L’équipe rouge de Microsoft AI construit l’avenir d’une IA plus sûr
Microsoft Security Blog

Ram Shankar Siva Kumar, gestionnaire principal de programme :

Un élément essentiel d’un logiciel d’expédition sécurisé est le red teaming.
Il fait généralement référence à la pratique consistant à émuler des adversaires du monde réel et à leurs outils, tactiques et procédures pour identifier les risques, découvrir les angles morts, valider les hypothèses et améliorer la posture de sécurité globale des systèmes.
Microsoft a une riche histoire de red teaming sur les technologies émergentes dans le but d’identifier de manière proactive les défaillances de la technologie.
Alors que les systèmes d’IA devenaient de plus en plus répandus, Microsoft a créé en 2018 l’AI Red Team : un groupe d’experts interdisciplinaires qui se consacrent à penser comme des attaquants et à sonder les systèmes d’IA à la recherche de défaillances.

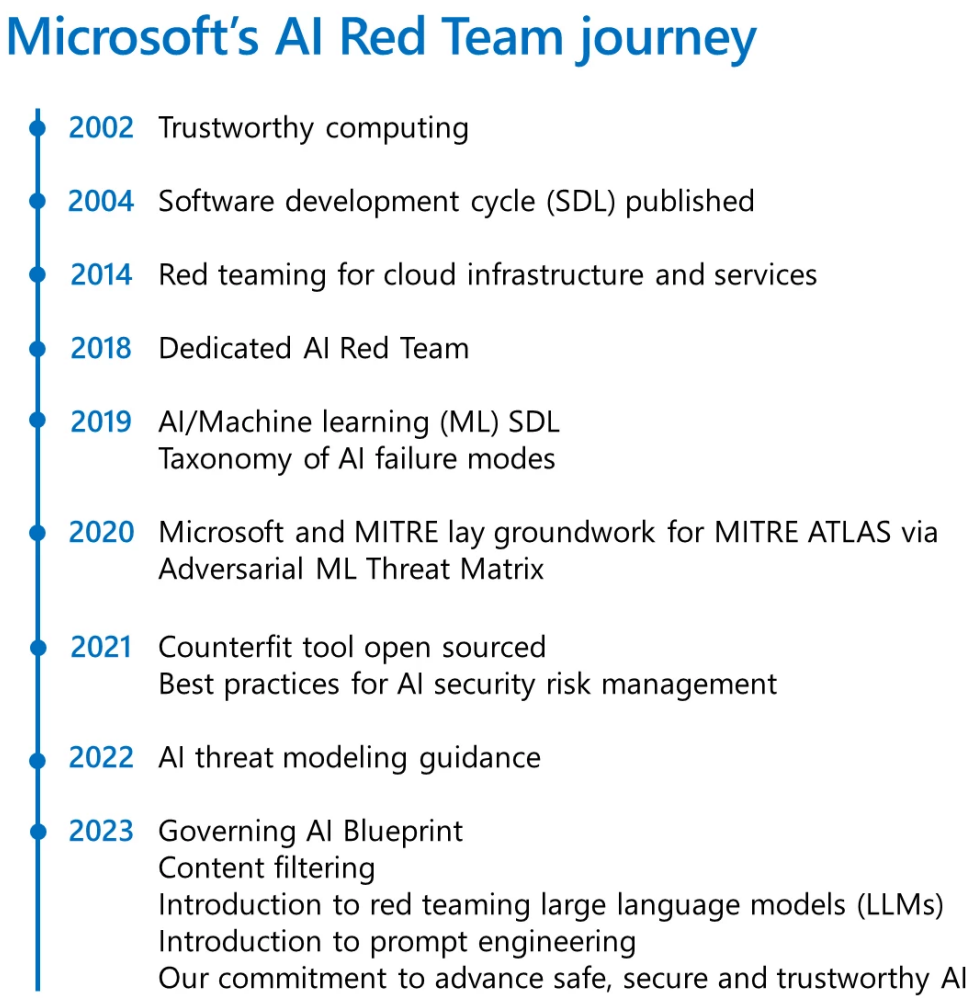
Nous partageons les meilleures pratiques de notre équipe afin que d’autres puissent bénéficier des enseignements de Microsoft.
Ces bonnes pratiques peuvent aider les équipes de sécurité à rechercher de manière proactive les défaillances des systèmes d’IA, à définir une approche de défense en profondeur et à créer un plan pour faire évoluer et développer votre posture de sécurité au fur et à mesure de l’évolution des systèmes d’IA générative.

