Google

Lignes directrices pour un apprentissage automatique responsable

Lignes directrices

Faites preuve de prudence lorsque les flux de contrôle sont modifiés sur la base des résultats des modèles d’apprentissage automatique.

Les flux de travail qui exploitent les résultats d'une primitive d'apprentissage automatique peuvent être modifiés pour produire des surfaces d'attaque difficiles à sécuriser. Il n'est pas recommandé d'intégrer des données dans un flux de travail d'exécution qui n'est pas explicitement prévu.

Évitez d’accéder directement aux modèles et aux métadonnées de ces modèles.

Toute interaction directe avec les primitives d'apprentissage automatique peut générer un certain nombre de surfaces d'attaque : exfiltration, inversion, etc. Maintenez ces objets à l'écart des utilisateurs, n’autorisez que des interactions contrôlées et approuvées et veillez à ce que les résultats soient associés à une ressource existante, comme une base de données.
Ne fournissez pas les résultats directement à l’utilisateur.

Assurez-vous que les modèles incluent une fonction de contrôle d’intégrité à la phase d’entraînement et après leur mise en œuvre.

Les logiciels peuvent être soumis à des modifications à différentes étapes du processus de développement et de déploiement; il en va de même pour l'apprentissage automatique. Vérifiez plutôt deux fois qu’une. N’exécutez que ce qui est prévu : authentifiez-vous avant toute opération.

La complexité des fonctionnalités peut rendre les contrôles difficiles.

Les interactions doivent être brèves, directes et succinctes. Les flux de travail étendus et complexes présentent un risque d'exposition aux attaques bien plus élevé.

Commencez par appliquer une logique extensionnelle au cours de la phase d’entraînement, puis passez à une logique intentionnelle pour tout ce qui a trait aux fonctions et à l’inférence, si nécessaire.

L'utilisation d'un ensemble de données vérifiées et connues pour le processus d'entraînement est un bon moyen d'éviter l'empoisonnement de données et d'autres attaques. Toutefois, pour créer des fonctions dotées de plus grandes capacités d'inférence, de nouvelles données doivent être ingérées; testez soigneusement les résultats de l’entraînement en vous basant sur les résultats anticipés à partir de bons intrants afin d'améliorer la résilience des systèmes.

Veillez à ce que les tests d'intégration soient créés en fonction du cas d'utilisation prévu de la primitive.

Les flux de travail d’apprentissage automatique nécessitent des tests d’intégration et des essais unitaires comme tout autre logiciel. La nature heuristique des logiciels d’apprentissage automatique peut nuire à l'élaboration d'assertions fiables; il est donc recommandé de se centrer sur les objectifs de résultats à long terme.

Soyez à l'affût des attaques qui utilisent la synthèse automatique, en particulier celles qui menacent les contrôles visant à vous authentifier.

Demandez-vous toujours si une machine est capable de faire une chose avant de présumer de la fiabilité de la voix, de la vidéo et du texte; surtout lorsque ceux-ci sont utilisés pour demander les droits d'accès.

Utilisez des politiques communes pour garder une longueur d’avance sur le modèle économique et assurer un modèle de sécurité solide et peu coûteux.

Si le coût d'une attaque est inférieur au coût des mesures correctives qui en découlent, le contrôle est probablement inefficace. Ajustez les coûts de sécurité automatiquement selon l’impact des attaques fructueuses, en augmentant progressivement les coûts imposés à l'auteur de l'attaque.

Sécurité accrue avec Google

Google

Évitez de confier à l'utilisateur le contrôle des modèles entraînés à partir de données exclusives, par exemple sur un appareil.

Tenez compte du cas d'utilisation impliqué dans la création de la primitive d'apprentissage automatique, et partez du principe général que les données d'entraînement qui ont permis de la créer peuvent être récupérées. Adoptez la même approche pour planifier la continuité des activités et évaluer les risques lors du développement de produits.

La limitation du taux permet d'éviter les attaques par extrapolation et de prévenir le gaspillage des ressources.

Les outils de contrôle traditionnels, comme les listes de contrôle d'accès, la limitation du taux et l'authentification, restent efficaces; combinez-les lors de la conception et de la mise en œuvre des technologies d'apprentissage automatique.