Generalitat de Catalunya
Departament de la Presidència
Direcció General de Serveis Digitals i Experiència Ciutadana

Marc de gestió de riscos d'intel·ligència artificial de l'Administració de la Generalitat de Catalunya

Aprovat per la Comissió d'Intel·ligència Artificial de l'11 de desembre de 2024

Última actualització: 12 de maig de 2025

C. del Foc, 57, edifici B
08038 Barcelona
Tel. 93 857 40 00

1. Context, objecte i destinataris

1.1. Context

El Marc de gestió de riscos d'intel·ligència artificial (d'ara endavant, el Marc) s'ha de situar en l'escenari normatiu i estratègic actual, en què l'Administració de la Generalitat i el seu sector públic estan immersos en un procés de transformació digital amb l'impuls de tecnologies com ara la intel·ligència artificial (IA).

El Decret 76/2020 de 4 d'agost, d'Administració digital és l'instrument jurídic que persegueix la màxima eficiència en l'àmbit de les relacions internes i externes, estableix el marc de la governança de les dades com a element estructural del funcionament de l'Administració digital i, per tant, l'estratègia de governança de les dades, de la qual aquest marc forma part.

El Model de governança de les dades ha estat aprovat per Acord de Govern 158/2023, de 25 de juliol i, d'entre els seus àmbits regulatoris, es troba l'analítica i l'ètica de les dades. En aquest sentit, en desenvolupament del Decret 76/2020 de 4 d'agost, s'ha d'aprovar un protocol de governança de les dades que organitza i desplega aquest model, incloent-hi aspectes tècnics i organitzatius. Aquest protocol es concreta en polítiques que permeten als rols involucrats desenvolupar les seves funcions de manera transversal i homogènia; i aquestes estan acompanyades de guies que detallen la metodologia i els procediments operatius.

Així doncs, la política d'analítica, que desenvoluparà també la política d'intel·ligència artificial, preveu el desenvolupament de nous serveis digitals analítics i predictius que permetin donar resposta a necessitats concretes i ajudar en la presa de decisions. L'analítica de dades té per objectiu desenvolupar les actuacions relacionades amb el governança de les dades en relació amb l'analítica, el tractament massiu de dades i la intel·ligència artificial, entre d'altres, a l'organització.

Al seu torn, la Política referida a l'ètica contribueix a alinear el Model de governança de les dades amb els principis ètics de governança, gestió i consum de dades, que aporta més transparència, qualitat i equitat de les dades per a la ciutadania.

Així mateix, el Reglament (UE) 2024/1689 del Parlament Europeu i del Consell, de 13 de juny de 2024, conegut com a Reglament europeu d'Intel·ligència Artificial (RIA), estableix la necessitat de gestionar i controlar de forma diligent els riscos durant tot el cicle de vida d'aquests sistemes.

1.2. Objecte

El document té com a objectiu facilitar que l'Administració de la Generalitat de Catalunya adopti sistemes d'intel·ligència artificial (IA) fiables i centrats en les persones, tot garantint la protecció dels drets fonamentals de la ciutadania. En aquest sentit, es vol minimitzar els impactes negatius i maximitzar els positius dels sistemes d'IA, a través de la gestió efectiva dels riscos per generar confiança i proporcionar beneficis a la ciutadania. També pretén capacitar els desenvolupadors i persones usuàries d'IA per entendre els impactes, responsabilitzar-se de les limitacions dels seus sistemes i millorar-ne el funcionament.

El marc està dissenyat per ser flexible i adaptatiu davant nous riscos, caracteritzant-se per la innovació i iteració constants. Els resultats s'integraran en les polítiques, guies, directrius i orientacions de l'estratègia de governança de dades de l'Administració de la Generalitat i el seu sector públic.

Per a dubtes sobre la terminologia, es pot consultar l'Annex I.

1.3. Destinataris

El Marc ha d'establir un diàleg constructiu entre tecnologia, govern i societat. Per això, identificar i entendre els destinataris del document és essencial perquè les intervencions i polítiques proposades siguin pertinents i efectivament implementades, amb l'objectiu d'assegurar que els beneficis de la IA es maximitzin mentre que els seus riscos es minimitzen.

En aquest sentit, el principal destinatari és el **conjunt de l'Administració de la Generalitat i el seu sector públic**, així com els **proveïdors de l'Administració** en matèria d'IA. En qualsevol cas, la **ciutadania** també ha de poder incorporar les seves perspectives i preocupacions per assegurar que el desenvolupament i implementació de la IA es realitzi de manera transparent i justa, a través de la cocreació de serveis.

2. Rols i responsabilitats

Per desenvolupar un model de governança de la IA adequat, cal establir una estructura organitzativa que inclogui els agents implicats i les línies jeràrquiques corresponents. Els rols i responsabilitats presentats són una primera aproximació centrada en la gestió de riscos, que caldrà actualitzar i adaptar a mesura que es treballin, desenvolupin i provin les polítiques i guies de manera col·laborativa.

En l'àmbit central, dins la Direcció General competent en matèria d'intel·ligència artificial, ha d'existir:

- **Suport transversal a la governança algorísmica:** Identifica els riscos en l'àmbit algorísmic i proposa mesures mitigadores, a petició dels responsables de sistemes IA i altres algorismes. Així mateix, aplica els criteris ètics i els requeriments previstos a la Política i Guia metodologia d'ètica de les dades i s'encarrega d'escalar al Comitè d'Ètica de les Dades els riscos nous identificats. En aquest sentit, ha de comptar amb el suport d'una oficina externa experta pel que fa a l'avaluació algorísmica.

En l'àmbit federal, com a mínim, s'hi impliquen els següents rols:

- **Responsable del sistema d'IA**: Perfil operatiu que té un coneixement profund del seu àmbit de negoci, per la qual cosa és el responsable directe d'aplicar i aterrar el Marc de gestió de riscos al seu sistema d'IA.

- **Referent de Governança de les dades i IA:** Assessora i dona suport als responsables del sistema d'IA, vetlla i supervisa que es compleixin les directrius i els estàndards de governança i la gestió de riscos en tot l'àmbit departamental i el seu sector públic.

- **Responsables de protecció de dades i ciberseguretat:** Rols departamentals encarregats de vetllar per la protecció de dades i la ciberseguretat, que han de tenir un paper imprescindible en la implementació dels sistemes i la gestió dels seus riscos. Acompanyen, en l'àmbit de les seves competències, els responsables dels sistemes d'IA en el procés.

