Un framework sécurisé basé sur l'IAG (Intelligence Artificielle Générative)

Co-directeurs de la thèse :
Samiha Ayed (UTT)
Lamia Chaari Fourati (Université de Sfax)

Sujet :

L'intelligence artificielle a changé ces dernières années les tendances dans les approches solutionnant des problématiques de cybersécurité. Plus récemment, l'intelligence artificielle générative (IAG) a encore renforcée le pouvoir de l'IA dans les différents domaines d'usage. Dans le domaine de la cybersécurité, l'usage de l'IA générative peut facilement renforcer l'usage du deepfake et le clonage vocal. Pour mener des attaques à large échelle, ces types d'offensives se servent beaucoup de la faiblesse humaine au lieu des failles dans un environnement informatique (réseaux ou systèmes d'information). L'usage de ces techniques par les attaquants peut être exploité dans des campagnes de manipulation sociale et de désinformation. Les cyber-attaques peuvent de plus en plus se basées sur l'IA générative pour la création d'attaques plus sophistiquées comme pour la génération de faux emails, de fausses vidéos (deepfakes) ou aussi de fausses données. De ce fait la détection de ces attaques devient de plus en plus difficile. En face de ces nouvelles menaces, il est nécessaire de proposer des nouvelles contremesures qui sont capables de détecter ces approches. Dans cette thèse, nous nous focalisons sur le lien entre les attaques adverses avec des modèles de l'intelligence artificielle générative. Nous proposerons des modèles génératifs pour détecter des schémas de comportement suspects ou pour générer des données de test plus réalistes afin d'évaluer la résilience des systèmes de sécurité. Pour la mise en pratique des solutions qui seront proposées, nous utiliserons les modèles d'IAG pour générer des données d'entraînement supplémentaires afin d'améliorer les modèles de détection des menaces. Cela peut aider à renforcer la robustesse des systèmes de cybersécurité en exposant les modèles à un large éventail de scénarios possibles.

Les étapes de cette thèse sont comme suit :

1- Une étude approfondie de l'état de l'art sur les techniques d'attaques qui se basent sur l'IAG (comme par exemple les attaques sur la reconnaissance faciale, les attaques adverses sur les modèles d'IA, les attaques de phishing améliorées, la génération de malwares, etc.).

2- La proposition d'une taxonomie pour les différents modèles basés sur l'IAG pour contrer les cyber-attaques basées sur des modèles génératifs. Cela peut inclure l'utilisation de techniques d'analyse comportementale, de détection de deepfakes et de modèles de machine learning spécialisés pour identifier des schémas inhabituels.

3- Proposition d'un modèle basé sur l'IAG qui permet de détecter et de réagir aux attaques adverses utilisées principalement pour manipuler délibérément les systèmes d'intelligence artificielle (IA) en modifiant ou en introduisant des données d'entrée de manière à induire des erreurs ou des comportements indésirables dans les résultats générés par ces systèmes.

4- Proposition d'un méta-modèle générique en se basant sur une classification définie pour les différents types d'attaques génératives.

5- Validation du modèle proposé avec un use-case d'un scénario réel en se basant sur une dataset respectant les besoins du modèle.

Les résultats de chaque partie de cette thèse doivent être publiés dans des articles de journaux ou de conférences internationales de bonne qualité.