La pratique du red teaming de l'IA a évolué pour prendre un sens plus large : elle couvre non seulement la recherche de failles de sécurité, mais inclut également la recherche d'autres défaillances du système, telles que la génération de contenu potentiellement dangereux.
Les systèmes d'IA comportent de nouveaux risques, et le red teaming est essentiel pour comprendre ces nouveaux risques, tels que l'injection rapide et la production de contenu non fondé.
L'équipe rouge de l'IA n'est pas seulement un atout chez Microsoft ; il s'agit d'une pierre angulaire de l'IA responsable dès la conception : comme l'a annoncé Brad Smith, président et vice-président de Microsoft, Microsoft s'est récemment engagé à ce que tous les systèmes d'IA à haut risque soient soumis à un red teaming indépendant avant d'être déployés.

L'objectif de ce blog est de contextualiser pour les professionnels de la sécurité comment le red teaming de l'IA se recoupe avec le red teaming traditionnel, et en quoi il diffère. Nous espérons que cela permettra à un plus grand nombre d'organisations de mettre en place leurs propres systèmes d'IA et de mieux tirer parti de leurs équipes rouges traditionnelles et de leurs équipes d'IA existantes.

Le Red Teaming contribue à rendre la mise en œuvre de l'IA plus sûre

Au cours des dernières années, l'équipe rouge de l'IA de Microsoft n'a cessé de créer et de partager du contenu pour permettre aux professionnels de la sécurité de réfléchir de manière globale et proactive à la manière de mettre en œuvre l'IA en toute sécurité.
En octobre 2020, Microsoft a collaboré avec MITRE ainsi qu'avec des partenaires industriels et universitaires pour développer et publier Adversarial Machine Learning Threat Matrix, un cadre permettant aux analystes de sécurité de détecter, de répondre et de remédier aux menaces.
Toujours en 2020, nous avons créé et mis en open source Microsoft Counterfit, un outil d'automatisation pour tester la sécurité des systèmes d'IA afin d'aider l'ensemble de l'industrie à améliorer la sécurité des solutions d'IA.
Par la suite, nous avons publié le cadre d'évaluation des risques liés à la sécurité de l'IA en 2021 pour aider les organisations à faire évoluer leurs pratiques de sécurité autour de la sécurité des systèmes d'IA, en plus de mettre à jour Counterfit.
Plus tôt cette année, nous avons annoncé des collaborations supplémentaires avec des partenaires clés pour aider les organisations à comprendre les risques associés aux systèmes d'IA afin qu'elles puissent les utiliser en toute sécurité, y compris l'intégration de Counterfit dans les outils MITRE et des collaborations avec Hugging Face sur un scanner de sécurité spécifique à l'IA disponible sur GitHub.



L'équipe rouge de l'IA liée à la sécurité fait partie d'un effort plus large d'équipe rouge de l'IA responsable (RAI) qui se concentre sur les principes d'équité, de fiabilité et de sûreté, de confidentialité et de sécurité, d'inclusion, de transparence et de responsabilité de Microsoft en matière d'IA.

Le travail collectif a eu un impact direct sur la façon dont nous expédions les produits d'IA à nos clients.

Par exemple, avant la sortie de la nouvelle expérience de chat Bing, une équipe de dizaines d'experts en sécurité et en IA responsable de l'entreprise a passé des centaines d'heures à rechercher de nouveaux risques de sécurité et d'IA responsable.

Cela *s'ajoutait* aux pratiques régulières et intensives de sécurité logicielle suivies par l'équipe, ainsi qu'à l'équipe rouge du modèle GPT-4 de base par les experts de la RAI avant le développement de Bing Chat.

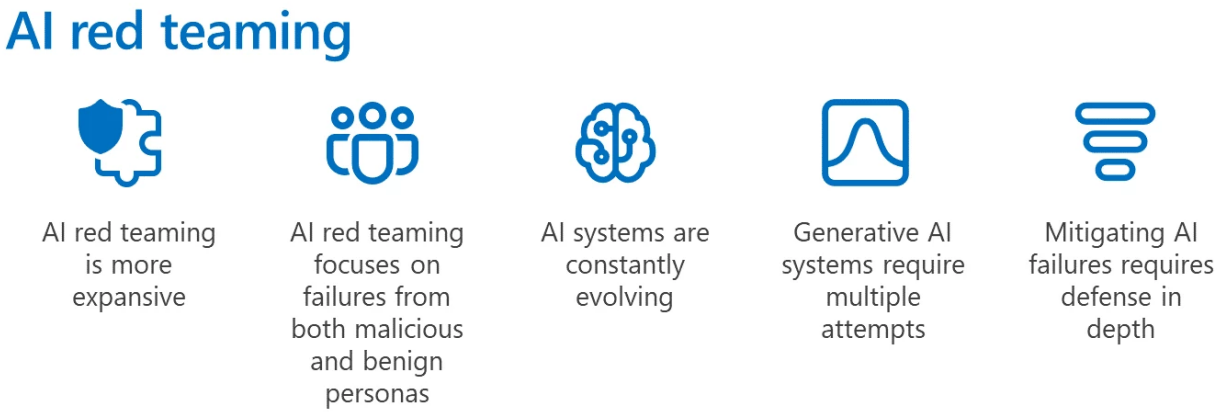
Les résultats de notre équipe rouge ont permis de mesurer systématiquement ces risques et d'élaborer des mesures d'atténuation de portée avant l'expédition du produit.