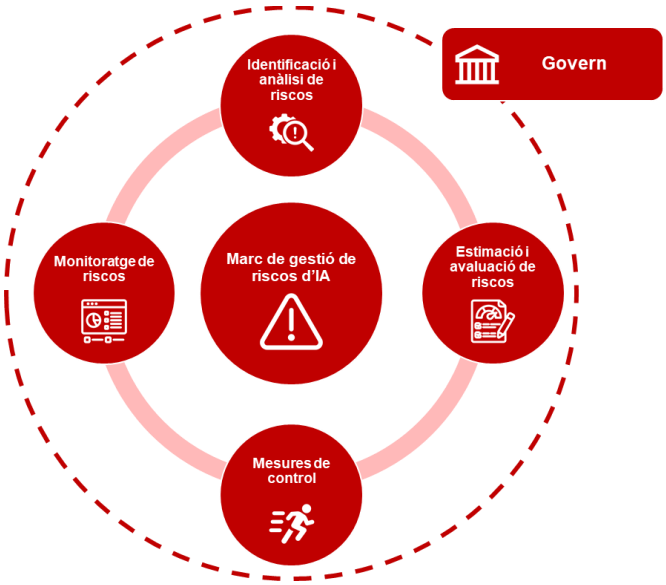
- **Rols de suport:** Existeixen altres perfils que habitualment ofereix el proveïdor tecnològic que poden intervenir en la gestió de riscos. En aquest sentit, sense voluntat exhaustiva, poden ser persones arquitectes de dades, enginyeres de dades, científiques de dades, enginyeres d'aprenentatge automàtic, entre d'altres.

3. Procés de gestió de riscos

La gestió efectiva de riscos d'IA és un procés continu que s'estén per totes les etapes del cicle de vida dels sistemes d'IA, adaptant-se oportunament a les evolucions i canvis de l'entorn.

En aquest sentit, el Marc de gestió de riscos d'IA està integrat per cinc fases clau amb una sèrie d'accions i resultats concrets per cadascuna:

- Governança.
- Identificació i anàlisi.
- Avaluació i estimació.
- Monitoratge.
- Mesures de control.



Il·lustració 1: Marc de Gestió de Riscos d'IA

D'acord amb el Marc, l'ordre de les fases es pot adaptar en funció del valor que aporti i de les necessitats i prioritats de cada responsable del sistema d'IA consideri. Habitualment, després de definir els resultats en la fase de Governança, es pot començar per la fase d'Identificació i anàlisi, seguit de l'Avaluació i estimació i les Mesures de control.

S'adoptarà un enfocament proactiu en la implementació de mesures de gestió de riscos, que asseguri un procés de mitigació de riscos integral i continu al llarg del desenvolupament i desplegament de sistemes d'IA.

En qualsevol cas, el procés ha de ser iteratiu, a través de la revisió i ajustament de les fases i l'establiment de connexions transversals per garantir una gestió de riscos d'IA eficient.

3.1. Identificació i anàlisi de riscos

En aquesta fase, es duu a terme una identificació i una anàlisi sistemàtica dels riscos coneguts i previsibles associats a cada sistema d'IA.

3.1.1. Comprensió holística del sistema d'IA

En primer lloc, és imprescindible, per a cada sistema d'IA, detectar i definir les capacitats, ús específic, beneficis esperats i costos per a una correcta gestió dels riscos.

Tota aquesta comprensió s'ha de documentar de tal manera que es faci **l'anàlisi i determinació** de les següents qüestions:

- **Necessitats i actors:** Identificació de les necessitats concretes que condueixen a la decisió d'implementar el sistema d'IA, així com de la conveniència d'usar aquesta tecnologia digital avançada i de les alternatives existents. Igualment, s'ha d'analitzar i determinar els possibles actors o parts interessades impactades per la implementació i ús d'aquest sistema.

- **Àmbit d'aplicació:** Determinació dels sectors o àmbits en què s'implementarà i s'utilitzarà cada sistema, així com de les tasques específiques a dur a terme.

- **Cost-benefici:** Anàlisi i constatació que els beneficis identificats justifiquen els costos que suposarà el desenvolupament, la implementació i el manteniment del sistema d'IA en conjunt, així com dels costos en un sentit ampli que poden sorgir a causa del seu mal funcionament.

- **Rols implicats i supervisió humana:** Definició, diferenciació i documentació clara dels diversos rols i responsabilitats humanes en usar, interactuar o administrar sistemes d'IA, així com els processos de supervisió relacionats.

3.1.2. Identificació i nivell de risc del sistema d'IA

D'acord amb l'article 3 del RIA, el risc és "la combinació de la probabilitat que es produeixi un perjudici i la gravetat d'aquest perjudici". La identificació i anàlisi de riscos implica l'ús sistemàtic de la informació disponible per identificar perills, és a dir, qualsevol "font potencial de dany". Un "dany" és "qualsevol efecte advers sobre la salut, seguretat i drets fonamentals".

El RIA estableix una classificació dels riscos associats amb l'ús i la implementació de tecnologies d'intel·ligència artificial, amb l'objectiu de garantir la seguretat i la protecció dels drets fonamentals dels individus.

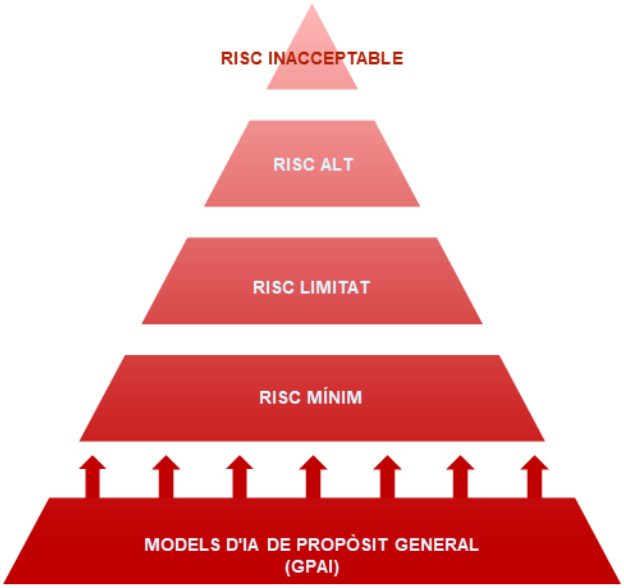
Aquesta classificació es divideix en quatre categories principals de **nivell de risc**, de més a menys limitacions:

- **Inacceptable (article 5):** Sistemes d'IA que presenten riscos tan elevats que la seva utilització no es considera admesa dins d'una societat democràtica, per la qual cosa estan prohibits.

- **Alt (article 6 i annex III):** Sistemes d'IA que presenten riscos molt rellevants per a la salut, la seguretat o els drets fonamentals de les persones físiques. En aquest sentit, s'estableix una regulació més estricta i controls més detallats per mitigar els possibles impactes adversos.

