Université de Montréal
et du monde.

Maison
des affaires
publiques et
internationales

Faculté des arts
et des sciences

Compte-rendus des
événements de la Maison


# The Coming AI Hackers

# The Coming AI Hackers

Cette conférence a eu lieu le **2 février 2026** à l'Université de Montréal. Elle a été organisée par le Centre international de criminologie comparée (CICC) en collaboration avec le Centre de recherche en droit public (CRDP), la Chaire L.R. Wilson et la Maison des affaires publiques et internationales.

## Objectifs :

Dans un contexte marqué par l'omniprésence croissante de l'intelligence artificielle, cet événement visait à analyser les enjeux associés à l'émergence de pirates informatiques spécialisés dans l'IA, communément désignés en anglais comme « AI hackers », en mettant l'accent sur l'intégrité des systèmes informatiques.

Il s'articulait autour de la question suivante :

- Que se passe-t-il lorsque les systèmes d'IA commencent à rechercher des failles dans les règles ?

**Animation :** David Decary-Hetu (Université de Montréal)

**Conférencier :** Bruce Schneier (Université Harvard)

## Thèmes abordés :

- **Hacking :** Le piratage informatique est un processus créatif et malveillant qui consiste à identifier une vulnérabilité dans un système informatique afin de l'exploiter et d'attaquer ce système.
- **AI Hackers :** Avec la présence incontournable de l'IA dans divers secteurs de la vie (banque, finance, santé, économie, politique…), les capacités des pirates informatiques se développent et rendent les différents systèmes informatiques plus vulnérables.
- **Vulnérabilité :** La vulnérabilité fait référence à l'exposition au piratage et à l'existence de failles dans un système informatique pouvant être exploitées pour mener des attaques.
- **Intégrité :** L'intégrité fait référence à la manière dont on maintient un système à l'intérieur d'une enveloppe. Les systèmes informatiques doivent fonctionner de manière fiable, cohérente et conforme à ce pour quoi ils ont été développés. Tous les éléments du système doivent être utilisés de façon intègre (intégrité des données, du modèle et du système…). Elle constitue la garantie que les données ne seront pas modifiées sans autorisation.
- **Sécurité :** La sécurité ne consiste pas seulement à arrêter les attaquants, mais aussi à prévenir les causes de défaillance des systèmes. Les défaillances d'intégrité peuvent causer des dommages considérables. Il s'agit d'événements liés à la sécurité, quelle que soit leur intention.

## Opinions et messages:

- Il existe deux types de défaillances de l'intégrité de l'IA. Le premier cas se produit lorsque l'IA exécute correctement une requête erronée : elle interprète mal les données d'entrée

ou produit des réponses incorrectes en raison de la manière dont elle a été conçue et entraînée. Le deuxième cas survient lorsqu'un adversaire manipule délibérément l'IA afin qu'elle exécute une requête erronée. Toute IA déployée dans un environnement hostile – c'est-à-dire la quasi-totalité d'entre elles – sera vulnérable à ce type d'attaque.

- Le piratage est souvent assimilé uniquement aux systèmes informatiques, mais il peut aussi s'étendre à tout autre système structuré ou organisé par des règles, tel que le système fiscal ou les marchés financiers. Ces systèmes, bien qu'ils ne soient pas directement des systèmes informatiques, peuvent être piratés.
- Les systèmes de piratage s'amélioreront au fur et à mesure que les techniques d'IA deviendront plus sophistiquées. Plus les systèmes d'IA seront capables de résoudre des problèmes complexes et d'apprendre par eux-mêmes, plus le risque de piratage sera grand.

## Points saillants à retenir :

1. L'essor de l'IA transforme profondément le paysage du piratage, en donnant aux attaquants des outils plus puissants pour identifier et exploiter les vulnérabilités des systèmes.
2. La sécurité ne consiste pas uniquement à bloquer les attaquants, mais aussi à prévenir les défaillances internes et les atteintes à l'intégrité, qu'elles soient intentionnelles ou non.
3. L'intégrité est centrale : elle implique de maintenir un système dans les limites prévues par sa conception et de garantir que les données et les modèles ne soient pas modifiés sans autorisation.
4. Les défaillances d'intégrité de l'IA peuvent être involontaires (erreurs liées à la conception ou à l'entraînement) ou résulter d'attaques délibérées dans des environnements hostiles.
5. Le piratage dépasse le cadre strictement informatique : tout système structuré par des règles – y compris économiques ou institutionnels – peut être « piraté » s'il présente des vulnérabilités exploitables.

## Conclusion et perspectives :

La conférence met en avant les différents risques auxquels les systèmes informatiques sont exposés avec l'arrivée de l'IA. Bruce Schneier termine ses propos en se posant un certain nombre de questions :

- Confidentialité : Pouvons-nous construire un réseau sécurisé à partir d'éléments non sécurisés dans un monde non sécurisé ?
- Intégrité : Pouvons-nous construire un système intègre à partir d'éléments non intègres dans un monde non intègre ?

Écrit par Eddy Damaris Nono Defo, étudiante à la maîtrise en environnement et développement durable
Révision par Johannes Müller Gomez, postdoctorant, Maison des affaires publiques et internationales