Conseils et ressources pour le red teaming

Le red teaming de l'IA a généralement lieu à deux niveaux : au niveau du modèle de base (par exemple, GPT-4) ou au niveau de l'application (par exemple, Security Copilot, qui utilise GPT-4 dans le back-end).

Les deux niveaux apportent leurs propres avantages : par exemple, le red teaming du modèle permet d'identifier dès le début du processus comment les modèles peuvent être utilisés à mauvais escient, d'évaluer les capacités du modèle et de comprendre les limites du modèle.

Ces informations peuvent être intégrées au processus de développement du modèle afin d'améliorer les futures versions du modèle, mais aussi de donner un coup de pouce aux applications pour lesquelles il est le plus adapté.
Le red teaming de l'IA au niveau de l'application adopte une vue système, dont le modèle de base est une partie.
Par exemple, lors de l'utilisation de Bing Chat par l'IA, l'ensemble de l'expérience de recherche alimentée par GPT-4 était dans le champ d'application et a été sondé à la recherche d'échecs.
Cela permet d'identifier les défaillances au-delà des mécanismes de sécurité au niveau du modèle, en incluant les déclencheurs de sécurité spécifiques à l'application.



Ensemble, l'analyse des risques liés à la sécurité et à l'IA responsable fournit un instantané unique de la façon dont les menaces et même l'utilisation bénigne du système peuvent compromettre l'intégrité, la confidentialité, la disponibilité et la responsabilité des systèmes d'IA.
Cette vision combinée de la sécurité et de l'IA responsable fournit des informations précieuses non seulement pour identifier les problèmes de manière proactive, mais aussi pour comprendre leur prévalence dans le système grâce à des mesures et éclairer les stratégies d'atténuation.
Vous trouverez ci-dessous les principaux enseignements qui ont contribué à façonner le programme AI Red Team de Microsoft.

1. **Le red teaming de l'IA est plus étendu**. L'équipe rouge de l'IA est désormais un terme générique pour sonder à la fois la sécurité et les résultats de l'IRA.
   Le red teaming de l'IA recoupe les objectifs traditionnels du red teaming dans la mesure où le composant de sécurité se concentre sur le modèle en tant que vecteur.
   Ainsi, certains des objectifs peuvent inclure, par exemple, de voler le modèle sous-jacent.
   Mais les systèmes d'IA héritent également de nouvelles vulnérabilités de sécurité, telles que l'injection rapide et l'empoisonnement, qui nécessitent une attention particulière.
   En plus des objectifs de sécurité, l'équipe rouge de l'IA comprend également l'examen des résultats tels que les problèmes d'équité (par exemple, les stéréotypes) et les contenus préjudiciables (par exemple, la glorification de la violence).
   L'équipe rouge de l'IA permet d'identifier ces problèmes à un stade précoce afin que nous puissions hiérarchiser nos investissements en matière de défense de manière appropriée.
2. **Le red teaming de l'IA se concentre sur les défaillances de personnes malveillantes et bénignes**.
   Prenons le cas du nouveau Bing de l'équipe rouge. Dans le nouveau Bing, l'équipe rouge de l'IA s'est

non seulement concentrée sur la façon dont un adversaire malveillant peut subvertir le système d'IA via des techniques et des exploits axés sur la sécurité, mais aussi sur la façon dont le système peut générer du contenu problématique et préjudiciable lorsque les utilisateurs réguliers interagissent avec le système. Ainsi, contrairement au red teaming de sécurité traditionnel, qui se concentre principalement sur les adversaires malveillants, le red teaming de l'IA prend en compte un ensemble plus large de personas et d'échecs.