- **Limitat:** Els sistemes d'IA inclosos en aquesta categoria poden tenir un impacte significatiu en els drets individuals. En qualsevol cas, són considerats gestionables mitjançant mesures de protecció i controls adequats, basades essencialment en accions de transparència (article 50).

- **Mínim:** Aquesta categoria engloba sistemes d'IA que tenen un impacte mínim o nul sobre els drets i llibertats de les persones. Són la majoria dels sistemes d'IA i es recomana l'adopció de codis de conducta voluntaris (article 95).



Il·lustració 2: Piràmide de riscos del RIA

3.1.3. Identificació de riscos

En primer lloc, cal **identificar tots els riscos** passats, presents i futurs. En aquest sentit, el RIA diferencia entre el risc conegut i el risc previsible perquè s'incloguin ambdós casos en la identificació dels riscos:

- Un risc és **conegut** quan el dany s'ha produït en el passat o és segur que ho farà en el futur. A més, serà conegut si l'Administració de la Generalitat ho pot saber amb un esforç raonable.

- Un risc és **previsible** si encara no s'ha esdevingut, però ja pot ser identificat. En aquest sentit, com més gran sigui l'impacte potencial del risc, més esforç haurà d'invertir l'Administració a preveure'l.

Si bé el RIA no especifica com s'han d'identificar els riscos, el Marc es basa en tècniques i mètodes existents com taxonomies de riscos, bases de dades d'incidents o anàlisi d'escenaris. Així mateix, a l'Annex II hi ha una relació de riscos classificats pel moment del cicle de vida del sistema d'IA en què es poden produir i possibles mesures de control.

Addicionalment, en cas que el sistema d'IA impliqui el tractament de dades de caràcter personal també s'ha tenir en compte l'establert als articles 32, 35 i 36 de l'RGPD per a la identificació del nivell de risc. En aquests supòsits, s'ha de considerar tant el Marc de Ciberseguretat per a la Protecció de Dades com la metodologia d'**avaluació d'impacte relativa a la protecció de dades** desenvolupades per l'Agència de Ciberseguretat de Catalunya.

Per altra banda, en cas que el sistema d'IA sigui d'alt risc, s'haurà de dur a terme l'**avaluació d'impacte de drets fonamentals,** d'acord amb l'article 27 del RIA. Segons l'apartat quart de l'article, si ja s'ha complert amb les obligacions amb l'avaluació d'impacte de protecció de dades, aquesta avaluació de drets fonamentals serà complementària de la primera.

3.2. Estimació i avaluació de riscos

Després d'haver identificat i analitzat els riscos, aquests s'han d'estimar i avaluar. D'aquesta manera, s'ha de considerar que el risc és resultat de la combinació entre probabilitat i severitat:

- **Probabilitat:** La freqüència amb què s'espera que es produeixi un fet de risc.

- **Severitat o acceptabilitat:** L'impacte potencial o les conseqüències negatives de la materialització d'aquest risc.

Així doncs, per a cadascun dels riscos identificats, s'ha d'avaluar si és probable i si és sever. Per tant, l'operació consisteix a estimar la probabilitat, avaluar la severitat i determinar el risc a través de la fórmula "**risc = probabilitat x severitat**".

En funció del resultat de l'operació, el risc serà més o menys crític, i haurà d'implicar una major o menor gestió. Una forma senzilla d'identificar-ho és a través d'una matriu d'avaluació, un instrument visual per categoritzar els riscos en funció d'aquests dos paràmetres.

Per altre costat, una bona manera d'assegurar que tots els riscos s'han considerat degudament són les **llistes de comprovació**. En aquest sentit, l'Estratègia d' intel·ligència artificial de Catalunya, Catalonia.AI, té entre els seus pilars fonamentals l'Observatori d'Ètica en Intel·ligència Artificial de Catalunya (OEIAC). Aquest ha desenvolupat **el Model PIO**

(Principis, Indicadors i Observables), una eina pionera que conté recomanacions ètiques i obligacions legals, i afavoreix el compliment de la normativa i regulació actual sobre la IA a través d'un procés de verificació exhaustiu.

Aquest model es basa en set pilars: transparència i explicabilitat, justícia i equitat, seguretat i no maleficiència, responsabilitat i retiment de comptes, privacitat, autonomia i sostenibilitat. A més, és una eina que afavoreix el compliment de la normativa i regulació actual sobre riscos associats a la IA a través d'un procés de verificació exhaustiu; permet identificar accions adequades o inadequades i sensibilitzar a través dels usos ètics i responsables de les dades i sistemes d'IA.

3.3. Mesures de control

Es poden identificar pràctiques comunes aplicables de manera general, tot i que cada risc identificat exigeix un pla d'acció adaptat a les seves característiques i contextos concrets. En aquest sentit, les estratègies generals de control de riscos són:

- **Evitació dels riscos:** Si el risc potencial sobrepassa el valor esperat o les conseqüències negatives són inacceptables, pot ser necessari no desenvolupar aquest sistema d'IA.
- **Prevenció dels riscos:** Implementació d'accions proactives per evitar que els riscos es materialitzin.
- **Mitigació dels riscos:** Implantació de pràctiques de disseny i desenvolupament que minimitzin la probabilitat de fallades o limitin la severitat dels errors.
- **Transferència dels riscos:** Se suma un tercer que pugui controlar el risc.
- **Acceptació dels riscos:** El risc és baix o les mesures per mitigar-lo són desproporcionades en relació amb el benefici que s'obté.

Per altra banda, s'ha de considerar l'**existència del risc residual**, és a dir, la part del risc restant després d'haver aplicat les mesures de gestió i control en el risc inicial. En aquest sentit, cal tenir en compte que no sempre és possible eliminar per complet el risc després de l'aplicació dels controls pertinents.

Així mateix, les mesures de control seran diferents en cada cas. Tot amb tot, serà útil tenir en compte les següents mesures per a tots els sistemes:

- **Transparència**: L'accés al Registre intern per part dels empleats públics i de la versió pública per a la ciutadania permet implicar altres actors, més enllà de la unitat i del seu àmbit departamental, en la detecció de riscos.

- **Supervisió humana:** Tant si el sistema requereix intervenció directa en el disseny humana (*human-in-the-loop*) o no (*human-over-the-loop*), aquests han de tenir un control humà al darrere que vetlli pels possibles riscos, l'ètica i el compliment normatiu.

- **Manteniment de les garanties administratives habituals:** Les decisions administratives tenen un sistema de garanties per assegurar els drets de les persones, essencialment a través dels recursos. En aquest sentit, els casos en què intervé un sistema d'IA han d'aplicar les mateixes garanties que en una actuació ordinària.

