 va ser la consultora Accenture la que es va veure afectada per l'ús d'eines d'intel·ligència artificial externes. Una vegada més va ser degut al fet que un grup d'empleats va implementar una solució d'intel·ligència artificial sense la revisió adequada dels equips de seguretat. L'eina, que estava dissenyada per analitzar

volums grans de dades, va accedir a informació confidencial de clients de perfil alt, incloses dades financeres i comercials.

L'ús no autoritzat d'aquesta eina externa va provocar que les dades no fossin processades sota les polítiques de seguretat de l'empresa. L'eina emmagatzemava alguns arxius sense el xifratge adequat, cosa que va permetre l'exposició d'informació delicada. El resultat d'aquest incident es va traduir en una filtració de dades important, cosa que va alertar les autoritats reguladores sobre el possible incompliment de les lleis de privacitat de les dades, com ara el RGPD a Europa.

Malgrat que Accenture va reaccionar ràpidament, aquest cas va demostrar que la manca de polítiques estrictes per a l'ús de la intel·ligència artificial pot posar en risc la privacitat de les dades dels clients i la reputació de l'empresa.

6. ESTRATÈGIES PER MITIGAR ELS RISCOS DEL SHADOW AI

6.1 Implementar una política clara de governança de la intel·ligència artificial

Les empreses han d'establir directrius clares sobre l'ús d'eines d'intel·ligència artificial, incloses les llistes de plataformes aprovades, les restriccions sobre quines dades es poden processar, i els procediments d'avaluació per a eines noves.

6.2 Educació i conscienciació

Els empleats han de rebre capacitació sobre els riscos del Shadow AI i com identificar plataformes no segures. Això també inclou ensenyar-los a avaluar polítiques de privacitat i termes d'ús abans d'adoptar una eina nova.

6.3 Invertir en eines corporatives aprovades

El fet de proporcionar eines d'intel·ligència artificial als empleats que siguin aprovades, segures i eficaces, reduirà la necessitat de recórrer a solucions no autoritzades.

6.4 Monitoratge proactiu de la infraestructura tecnològica

El departament de TI ha d'implementar solucions de monitoratge continuat per identificar i bloquejar l'ús d'eines no aprovades a la xarxa corporativa.

6.5 Establir acords clars amb proveïdors

Abans d'adoptar eines externes, les empreses han de negociar acords que garanteixin la protecció de les seves dades i el compliment normatiu.

7. CLÀUSULA DE CONFIDENCIALITAT

Aquest document és propietat de l'Agència Nacional de Ciberseguretat d'Andorra. Tota la informació que conté és confidencial, aquesta informació s'actualitzarà regularment per reflectir els possibles canvis dels productes i no podrà ser copiada o revelada a terceres persones sigui totalment o en part, sense consentiment previ exprés de l'Agència Nacional de Ciberseguretat d'Andorra.