3. **Les systèmes d'IA sont en constante évolution.** Les applications de l'IA changent régulièrement. Par exemple, dans le cas d'une application de modèle de langage volumineuse, les développeurs peuvent modifier la métaprompte (instructions sous-jacentes au modèle ML) en fonction des commentaires.
   Alors que les systèmes logiciels traditionnels changent également, d'après notre expérience, les systèmes d'IA changent à un rythme plus rapide.
   Il est donc important de poursuivre plusieurs cycles de red teaming des systèmes d'IA et de mettre en place des systèmes de mesure et de surveillance systématiques et automatisés au fil du temps.
4. **L'association rouge de systèmes d'IA générative nécessite plusieurs tentatives**. Dans le cadre d'un engagement de Red Teaming traditionnel, l'utilisation d'un outil ou d'une technique à deux moments différents sur la même entrée produirait toujours le même résultat.
   En d'autres termes, en général, le red teaming traditionnel est déterministe.
   Les systèmes d'IA générative, en revanche, sont probabilistes.
   Cela signifie que l'exécution de la même entrée deux fois peut fournir des sorties différentes.
   C'est à dessein, car la nature probabiliste de l'IA générative permet une plus large gamme de résultats créatifs.
   Cela rend également le red teaming délicat, car une invite peut ne pas conduire à l'échec de la première tentative, mais réussir (en faisant apparaître des menaces de sécurité ou des préjudices RAI) dans la tentative suivante.
   L'une des façons dont nous avons expliqué cela est, comme Brad Smith l'a mentionné dans son blog, de poursuivre plusieurs séries d'équipes rouges dans la même opération.
   Microsoft a également investi dans l'automatisation qui permet de faire évoluer nos opérations et dans une stratégie de mesure systémique qui quantifie l'étendue du risque.
5. **L'atténuation des défaillances de l'IA nécessite une défense en profondeur.** Tout comme dans le domaine de la sécurité traditionnelle, où un problème tel que le phishing nécessite diverses mesures d'atténuation techniques, telles que le renforcement de l'hôte pour qu'il identifie intelligemment les URI malveillants, la résolution des défaillances détectées via le red teaming de l'IA nécessite également une approche de défense en profondeur.
   Cela implique l'utilisation de classificateurs pour signaler le contenu potentiellement dangereux à l'aide de metaprompt pour guider le comportement afin de limiter la dérive conversationnelle dans les scénarios conversationnels.

La technologie de construction responsable et sécurisée fait partie de l'ADN de Microsoft. L'année dernière, Microsoft a célébré le 20e anniversaire du mémo Trustworthy Computing qui demandait à Microsoft de fournir des produits « aussi disponibles, fiables et sécurisés que des services standard tels que l'électricité, les services d'eau et la téléphonie ».
L'IA s'annonce comme la technologie la plus transformationnelle du 21e siècle.

Et comme toute nouvelle technologie, l'IA est soumise à de nouvelles menaces.
Gagner la confiance des clients en protégeant nos produits reste un principe directeur alors que nous entrons dans cette nouvelle ère - et l'équipe rouge de l'IA est au centre de cet effort.
Nous espérons que cet article de blog inspirera d'autres personnes à intégrer l'IA de manière responsable et sûre via le red teaming.

Ressources

Le red teaming de l'IA fait partie de la stratégie plus large de Microsoft visant à fournir des systèmes d'IA de manière sécurisée et responsable.
Voici d'autres ressources pour vous donner un aperçu de ce processus :

- Pour les clients qui créent des applications à l'aide de modèles Azure OpenAI, nous avons publié un guide pour les aider à constituer une équipe rouge d'IA, à définir la portée et les objectifs, et à exécuter les livrables.
- Pour les intervenants en cas d'incident de sécurité, nous avons publié une barre de bogues pour trier systématiquement les attaques sur les systèmes de ML.
- Pour les ingénieurs ML, nous avons publié une liste de contrôle pour effectuer une évaluation des risques liés à l'IA.
- Pour les développeurs, nous avons publié des conseils de modélisation des menaces spécifiquement pour les systèmes de ML.
- Pour tous ceux qui souhaitent en savoir plus sur l'IA responsable, nous avons publié une version de notre Norme sur l'IA responsable et de notre évaluation de l'impact, entre autres ressources.
- Pour les ingénieurs et les décideurs, Microsoft, en collaboration avec le Berkman Klein Center de l'Université de Harvard, a publié une taxonomie documentant divers modes de défaillance de l'apprentissage automatique.
- Pour l'ensemble de la communauté de la sécurité, Microsoft a organisé le concours annuel d'évasion de l'apprentissage automatique.
- Pour les clients Azure Machine Learning, nous avons fourni des conseils sur la sécurité et la gouvernance de l'entreprise.

Contributions de Steph Ballard, Forough Poursabzi, Amanda Minnich, Gary Lopez Munoz et Chang Kawaguchi.

Recherche et mise en page par:
Michel Cloutier
CIVBDL
20231109
"C'est ensemble qu'on avance"