En aquesta fase, s'establiran els indicadors i el pla de recuperació que s'assenyalen a l'apartat 3.4.

3.4. Monitoratge de riscos

3.4.1. Eines de monitoratge

El monitoratge és la fase que implica el procés continu de seguiment i revisió dels riscos per mantenir una vigilància efectiva, així com per garantir una resposta adequada davant qualsevol canvi d'estatus o impacte potencial. En aquest sentit, es monitora per a facilitar:

- **Generació d'alertes**: L'automatització d'alertes i la seva revisió permet anticipar riscos. Per exemple, monitorar el consum d'un sistema respecte de la potència contractada al núvol pot ajudar a detectar la necessitat de contractar-ne més.
- **Millora contínua:** El monitoratge ha de permetre una revisió periòdica exhaustiva del funcionament del servei.

La selecció de l'eina de monitoratge s'ha de fer en funció de la criticitat, tenint en compte que tots els sistemes amb un risc alt n'hauran de tenir obligatòriament. Aquest seguiment s'ha de dur a terme per a cada sistema d'IA, però en alguns casos es podran agregar les dades. Per exemple, a través d'un quadre de comandament de tots els bots de conversa de l'Administració de la Generalitat.

En aquest sentit, els **indicadors claus de rendiment** són una eina per mesurar l'efectivitat de les estratègies de mitigació de riscos implementades en els sistemes d'IA. Aquests indicadors monitoraran per a cada sistema, com a mínim, les següents qüestions:

- Funcionament del sistema.
- Riscos associats (per exemple, el cost del núvol contractat o l'ús de llenguatge agressiu d'un bot de conversa).
- Funcionament del sistema de gestió de riscos.
- Incidències i temps de reacció.

Més enllà dels indicadors claus de rendiment, existeixen **altres eines de monitoratge** que poden ser útils en diferents casos, com ara els informes de seguiment, l'elaboració de taules i gràfics, panells de control interactius o matrius de risc, entre d'altres.

3.4.2. Pla de recuperació

Un pla de recuperació és un conjunt de protocols i procediments dissenyats per restaurar la funcionalitat i les operacions d'un sistema d'IA després d'un incident o crisi:

- Un **incident** es defineix com qualsevol esdeveniment inesperat o no planificat que afecta la funcionalitat o l'operativitat normal d'un sistema d'IA, però que es pot gestionar i resoldre amb procediments establerts i genera un mínim impacte en el servei.

- Una **crisi** és la condició inestable que implica un canvi abrupte o significatiu imminent que requereix atenció i mesures urgents, especialment per a protegir els drets fonamentals de les persones.

En aquest sentit, els incidents es gestionen de forma federal, en l'àmbit de cada departament. En canvi, les crisis caldrà abordar-les de manera coordinada entre els rols federals i els rols centrals que es determinin, als quals s'haurà d'avisar al més aviat possible.

S'establirà un pla de recuperació marc, que cada unitat adaptarà a cadascun dels seus sistemes. En qualsevol cas, cada pla ha de contenir, com a mínim, els següents elements comuns:

- **Avaluació de danys:** Determinació de l'extensió de l'impacte en els sistemes d'IA i les seves causes.
- **Mitigació dels efectes:** Aplicació de mesures immediates per afrontar i reduir els danys causats.
- **Restabliment d'operacions:** Represa de les operacions crítiques a través de sistemes de suport o solucions alternatives.
- **Comunicació:** S'informa els actors implicats de la incidència, de l'estat de recuperació i del restabliment de les operacions. S'ha d'avisar els rols centrals quan la situació es produeixi en un sistema amb el nivell de risc identificat com a alt.
- **Revisió de seguretat:** Cal assegurar que els sistemes d'IA siguin segurs i estables, i que la crisi no es pugui tornar a produir.

Tot amb tot, si es vol assegurar el seu èxit, s'han de dur a terme simulacres per avaluar la capacitat de resposta del pla davant els diferents escenaris d'incidència. Així mateix, la planificació ha de ser avaluada periòdicament en funció dels canvis tecnològics i l'experiència acumulada.

3.4.3. Millora contínua, retroalimentació i lliçons apreses

L'objectiu d'aquesta etapa és capturar, analitzar i aplicar lliçons apreses de la gestió de riscos d'IA per millorar contínuament les pràctiques i polítiques de riscos. Aquest cicle de retroalimentació assegura que l'Administració no només reaccioni als incidents de riscos, sinó que proactivament ajusti els seus mètodes de gestió a partir de les experiències.

En aquest sentit, l'etapa consta dels següents passos ordenats:

- **Recol·lecció de dades:** Compilació sistemàtica de dades i observacions de totes les fases del procés de gestió de riscos, inclosos incidents, mitigacions exitoses i fallides, i retorn de les parts interessades.
- **Anàlisi d'informació:** Avaluació de les dades recollides per identificar patrons, causes subjacents d'èxits o fracassos, i oportunitats de millora.
- **Disseminació de coneixements:** Sistematització i compartició de les lliçons apreses amb el conjunt de l'Administració per garantir que el coneixement influeixi en la pràctica.

- **Implementació de millores:** Aplicació pràctica de les lliçons apreses per ajustar les estratègies i processos de gestió de riscos.
- **Avaluació de l'impacte:** Mesurar l'efectivitat de les millores implementades per tancar el cicle de retroalimentació.

3.5. Governança

La governança de riscos és un component transversal que impregna tot el procés de gestió de riscos d'IA, a través de la facilitació i reforç de les altres fases del procés.

El **canvi de cultura i organització administratives** és imprescindible per implementar correctament els principis i pràctiques d'IA. Aquest fet es materialitza a través de:

- L'establiment d'uns **principis ètics de l'ús de la IA** a l'Administració.
- La **capacitació i sensibilització** dels empleats públics en l'àmbit de les implicacions ètiques de la IA.
- La definició de **bones pràctiques**, a través de la cocreació dels sistemes d'IA, el compliment de la normativa vigent i la perspectiva ètica.

En el context de l'Administració de la Generalitat, aquest enfortiment es duu a terme en consonància amb el treball del Comitè d'Ètica de les Dades i el Model de governança de la dada, i del Pla de capacitació digital del personal.

En qualsevol cas, la coordinació amb el Comitè d'Ètica de les Dades és clau. Per això, s'haurà de definir un flux entre aquest òrgan i la gestió de riscos per a respondre les necessitats i situacions noves que apareguin i presentin dubtes o calgui avaluar des d'aquesta perspectiva.

