MOOC
Les fondamentaux de l'IA

Les risques de l'IA

Christelle GIBON

Explorons les risques de l'intelligence artificielle et leurs mécanismes.

L'intelligence artificielle n'est pas un simple objet technique. Elle se nourrit de nos données, s'imprègne de nos comportements et s'enrichit de notre expertise. Dans ce processus d'apprentissage, elle développe une capacité à simuler nos propres facultés cognitives, devenant ainsi un miroir de notre intelligence : elle manie le langage avec une aisance souvent déconcertante, elle est capable de dialoguer et d'écrire comme nous. Désormais, nous communiquons avec elles en langage naturel, aussi simplement qu'avec un autre humain.

Dans cette évolution, l'IA devient de plus en plus autonome, pouvant prendre en charge une part croissante de nos tâches quotidiennes à travers des agents IA de plus en plus sophistiqués.

Toutes ces caractéristiques font de l'IA un objet d'une nature particulière, dont les impacts sont multi-dimensionnels.

Les facteurs de risque inhérents à la technologie

Commençons par examiner les facteurs de risque inhérents à la technologie elle-même.

1. Les données

Tout d'abord, les données : pas d'IA sans Data !

Les données sont à la base du fonctionnement de l'IA mais constituent aussi son talon d'Achille.

Imaginez un instant que vous deviez prendre une décision importante en ne disposant que d'informations partielles ou inexactes. C'est exactement ce qui peut arriver à une IA entrainée avec des données biaisées, incomplètes ou obsolètes.

Prenons un exemple concret : une IA utilisée pour le recrutement. Si elle a été entraînée sur des données historiques reflétant d'anciennes inégalités, elle risque de reproduire ces discriminations en privilégiant certains profils au détriment d'autres.

2. La nature probabiliste des algorithmes

Un autre point clé à comprendre est que l'IA faite d'apprentissage machine, fonctionne sur un mode probabiliste. Contrairement à une calculatrice qui fournit toujours un résultat exact, une IA propose une réponse basée sur des probabilités et une marge d'erreur inhérente.

En d'autres termes, utiliser une IA pour prendre des décisions, c'est raisonner en termes de statistiques et de probabilités. Cela signifie qu'il peut y avoir des erreurs, des approximations… et parfois même des résultats inattendus ou absurdes : les fameuses hallucinations !

3. L'opacité des systèmes

Un autre enjeu de taille est ce qu'on appelle "l'effet boîte noire".

Les mécanismes et raisonnements de certains algorithmes d'IA, notamment les plus sophistiqués, sont difficiles, voire impossibles à expliquer clairement.

Cette opacité pose un problème particulier dans le service public, où la transparence est un principe à valeur constitutionnelle : chaque citoyen doit comprendre les décisions prises par l'administration pour pouvoir les contester le cas échéant.

4. Le piège de la surconfiance

Un autre facteur de risque, et non des moindres, est l'illusion d'infaillibilité.

L'IA donne des réponses cohérentes, vraisemblables que l'on peut avoir tendance à considérer comme nécessairement justes.

Or, on l'a vu, une IA peut se tromper.

Cette surconfiance peut éroder notre sens critique. Même dans les systèmes prévoyant une supervision humaine, la tentation devient forte de s'effacer et de privilégier la réponse de la machine au détriment du jugement humain.

Quels risques découlent de ces mécanismes et peuvent concerner directement les collectivités territoriales ?

1. Atteinte aux droits fondamentaux et à nos libertés individuelles

L'IA peut potentiellement affecter un large éventail de droits fondamentaux.

- **Respect de la vie privée**

Pour fonctionner, de nombreux systèmes d'IA nécessitent une collecte massive de données personnelles. Cela pose une question cruciale : comment garantir la confidentialité et protéger ces informations sensibles ?

Par exemple : un assistant virtuel qui collecte en continu des informations sur les citoyens pour optimiser un service public. A qui ces données sont-elles accessibles ? Et comment sont-elles utilisées ?

La CNIL (gendarme français en la matière) travaille actuellement sur des cadres de protection adaptés à l'IA pour s'assurer que les droits des citoyens sont respectés.

- **Biais et discriminations**

Un autre risque majeur est la reproduction et l'amplification des préjugés existants par les systèmes d'IA.

