Utiliser un système d'IA en production

05 avril 2022

Garantir la qualité et la transparence du système au cours de son utilisation.

Permettre à l’humain de garder la main

Si les systèmes d'IA peuvent permettre d'automatiser des tâches et être parfois plus performants qu'un opérateur humain, **une personne doit toutefois être mobilisée lorsque cela est nécessaire** pour empêcher les détournements et la systématisation de certaines erreurs. Les mesures prises dans ce but doivent être identifiées, effectives et pérennes.

Les questions suivantes pourront aider le responsable de traitement à mettre en œuvre le cadre favorable au respect de ces conditions.

☐ Une supervision par un opérateur humain est-elle prévue ?

☐ Des mécanismes ont-ils été prévus pour que l'opérateur puisse modifier manuellement une décision prise par le système d'IA (par ex. : modifier le profil donné à l’utilisateur d’une plateforme de partage de vidéos) ou en arrêter le fonctionnement (par ex. : couper l’accès d’un *chatbot* en ligne dont l’apprentissage automatique aurait conduit à une dérive du système) ?

Lesquels ?

Quel type d'intervention est prévue ?

☐ *Human-in-the-Loop* (capacité d'intervention humaine dans chaque cycle de décision du système)

☐ *Human-on-the-Loop* (capacité d'intervention humaine pendant le cycle de conception du système et la surveillance du fonctionnement du système)

☐ *Human-in-Command* (capacité de superviser l'activité globale du système d'IA et de décider quand et comment utiliser le système d'IA dans une situation donnée)

☐ Ces mécanismes font-ils l'objet de protocoles clairement formulés et connus de tous ?

☐ Sont-ils intégrés de manière naturelle dans le traitement ?

☐ Les ressources humaines et matérielles nécessaires ont-elles été prévues ?

☐ Les personnes en charge de cette tâche sont-elles clairement identifiées ?

☐ Sont-elles en mesure d'exercer leur contrôle sur le système d'IA facilement ?

☐ Les personnes ont-t-elles reçu une formation suffisante ?

☐ L’information qui leur est fournie lors de la supervision est-elle suffisante ?

☐ Les outils permettant le contrôle par l'opérateur ont-ils été suffisamment éprouvés ?

☐ Une identification des cas où l'intervention humaine est nécessaire a-t-elle été mise en place (via le calcul d'un indicateur de confiance par exemple) ?

☐ Une étude de la proportion de cas soumis à l’humain et de l’efficacité de la supervision est-elle prévue ?

Comment le risque lié au biais d'automatisation (tendance d’un opérateur à accorder une confiance trop importante à un procédé automatisé) a-t-il été pris en

compte ?

☐ Des mécanismes visant à compenser ce biais ont-ils été mis en œuvre ?

Lesquels ?

Mettre en œuvre la transparence pour assurer la confiance

Afin d'instaurer la confiance entre les personnes concernées et le responsable du système d'IA, **le plus haut niveau de transparence doit être mis en place**, tant pour expliquer le fonctionnement du système lui-même que pour expliquer les décisions individuelles. La journalisation permettant de faciliter l'explicabilité ne doit en revanche pas se faire au détriment du respect de la vie privée des opérateurs ou des personnes concernées.

☐ Une journalisation des éléments (données utilisées pour l'inférence, indicateurs de confiance, version du système, etc.) servant à la prise de décision par le système d'IA existe-t-elle ?

☐ Les éléments journalisés permettent-ils d'expliquer, à posteriori, une décision particulière prise par le système d'IA ?

☐ Ces informations sont-elles conservées pour une durée justifiée ?

☐ Sont-elles limitées aux strictes catégories nécessaires à l'explication de la décision ?

☐ Le fonctionnement du système d'IA est-il expliqué aux personnes amenées à interagir avec elle ?

☐ L'explication est-elle rendue suffisamment claire et compréhensible, grâce à des cas concrets par exemple, et en expliquant les limitations, hypothèses et cas extrêmes du système ? (par ex. un robot industriel assistant un humain dans sa tâche saura perforer mais pas visser)

☐ Des outils techniques et méthodologiques sont-ils utilisés pour permettre l'explicabilité du système ?

☐ Le code est-il ouvert (*open source*) ?

Assurer la qualité du traitement

En dehors des mesures de transparence et de supervision mises en place, des mesures techniques doivent également assurer que la qualité des sorties du système d'IA est maintenue au cours de sa durée de vie.

☐ Une analyse automatique des logs de journalisation existe-t-elle afin d'alerter l'utilisateur et/ou la personne en cas de défaillance, de fonctionnement anormal ou d'attaque ?

☐ Un contrôle de la qualité et de la correspondance des données collectées en environnement réel avec les données d'apprentissage et de validation est-il maintenu au cours de l'utilisation du système d'IA ?

☐ La qualité des sorties du système d'IA est-elle contrôlée au cours du cycle de vie du système d'IA ?

Les risques spécifiques

Certains risques spécifiques à certains systèmes d'IA demandent une attention particulière. Les exemples ci-dessous permettent d'identifier certains enjeux.

- [ ] Dans le cas d'un apprentissage en continu, la qualité des données utilisées pour l'entraînement est-elle contrôlée tout au long du cycle de vie du système ?
- [ ] Dans le cas d'un apprentissage fédéré, des mesures sont-elles prises pour lutter contre le caractère non-indépendant et identiquement distribué des données ? (par ex. : apprentissage fédéré d'un modèle de reconnaissance vocale sur des assistants vocaux)
- [ ] Dans le cas où un système d'IA est utilisé pour collecter des données personnelles par exploration des données ou *data mining*, une vérification permet-elle de s'assurer que seules les données des personnes concernées sont collectées ?
- [ ] Une vérification permet-elle de s'assurer de leur intégrité et de leur véracité (par la vérification de l'authenticité des sources, et de la qualité de l'extraction des données par exemple) ?
- [ ] Dans le cas d'un algorithme de recommandation de contenu, les suggestions pourraient-elles influencer les opinions des personnes (politiques par exemple) ?
- [ ] Pourraient-elles aller à l'encontre de l'intérêt de certaines personnes (morales ou physiques) ?

Imprimer cette check-list

Aucune information n'est collectée par la CNIL.

< Fiche précédente : développer et entraîner un algorithme

Sommaire

Fiche suivante : sécuriser le traitement >

Vous souhaitez contribuer ?

Écrivez à ia[@]cnil.fr

Haut de page