3.5.1. Principis

La governança ètica constitueix un eix central de l'estratègia de la governança de les dades de l'Administració de la Generalitat, per la qual cosa és necessari disposar i aplicar principis i conductes ètiques garanteix un ús de la IA fiable, segur i ètic.

En aquest sentit, la Comissió d'Intel·ligència Artificial, a proposta del Comitè d'Ètica de les Dades, va aprovar els **principis i criteris ètics en l'ús de les dades i la intel·ligència artificial** l'11 de desembre de 2024, que són els següents:

- **Justícia, equitat i no-discriminació**. Assegurar l'ús equitatiu, just i respectuós de les dades (equitat, no-discriminació i respecte als drets humans).
- **Acció i supervisió humana**. Garantir la protecció dels drets fonamentals mitjançant la intervenció humana i l'ús responsable de les dades.
- **Confidencialitat, privacitat i seguretat**. Protegir la informació personal, garantir la seguretat digital i custodiar la informació confidencial.

- **Governança de les dades avançades i de sistemes d'IA**. Supervisar que el disseny, desenvolupament, implementació i ús avançat de dades i de sistemes d'IA i algorismes es porta a terme en coherència amb la protecció dels drets humans i la responsabilitat pública, i que els sistemes d'IA són transparents.
- **Governança dinàmica de l'ètica**. Fomentar una avaluació constant de l'ètica mitjançant l'ús de dades i tecnologies d'IA i garantir l'actualització del desenvolupament ètic de la IA, la seva implementació i el seu foment.
- **Transparència i retiment de comptes**. Fomentar la difusió amb caràcter permanent i actualitzat i la claredat i el retiment de comptes en l'ús de dades, IA i altres algorismes, així com la participació ciutadana i els processos de treball transparents.
- **Desenvolupament sostenible i protecció del medi ambient**. Promoure pràctiques tecnològiques que suportin la sostenibilitat i protegeixin el medi ambient.

Aquest és el marc de referència en aquest àmbit de l'Administració de la Generalitat i el seu sector públic. En aquest sentit, d'acord amb el Model de governança de les dades de l'Administració de la Generalitat, la implementació d'aquests principis es facilitarà a través de directrius i protocols específics, que inclouen auditories regulars per detectar i corregir biaixos, així com processos de revisió que assegurin l'adequació i l'actualització permanent dels algorismes utilitzats.

3.5.2. Capacitació i sensibilització

Per tal de capacitar, formar i sensibilitzar els empleats públics en l'àmbit de la IA, l'ètica i la gestió de riscos cal:

- **Desenvolupament d'itineraris formatius**: Cada rol ha de disposar de la formació contínua adequada per a la seva implicació. En aquest sentit, els itineraris formatius han de ser inclusius i accessibles, a través d'una varietat de formats i plataformes per maximitzar el seu abast i efectivitat. En funció dels rols:
    - **Tots els empleats públics:** Han de comprendre d'una forma bàsica com funcionen els sistemes d'IA i com hi interactuen en les seves tasques del dia a dia.
    - **Perfils específics:** Necessiten estar al dia de les darreres tecnologies i pràctiques de l'àmbit, per la qual cosa els caldrà una formació més avançada.
- **Programació d'accions de sensibilització:** Les accions han d'identificar i comunicar riscos potencials, promoure una cultura d'ètica de la IA i proporcionar orientació sobre la gestió de riscos.

3.5.3. Bones pràctiques

La participació i involucració dels actors implicats en el cicle de vida d'un sistema d'IA és cabdal per a una gestió de riscos efectiva i responsable. D'aquesta manera, per a cada sistema **els actors seran diferents**, però sovint els seus perfils seran molt diversos: des de jurídics, a tecnològics; i d'unitats de caràcter transversal a unitats amb competències específiques. Aquest fet, que podria semblar una dificultat, ha de servir per impulsar projectes amb una **mirada transversal i multidisciplinària.**

Per això, s'estableix la **cocreació**, és a dir, el disseny participatiu, **com a metodologia** per a la creació i posada en servei dels sistemes. D'aquesta manera, totes les parts s'involucren en el procés, hi contribueixen de manera significativa, en coneixen el context i abast i, d'aquesta manera, fomenten una gestió de riscos del sistema en els seus àmbits respectius.

Com que el sistema neix de la implicació dels diferents actors, respon a les seves necessitats i, per tant, és més útil i usable. Així mateix, pel caràcter multidisciplinari, s'hi incorporen les diferents perspectives que cal tenir en compte, cosa que pot afavorir la reducció del risc de biaixos i problemes futurs.

Tots els sistemes d'IA han d'**operar dins dels paràmetres de la normativa vigent**, a través del foment dels drets fonamentals de les persones, així com han de poder adaptar-se ràpidament als canvis normatius. Per això, els equips encarregats del disseny i desenvolupament de sistemes d'IA han de treballar amb les assessories jurídiques de les unitats corresponents, així com amb les persones delegades de protecció de dades i responsables de ciberseguretat.

Així mateix, més enllà de l'imprescindible compliment de la legalitat, és essencial **incorporar la perspectiva de l'ètica des del disseny** en tot el procés de gestió de riscos. L'ètica implica pensar en les conseqüències de les accions que es duen a terme i té una relació directa amb la presa de decisions i els criteris i valors que s'hi incorporen.

Annex I. Definicions

Per facilitar una major comprensió del Marc es presenten a continuació una sèrie de definicions del document, la major part de les quals basades en el RIA:

Criticitat: Importància d'un risc potencial associat amb l'ús d'un sistema d'IA, tenint en compte la probabilitat que es produeixi multiplicada per la seva severitat. La criticitat ajuda a prioritzar les accions de mitigació i els recursos destinats a la gestió dels riscos.

Dades personals: les dades personals definides a l'article 4, punt 1, del Reglament (UE) 2016/679.

Finalitat prevista: l'ús al qual el proveïdor destina un sistema d'IA, inclosos el context i les condicions d'ús específics, tal com es determina en la informació facilitada pel proveïdor en les instruccions d'ús, el material promocional o de venda i les declaracions, així com en la documentació tècnica.

Human-in-the-loop: Procés en què una persona humana participa activament en la presa de decisions d'un sistema d'IA, a través de la supervisió i ajustament dels resultats generats per la màquina per garantir que siguin precisos i adequats.

Human-over-the-loop: Procés en què una persona humana supervisa el funcionament d'un sistema d'IA a un nivell més alt, i intervé només en casos excepcionals o quan el sistema necessita correccions per assegurar que funciona correctament i d'acord amb els objectius desitjats.

Incident greu: qualsevol incident o fallada de funcionament d'un sistema d'IA que provoqui directament o indirectament qualsevol de les següents situacions:

(a) la mort d'una persona o un dany greu per a la salut d'una persona.