Pourquoi ? Parce que l'IA apprend à partir des données qu'on lui fournit.

Si ces données sont biaisées, les décisions de l'IA le seront aussi !

Par exemple : une IA utilisée pour analyser les dossiers de demande d'aide sociale. Si les données d'entraînement montrent qu'historiquement, certains profils obtiennent plus souvent des refus, l'algorithme risque de perpétuer ces inégalités. Cela peut exclure involontairement des citoyens de leurs droits légitimes, ce qui est contraire au principe d'égalité de traitement que porte l'administration publique.

- **Manque de transparence et d'explicabilité**

Un autre problème central avec l'IA, c'est son opacité.

Comment expliquer une décision prise par un algorithme et la contester… si personne ne sait vraiment comment il fonctionne ?

Imaginez une IA qui décide de l'attribution d'une aide sociale. Un citoyen se voit refuser son aide, il demande des explications. Mais l'algorithme fonctionne comme une boîte noire : on ne sait pas précisément quels critères ont pesé dans la décision.

Assurer la transparence et l'explicabilité des algorithmes est essentielle dans le service public.

2. Manipulation de l'information et désinformation

Autre défi : celui de la manipulation de l'information et la désinformation.

Avec l'essor des IA génératives, il est devenu extrêmement facile de produire des contenus trompeurs avec un réalisme saisissant : textes imitant des articles officiels, images et vidéos truquées (les *deepfakes*), synthèses vocales imitant des voix réelles…

Imaginez par exemple la diffusion de fausses informations concernant des services publics locaux : les conséquences peuvent être désastreuses y compris pour la confiance des citoyens et l'image de la collectivité.

3. Sécurité, cybersécurité et détournement des IA

L'intelligence artificielle augmente la surface d'attaque des infrastructures informatiques parce qu'elle peut être connectée à de nombreuses bases de données internes et interagir avec des applications métiers.

Par ailleurs, certaines IA peuvent être détournées à des fins malveillantes. Par exemple, des IA capables de générer du texte peuvent être détournées pour créer des e-mails d'hameçonnage ultra-réalistes ciblant les collectivités territoriales. Elles sont devenues d'ailleurs des cibles majeures.

4. Enjeu de souveraineté

Pour les collectivités territoriales, un défi fondamental se dessine : garder la maîtrise de leurs outils et de leurs données pour préserver expertise et capacité d'action.

Or, le paysage actuel de l'IA est dominé par une poignée de géants technologiques non européens. Ils concentrent entre leurs mains le contrôle des données, des algorithmes et des infrastructures critiques.

Cette situation fait peser des risques majeurs sur nos collectivités territoriales :

- Une dépendance technologique croissante envers des acteurs étrangers
- L'exposition de données sensibles à des lois extraterritoriales
- L'impossibilité d'exercer un véritable contrôle sur des algorithmes opaques dont ni le fonctionnement ni les données ne peuvent être pleinement audités.

Face à ces enjeux, certaines administrations s'engagent dans le développement de solution locales et souveraines.

5. Impact Environnemental

L'IA a une empreinte écologique importante, souvent sous-estimée.

Pourquoi ? Les data centers qui fournissent la puissance de calcul nécessaire à l'entrainement et à l'utilisation de ces IA consomment énormément d'électricité. Le refroidissement de leurs serveurs demande des quantités massives d'eau. Et l'exploitation de ressources rares est nécessaire pour fabriquer le matériel.

Selon l'Agence internationale de l'énergie (AIE), les centres de données associés à l'IA ont consommé quasiment 2 % de la production mondiale, chiffre qui va continuer à augmenter et qui pose déjà des problèmes localement sur le réseau électrique comme en Irlande où les datacenters représentent déjà 20 % de la demande d'électricité.

Certaines approches dites d'IA frugale se mettent en place pour prendre en compte ce coût environnemental dans le choix de recourir ou non à de tels systèmes.

6. Impact sur l'organisation du travail

L'arrivée de l'IA générative a relancé le débat sur les risques de disparition du travail liés à l'automatisation. Des études annoncent des disparitions massives d'emplois, d'autres des créations nettes d'emploi. Ce que ces chiffres ne doivent pas masquer est que l'IA transforme déjà profondément nos environnements professionnels. L'automatisation, la délégation à la machine d'une partie de nos tâches présente parfois un potentiel d'amélioration mais aussi un certain nombre de risques : déqualification, intensification du travail, perte d'autonomie, déresponsabilisation, évolution du rôle des managers, partage du travail et d'autres.

