Google

# Indications pour un machine learning responsable

## Indications

### Agissez avec prudence lorsque les flux de contrôle sont modifiés en fonction des résultats des modèles de ML.

Il est possible de modifier les flux de travail fondés sur les données de sorties issues des valeurs primitives du machine learning pour créer des angles d'attaque difficiles à sécuriser. De même, il n'est pas recommandé d'inclure des données dans un flux de travail d'exécution qui n'est pas expressément prévu.

### Évitez de donner un accès direct aux modèles et aux métadonnées des modèles.

Les interactions directes avec les valeurs primitives du machine learning peuvent générer différents angles d'attaque : exfiltration, inversion, etc. Vous devez maintenir ces éléments à distance des utilisateurs, permettre uniquement des interactions contrôlées et approuvées et veiller à ce que chaque donnée de sortie soit associée à une fonctionnalité établie, comme une base de données. Ne fournissez aucune donnée de sortie directement aux utilisateurs.

### Veillez à ce que les modèles vérifient l'intégrité lors de l'entraînement ainsi qu'après la mise en œuvre.

Les logiciels peuvent être modifiés à de nombreux stades de leur développement et de leur déploiement. C'est aussi valable pour le machine learning. Vérifiez encore et encore. Exécutez uniquement les fonctions que vous savez prévues : l'authentification avant l'exploitation.

### Des fonctionnalités complexes peuvent générer des contrôles complexes.

Les interactions doivent être brèves, précises et succinctes. Les flux de travail longs et complexes augmentent les risques d'exposition à des attaques réussies.

### Utilisez d'abord la logique extensionnelle pour l'entraînement puis passez à la logique intentionnelle pour les fonctionnalités et l'inférence, selon les besoins.

Le recours à un ensemble approuvé et de qualité pour l'entraînement est un bon moyen d'éviter les attaques par empoisonnement, entre autres. Néanmoins, vous avez besoin de données supplémentaires pour augmenter les capacités d'inférence de vos fonctionnalités. En testant soigneusement les résultats de cet entraînement à partir de résultats attendus issus de données d'entrée de qualité, vous obtenez des systèmes plus résilients.

### Veillez à créer des tests d'intégration pour les cas d'utilisation prévus des valeurs primitives.

Comme avec tout autre logiciel, les flux de travail du machine learning doivent être soumis à des tests d'intégration et des tests unitaires. Compte tenu de la fonction heuristique du ML, il peut être difficile de formuler des affirmations fiables. Il est donc recommandé de se concentrer sur les résultats souhaités à long terme pour le flux de travail.

### Repérez les attaques qui s'appuient sur la synthèse automatique, en particulier pour les commandes ayant rapport à l'identité.

Posez-vous systématiquement la question « est-ce qu'une machine sait faire ça actuellement ? » plutôt que de supposer l'authenticité d'une voix, d'une vidéo ou d'un texte, surtout lorsqu'ils servent à accorder des droits d'accès.

### Utilisez des politiques mutuelles pour garder une longueur d'avance sur le modèle économique et garantir un modèle de sécurité robuste à moindre coût.

Si le coût d'une attaque est inférieur à celui des mesures nécessaires pour y remédier, alors les mesures de contrôle seront probablement inefficaces. Faites en sorte que les coûts de sécurité s'ajustent automatiquement en fonction des répercussions des attaques réussies afin d'augmenter progressivement les coûts pour les attaquants.

Google

**Évitez de mettre des modèles entraînés à partir de données propriétaires dans les mains d'un utilisateur, par exemple, sur un appareil.**

Analysez les cas d'utilisation mobilisés pour la création des valeurs primitives du machine learning et supposez qu'en théorie, il est possible de récupérer les données d'entraînement originelles. Utilisez ces informations pour vos plans de continuité de l'activité et les évaluations des risques dans le cadre du développement des produits.

**Utilisez la limitation du débit pour éviter les attaques par extrapolation et prévenir la consommation inutile de ressources.**

Les contrôles traditionnels sont encore efficaces dans l'ensemble. Superposez les niveaux de contrôle avec, par exemple, les listes de contrôle des accès, la limitation du débit et une authentification lors de la conception et de la mise en œuvre des technologies de machine learning.