(b) pertorbació greu i irreversible de la gestió i el funcionament d'infraestructures crítiques.

(b bis) incompliment de les obligacions derivades del Dret de la Unió destinades a protegir els drets fonamentals.

(bb) danys greus a la propietat o al medi ambient.

Indicació: Seqüència de text o línia de codi que l'usuari introdueix en un model d'intel·ligència artificial amb l'objectiu d'obtenir una resposta. Conegut en anglès com a "*prompt*".

Model d'IA de propòsit general: un model d'IA, fins i tot quan s'ha entrenat amb una gran quantitat de dades utilitzant l'autosupervisió a escala, que mostra una generalitat significativa i és capaç de realitzar de forma competent una àmplia gamma de tasques diferents, independentment de la forma en què es comercialitzi el model, i que pot integrar-se en una varietat de sistemes o aplicacions posteriors. Això no inclou els models d'IA que s'utilitzen abans de la seva comercialització per a activitats de recerca, desenvolupament i creació de prototips.

Posada en servei: el subministrament d'un sistema d'IA per al seu primer ús directament a l'usuari o per a ús propi a la Unió per a les finalitats previstes.

Priorització de risc: selecció específica i ordenada dels riscos detectats segons el seu nivell de preponderància i impacte, amb l'objectiu d'enfocar els recursos existents a primar el tractament d'aquells riscos més importants i danyosos enfront d'altres més lleus. La creació d'expectatives poc realistes sobre els riscos pot portar les organitzacions a localitzar recursos de manera ineficient per solucionar els riscos. Tanmateix, adoptant una cultura de gestió de riscos d'IA pot ajudar les organitzacions a adonar-se que no tots els riscos són igualment importants. És per això que es proposa una priorització de riscos, d'aquells més rellevants i el dany dels quals pot ser més gran per als actors implicats o impactats. D'aquesta manera, els recursos es distribueixen eficientment per solucionar i rebaixar i o mitigar els possibles danys derivats d'aquells riscos més rellevants.

Proveïdor: una persona física o jurídica, autoritat pública, agència o un altre organisme que desenvolupi un sistema d'IA o un model d'IA de propòsit general o que faci desenvolupar un sistema d'IA o un model d'IA de propòsit general i els comercialitzi o posi en servei sota el seu propi nom o marca comercial, sigui a títol onerós o gratuït.

Rendiment d'un sistema: la capacitat d'un sistema d'intel·ligència artificial per aconseguir la seva finalitat prevista.

Risc residual: risc que roman després d'haver acomplert les tasques de mitigació de riscos. Caldrà documentar els riscos residuals perquè el proveïdor d'aquest sistema consideri els riscos de desplegar el producte d'IA i pugui informar els seus usuaris finals dels impactes negatius d'interactuar amb ells.

Risc: combinació de la probabilitat que es produeixi un dany i la gravetat d'aquest dany.

Sistema d'IA: sistema basat en màquines dissenyat per funcionar amb diversos nivells d'autonomia i que pot mostrar capacitat d'adaptació després del seu desplegament i que, per a objectius explícits o implícits, infereix, a partir de l'entrada que rep, com generar sortides com ara prediccions, continguts, recomanacions o decisions que poden influir en entorns físics o virtuals.

Tolerància al risc: preparació d'una Administració o actor de la IA per suportar prou risc per a aconseguir els seus objectius. Pot ser influenciat per requisits legals o regulatoris establerts per organitzacions, indústries, comunitats o governs. Diferents organitzacions poden tenir diferents toleràncies al risc. Les organitzacions han de seguir les regulacions existents i guies per a definir els criteris de risc, tolerància i resposta establerts per requisits organitzatius, de domini, disciplina, sector o professionals.

Ús indegut raonablement previsible: l'ús d'un sistema d'IA d'una manera no conforme amb la seva finalitat prevista, però que pot resultar d'un comportament humà raonablement previsible o de la interacció amb altres sistemes, inclosos altres sistemes d'IA.

Annex II. Riscos del cicle de vida d'un sistema d'IA i mesures de control aplicables

A continuació es presenten els riscos associats a les cinc etapes del cicle de vida d'un sistema i les mesures de control que s'hi poden aplicar.

1. Descobriment de l'oportunitat

Fase de definició dels objectius i resultats desitjats del projecte, a la vegada que se seleccionen i alineen les expectatives de les parts implicades. Així mateix, es defineixen els recursos i passos decisius i els indicadors per mesurar-ne l'èxit. Aquesta presenta els següents riscos i possibles mesures de control a aplicar:

Nom del risc	Descripció del risc	Mesures de control del risc
Responsabilitat del sistema	Si no es documenten adequadament les decisions, pot ser difícil determinar la responsabilitat per un comportament inesperat o un ús indegut.	Implementar documentació detallada i assignació clara de responsabilitats.
Responsabilitat jurídica	Cal determinar qui és responsable del model fundacional, amb l'objectiu de respondre davant els organismes reguladors corresponents.	Definir la propietat i responsabilitats operacionals per a models fonamentals.
Propietat intel·lectual	El fet de no determinar la propietat del contingut generat per IA pot generar inseguretat jurídica sobre els drets de propietat intel·lectual relacionats amb els continguts generats.	Revisar i ajustar estratègies davant canvis en lleis de propietat intel·lectual. Desenvolupar estratègia per al contingut generat per IA de compliment normatiu en aquest àmbit.
Funcions i responsabilitats poc clares de la governança de les dades	La manca de claredat dins dels processos de governança de les dades, la qual cosa dona lloc a un impacte regulatori, financer, de resiliència i de reputació.	Establir estructures clares de governança i responsabilitats, d'acord amb el Model de governança de les dades i els instruments que el desenvolupen.
Impacte en el medi ambient	El consum de grans quantitats d'energia i d'aigua per a l'entrenament d'IA generen una petjada de carboni i hídrica que cal tenir en compte.	Implementar pràctiques d'eficiència energètica i ús d'energies renovables per reduir la petjada de carboni i hídrica dels centres de dades.
Manca d'alineament estratègic	Sense una estratègia clara en l'àmbit de la IA, pot ser problemàtic exposar l'Administració de la Generalitat a un risc significatiu al voltant de l'ús inadequat de la IA. A més, pot existir una manca d'alineament entre els sistemes d'IA i	Desenvolupar i mantenir una estratègia clara per a la implementació de la IA que alineï els objectius tecnològics amb els de les polítiques públiques.

	els objectius estratègics de l'Administració, així com una manca de confiança en l'ús d'aquests sistemes, la qual cosa pot dificultar la seva implementació i ús. Així mateix, la manca de coneixement sobre els processos i context de l'Administració pot obstaculitzar la correcta implementació i ús de la IA.	
Desalineament amb els valors culturals i ètics de l'Administració de la Generalitat	Les decisions preses per la solució d'IA han d'estar alineades amb els valors culturals i ètics de l'Administració, o provoquen decisions dolentes o incorrectes de les quals s'ha de fer responsable l'organització.	Establir revisions periòdiques dels models d'IA per garantir que les seves decisions reflecteixin els valors i ètica de l'Administració.