Maintenant que nous avons identifié ces risques majeurs, examinons comment ils peuvent être maitrisés.

Le cadre règlementaire existant

Plusieurs régulations européennes et françaises existent déjà pour encadrer l'usage de la donnée et l'IA comme :

- Le Règlement pour la protection des données personnelles (ou RGPD) qui protège les données personnelles des citoyens

- La loi pour une République numérique (LRN) qui modifie le code des relations entre le public et l'administration (CRPA) afin de faciliter l'accès aux données publiques et renforcer la transparence des algorithmes utilisés dans l'administration.

En Europe, en France et dans les collectivités, un cadre règlementaire mais aussi des initiatives se mettent en place pour encadrer l'usage de l'IA, réduire les risques et développer une IA qui bénéficie à toutes et tous.

Le nouveau cadre européen

Le Règlement sur l'Intelligence Artificielle, entré en vigueur depuis août 2024, se fonde sur les risques. Ses premières obligations sont applicables depuis février 2025 : sont dorénavant interdits certains systèmes d'IA jugés à risque inacceptable comme la manipulation comportementale, la police prédictive, la reconnaissance des émotions sur le lieu de travail et dans les établissements d'enseignement, ainsi que la notation sociale.

Les initiatives locales

Au-delà de ce cadre règlementaire qui s'impose à tous, un certain nombre de collectivités territoriales travaillent à développer leurs propres cadres d'usage en associant les agents et parfois les citoyens. Ils produisent des guides de bonnes pratiques, des chartes internes, des doctrines voire des stratégies qui comblent les lacunes règlementaires et permettent de prendre en considération les spécificités locales.

À Montpellier Métropole, par exemple, une charte éthique de l'IA a été co-construite avec les citoyens, intégrant leurs préoccupations sur la transparence et l'équité des algorithmes.

Les leviers d'action

En conclusion, voyons quelques leviers d'actions concrets pour anticiper et maitriser les risques liés à l'IA.

- **1. La gouvernance des données**

La qualité des systèmes d'IA repose avant tout sur les données qui les nourrissent. Une IA ne peut être plus fiable que les données sur lesquelles elle s'appuie.

Travailler à la structuration et la mise en qualité des données est donc un prérequis.

- **2. L'évaluation**

Avant de céder à l'attrait de l'IA, une démarche réflexive s'impose. L'IA est-elle la réponse adaptée ? N'existe-t-il pas des alternatives plus simples, plus frugales ? Parfois la solution première est celle de l'optimisation des processus et non l'ajout d'une nouvelle couche technologique.

Evaluer l'ensemble des impacts - sociaux, éthiques, environnementaux - en amont permet de décider de recourir ou non à ces systèmes de façon éclairée et de fixer des critères d'acceptabilité.

- **3. L'humain au centre**

L'IA doit demeurer un outil au service de l'humain, et non l'inverse, ce qui implique :

  - De concevoir des systèmes au service des capacités humaines, préservant l'autonomie et l'esprit critique
  - D'impliquer les utilisateurs finaux dès la conception
  - De maintenir un contrôle humain sur les systèmes et de prévoir un droit de retour en arrière

- **4. La formation et la montée en compétences de toutes et tous**

Une montée en compétence collective est nécessaire pour utiliser l'IA de façon à minimiser les risques. Cela implique :

  - Le développement d'une véritable culture de l'IA dans les organisations
  - La formation aux enjeux et aux bonnes pratiques
  - Et une veille active sur les évolutions technologiques et réglementaires

- **5. La transparence et l'explicabilité : le socle de la confiance**

Enfin, la transparence doit être une ligne directrice pour développer l'IA dans l'action publique. Usagers et agents doivent pouvoir comprendre :

- Les mécanismes qui sous-tendent les systèmes d'IA
- La nature et la finalité des données utilisées
- Les processus décisionnels mis en œuvre

Cette transparence n'est pas une simple obligation : elle est le fondement de notre confiance dans ces systèmes.

C'est sur ces fondations, alliant cadre réglementaire et bonnes pratiques locales, que nous pourrons bâtir une IA digne de confiance qui serve réellement l'intérêt général.