2. Descoberta i recollida de dades

Fase de recollida de dades rellevants i preparació per a l'aprenentatge del sistema d'IA. Aquesta fase acostuma a ser la que requereix més temps i té els següents riscos:

Nom del risc	Descripció del risc	Mesures de control del risc
Biaixos de les dades	Els biaixos històrics, de representació i socials (incloent-hi lingüístics) presents en les dades utilitzades per entrenar i afinar el model poden afectar negativament el comportament del model. L'entrenament amb dades esbiaixades podria conduir a resultats esbiaixats que poden representar injustament o discriminar determinats grups o individus per diverses raons, fets prohibits per la legislació vigent.	Implementar eines d'avaluació i mitigació de biaixos durant la preparació de dades.
Cura de les dades	Si no es recopilen o preparen les dades d'entrenament o ajustament de forma correcta, el resultat pot ser una desalineació dels valors o la intenció desitjats d'un model i el resultat real.	Alinear les dades amb intencions del model mitjançant gestió rigorosa.

Procedència de les dades	En cas de manca de procediments estandarditzats i establerts per verificar d'on provenen les dades, no hi ha garanties que les dades disponibles siguin el que diuen ser. Si les dades s'han recopilat, manipulat o falsificat de forma poc ètica, l'ús d'aquestes pot donar lloc a comportaments no desitjats en el model.	Adoptar mètodes estandarditzats per a la verificació de la fiabilitat de les dades.
Reentrenament posterior	La reutilització del resultat per tornar a entrenar un model no es pot dur a terme sense implementar una investigació humana adequada, que minimitzi les possibilitats que s'incorporin outputs no desitjats a les dades d'entrenament o ajust del model.	Integrar la supervisió humana en el procés de reutilització de dades.
Corrupció de dades a causa de la interacció no intencionada amb altres sistemes	La interacció entre la solució d'IA i altres entitats, fonts de dades o sistemes basats en el núvol pot donar lloc a canvis no desitjats en l'entrada o sortida de dades.	Desenvolupar protocols per gestionar interaccions entre sistemes d'IA i altres fonts.
No compliment de limitacions de transferència, ús i adquisició de dades	La normativa vigent pot limitar o prohibir la transferència, l'ús o la recopilació de certs tipus de dades per a casos d'ús específics d'IA. Aquestes restriccions poden afectar la disponibilitat de les dades necessàries per entrenar un model d'IA i poden donar lloc a dades mal representades, però el seu ús no és possible per part de l'Administració i s'han de buscar alternatives per suplir aquesta mancança.	Establir polítiques que compleixin amb lleis de transferència, ús i recopilació de dades.
Vulneració dels drets de les persones sobre les seves dades personals	La identificació o l'ús indegut de les dades pot donar lloc a la violació dels drets en aquest àmbit, per la qual cosa s'ha d'incorporar la protecció de dades personals des del disseny.	Implementar pràctiques de gestió de dades i privacitat per protegir informació de les persones usuàries, sota la supervisió i assessorament del Delegat o Delegada de Protecció de Dades (DPD).
Inclusió d'informació confidencial en la indicació	Divulgació d'informació personal o confidencial com a part d'una indicació (prompt) que s'envia al model. Les dades de la indicació es poden emmagatzemar o usar posteriorment per a altres finalitats, com l'avaluació i el reentrenament del model, però aquest fet s'ha d'analitzar des d'una òptica de protecció de dades. Si el sistema no es desenvolupa adequadament, el model podria revelar informació confidencial o IP en el resultat generat, o la informació confidencial dels usuaris finals es podria recopilar i emmagatzemar de forma involuntària.	Protegir la informació confidencial a través dels protocols de gestió i emmagatzematge de dades.

Reidentificació	Fins i tot amb l'eliminació de la informació d'identificació personal (PII) i la informació personal confidencial (SPI) de les dades, és possible que encara es puguin identificar les persones a causa d'altres característiques disponibles en les dades. Les dades que poden revelar dades personals o confidencials s'han de revisar adequadament.	Emprar tècniques avançades d'anonimització i gestió de dades.
Manca de consentiment informat	Existeix el risc que les recopilades per a l'entrenament de models d'IA no tinguin el consentiment informat del propietari de les dades. De vegades aquest ús és compatible amb la normativa vigent, però també pot ser poc ètic i generar riscos de manca de confiança ciutadana.	Mantenir estàndards ètics que assegurin el consentiment informat per a la recopilació de dades.
Atacs al sistema d'IA	Un sistema d'IA pot rebre diversos tipus d'atacs interns i externs mitjançant els quals es pot manipular, corrompre, introduir o extreure informació i dades emprades o generades pel sistema, cosa que pot generar impactes molt significatius i nocius, així com revelar i comprometre informació personal i confidencial. Hi ha diferents tipus d'atacs als sistemes d'IA: *data poisoning*, atac d'inferència de pertinença, atac d'interferència d'atributs, atac d'evasió, atac d'extracció, injecció de la indicació, filtració de la indicació, *prompt priming* i *jailbreaking*.	Enfortir la ciberseguretat del sistema per defensar-se contra atacs com enverinament de dades i atacs adversaris, d'acord amb les directrius i sota la coordinació de l'Agència de Ciberseguretat de Catalunya.

3. Desenvolupament del model

Fase iterativa que inclou la neteja de dades, enginyeria de característiques, disseny algorísmic en detall, prova i execució d'un conjunt de models per tractar de predir comportaments, així com el desenvolupament de documentació relacionada. En aquest sentit, un cop entrenat, el model ha d'avaluar-se per a veure'n el rendiment i poden dur-se a terme múltiples rondes de desenvolupament i refinament d'aquest. Els riscos d'aquesta etapa són:

Nom del risc	Descripció del risc	Mesures de control del risc
Errors de funcionament	A causa de diferents factors, un sistema d'IA generar contingut erroni que derivi en diferents impactes nocius. Per exemple, un llenguatge no adequat o amb prejudicis, informació incorrecta.	Fer proves i validacions extenses per detectar i corregir sortides perjudicials o incorrectes.

Al·lucinació	Generació de contingut fàcticament inexacte o fals, que poden induir a error les persones usuàries i incorporar-se a sistemes posteriors.	Implementar controls de validació i verificació de la informació generada per la IA per evitar la difusió de dades falses o incorrectes.
Manca de traçabilitat	Existeix el risc que el disseny no garanteixi que les activitats executades per o a través d'un bot de conversa o un model d'IA puguin rastrejar-se fins a un bot o un compte d'usuari específics, la qual cosa limita l'eficàcia de les activitats de registre i supervisió.	Implementar sistemes de registre i auditoria que permetin rastrejar les activitats i decisions de la IA.
No divulgació	No explicar que el contingut és generat per un model d'IA redueix la confiança i és enganyós, especialment si està fet intencionalment.	Implementar polítiques de transparència que exigeixin la divulgació de l'ús d'IA en la generació de continguts i decisions.
Atribució de fonts poc fiables	L'atribució de fonts és la capacitat del sistema d'IA per descriure a partir de quines dades d'entrenament ha generat els resultats. Atès que les tècniques actuals es basen en aproximacions, aquestes atribucions poden ser incorrectes. En aquest sentit, les explicacions de baixa qualitat dificulten que les persones usuàries, validadores de models i auditores entenguin i confiïn en el model.	Millorar els mètodes de validació de les dades per assegurar l'atribució fiable de les fonts utilitzades en l'entrenament de models.
Manca de transparència de les dades i del procés de desenvolupament del model	Cal disposar d'informació precisa sobre com s'han recopilat, conservat i usat les dades d'un model per entrenar-lo, així com de la resta de processos de desenvolupament. En cas contrari, podria ser més difícil explicar satisfactòriament el seu comportament i s'augmenta el risc d'un ús indegut del model. A més, és clau per al compliment normatiu, l'ètica de la IA i per guiar l'ús adequat dels models.	Documentar exhaustivament l'origen, la gestió i l'ús de les dades utilitzades per entrenar els models d'IA. Augmentar la documentació i explicabilitat dels models d'IA per facilitar la comprensió del seu funcionament i metodologia.
Manca de seguiment de la metodologia de disseny i desenvolupament	La metodologia de disseny i desenvolupament pot no estar configurada de manera consistent per garantir un bon lliurament d'un programa d'automatització d'IA. En aquest cas, s'haurà de seguir el procediment establert per evitar-ho.	Establir un protocol estàndard per al disseny i desenvolupament de models d'IA que asseguri la coherència i qualitat de les solucions.
Proves insuficients de la solució d'automatització	Les proves dels models poden no haver estat exhaustives i el resultat seria un sistema que no compleix amb els requisits i els objectius estratègics.	Augmentar la rigorositat i freqüència de les proves al llarg del cicle de desenvolupament d'IA.

4. Desplegament i consum del model

Fase d'implementació del model en un entorn de producció. Aquesta etapa pot implicar la integració del model amb els sistemes existents, la creació d'una aplicació o servei que usi el model o l'aprofitament de la informació a través d'un context sense connexió simultània, com un informe. Els riscos que presenta són:

Nom del risc	Descripció del risc	Mesures de control del risc
Deriva del model (*model drift*)	Reducció del rendiment del model per canvis en les dades o en les relacions entre les variables d'entrades i sortides, que pot conduir a una presa de decisions defectuosa i a prediccions incorrectes. En molts casos, sempre estan arribant nous punts de dades (noves variacions, nous patrons, noves tendències, etc.) que les dades històriques antigues no poden capturar. Pot aparèixer per diferents causes: deriva del concepte, deriva de la dada o canvis en la segmentació de dades.	Implementar el monitoratge continu per detectar i corregir la deriva del model.
Riscos en la posada en producció	Poden sorgir riscos a causa de la dificultat per complir amb els requisits de rendiment del model i de la capa del servei. Això es pot agreujar pel fet que la precisió de la producció sempre serà desigual de la precisió de l'equip de prova i encara divergirà mentre el model no s'actualitzi. En aquest sentit, s'haurà d'implementar una arquitectura de manteniment elaborada.	Abordar les discrepàncies de rendiment i actualitzar models regularment.
Errors en la implementació	Fallades que s'esdevenen quan el model desenvolupat i avaluat s'integra en un entorn operatiu real. Aquests errors comprometen la funcionalitat i seguretat del model, demanen proves exhaustives i monitoratge continu per a la seva identificació i correcció efectives.	Utilitzar protocols rigorosos de desplegament i proves per minimitzar errors durant la integració de models d'IA.
Dependència excessiva o insuficient	En les tasques en què les persones prenen decisions basades en suggeriments basats en la IA, la dependència excessiva o insuficient pot conduir a una mala presa de decisions a causa de la confiança inadequada en el sistema d'IA. Les conseqüències negatives poden augmentar amb la importància de la decisió.	Desenvolupar programes de capacitació i sensibilització en matèria d'IA que equilibrin la confiança i dependència en aquests sistemes.

5. Monitoratge del model

Fase de supervisió del model una vegada desplegat per poder introduir, provar, entrenar i implementar-ne de nous que permetin canviar estratègies i adaptar-se a altres entorns o necessitats. Sovint caldrà mantenir i actualitzar el model, la qual cosa pot implicar tornar a una fase anterior i s'ha de considerar una part normal del cicle de vida. Els seus riscos són:

Nom del risc	Descripció del risc	Mesures de control del risc
Manca de supervisió dels resultats dels models d'IA	Sense una supervisió adequada dels resultats del model, el sistema podria no comportar-se segons el previst quan es va dissenyar i implementar (supervisió dels resultats en relació amb els requisits originals de l'Administració, els requisits ètics, els valors corporatius, etc.) o tenir males respostes o temps de decisió.	Establir un marc de monitoratge continu que avaluï els resultats de la IA enfront dels estàndards i objectius originals.
Impactes d'un monitoratge inadequat en l'eficàcia general	Pot existir el risc que no s'apliquin els requisits d'auditabilitat o supervisió de les operacions d'IA, la qual cosa afecta la capacitat de supervisar l'eficàcia contínua de les operacions, així com l'anàlisi o la resposta oportuna a possibles incidents.	Desenvolupar un sistema integral de monitoratge i resposta per ajustar i millorar contínuament l'eficàcia dels sistemes d'IA.
Ineficiències en els mecanismes de monitoratge	Pot existir el risc que els mecanismes implementats per monitorar el funcionament i els resultats dels sistemes d'IA no siguin plenament efectius i no permetin detectar aquells riscos que no sigui han detectat en fases prèvies del cicle de vida del sistema.	Desenvolupar mecanismes efectius de monitoratge per assegurar que els models d'IA funcionin segons el previst i s'alineïn amb estàndards ètics i de negoci.