L’intelligence artificielle

Cybersécurité: Défis pour la Suisse politique

Etat des lieux

Les termes machine learning (apprentissage automatique) et intelligence artificielle (IA) couvrent un vaste domaine qui trouve son origine dans les statistiques et la recherche opérationnelle. Au fond, l’IA est une méthode permettant de prédire des résultats à partir d’ensembles de données. Il s’agit de créer des modèles qui déduisent automatiquement des modèles dans les données et les utilisent pour prendre des décisions. L’IA s’impose comme une technologie essentielle pour la société numérique et l’économie. Nous dépendons de plus en plus de la capacité de l’IA à apprendre des expériences passées, à argumenter, à découvrir un sens ou à classer des données complexes afin de prendre des décisions sensibles et d’automatiser les processus et la prise de décision.

Le développement de l’IA conduit toutefois à **l’intelligence artificielle hostile** (Adversarial Artificial Intelligence, AAI en anglais). L’agresseur utilise l’IA soit pour lancer des attaques afin de compromettre les modèles d’IA utilisés, soit pour adapter et automatiser des éléments d’attaques qui étaient auparavant tout simplement impossibles (deep fake ou hypertrucage) ou qui reposaient fortement sur des processus manuels.

Avec l’AAI, les modèles d’apprentissage automatique interprètent mal les données et se comportent d’une manière favorable à l’agresseur.

Recommandations

1. Les organisations doivent intégrer dans leur évaluation des risques leurs modèles d’IA ainsi que l’automatisation et la prise de décision basées sur l’IA.
2. Il est nécessaire de comprendre où l’adaptation des menaces perpétrées par l’AAI remet en question les attitudes actuelles et où concentrer les ressources limitées.
3. Intégrer la vérification et l’évaluation de la solidité des modèles d’IA dans le processus de prise de décision.
4. Les pouvoirs publics, les entreprises et le secteur de l’éducation doivent veiller à disposer et à développer la base de compétences nécessaire et la réserve de talents.

Pour compromettre le comportement d’un modèle, les agresseurs créent des «exemples/données hostiles» qui ressemblent souvent à des entrées normales, mais qui, au contraire, nuisent à la performance du modèle. Les modèles d’IA classent ensuite ces exemples de manière incorrecte et produisent des réponses incorrectes avec un haut degré de certitude.

Défis

La puissance de calcul peu coûteuse et l'abondance des données collectées ont permis aux modélisateurs et aux agresseurs de développer à faible coût des modèles d'IA de plus en plus complexes. La précision et la complexité des modèles d'IA étant en constante augmentation, de nombreux comportements des modèles échappent à la vaste compréhension humaine. La plupart de ces modèles d'IA sont devenus des boîtes noires. Si un agresseur peut définir un certain comportement inconnu des développeurs du modèle, il pourra ensuite exploiter ce comportement pour en tirer profit.

Plusieurs modèles d'IA, y compris les réseaux neuronaux les plus modernes, sont susceptibles d'être vulnérables aux exemples contradictoires. Cela signifie que ces modèles classifient mal des exemples qui ne diffèrent que légèrement (indétectable pour l'homme) des exemples correctement classifiés.

La vulnérabilité face à l'AAI devient l'un des principaux risques liés à l'utilisation de l'IA dans des environnements où la sécurité est critique. Des attaques contre les technologies de base telles que la vision par ordinateur, la reconnaissance optique de caractères (ROC), le traitement du langage naturel (TLN), la parole et la vidéo (deep fakes) et la détection des logiciels malveillants ont déjà été démontrées.

Voici quelques exemples de menaces venant de l'AAI:

1. Face swapping (échange de visages) dans les vidéos à l'aide d'algorithmes d'apprentissage automatique. Les services de deep fake sont déjà proposés en ligne pour quelques dollars.
2. Reconnaissance/classification d'images dans le domaine de la conduite autonome, p. ex. en cas de mauvaise interprétation des panneaux de signalisation ou des obstacles.
3. Manipulation de la reconnaissance de texte dans les services automatisés destinés au traitement de documents ou de paiements.
4. Capacité à apprendre comment contourner les mécanismes de détection et de contrôle des fraudes basés sur l'IA.
5. Distorsion malveillante des modèles d'IA pour favoriser ou discréditer certains groupes.

Nécessité d'agir

L'AAI cible des secteurs que nous n'avons jamais sécurisés, soit les modèles d'IA eux-mêmes. Les organisations doivent intégrer dans leur évaluation des risques leurs modèles d'IA ainsi que l'automatisation et la prise de décision basées sur l'IA. La défense contre l'AAI inclut des stratégies proactives et réactives. Les stratégies proactives renforcent les modèles d'IA face à des exemples hostiles, tandis que les stratégies réactives visent à les détecter lorsque le modèle d'IA est utilisé.

Nous devons comprendre ce nouvel environnement de menace en constante évolution et remettre de plus en plus en question les processus de prise de décision automatisée en matière d'IA.

Références

- Explaining and Harnessing Adversarial Examples
https://arxiv.org/pdf/1412.6572.pdf
- Ai Is The New Attack Surface
https://www.accenture.com/_acnmedia/Accenture/Redesign-Assets/DotCom/Documents/Global/1/Accenture-Trustworthy-AI-POV-Updated.pdf

What is adversarial artificial intelligence and why does it matter?
https://www.weforum.org/agenda/2018/11/what-is-adversarial-artificial-intelligence-is-and-why-does-it-matter/
- Deepfakes web β
https://deepfakesweb.com

Contact

Nicole Wettstein
Responsable du programme prioritaire Cybersécurité
+41 44 226 50 13

https://www.satw.ch/cybersecurity-defis

Impressum
Académie suisse des sciences techniques SATW
Contributions d'experts
Karl Aberer, EPFL | Umberto Annino, InfoGuard | Alain Beuchat, Banque Lombard Odier & Cie SA | Matthias Bossardt, KPMG | Adolf Doerig, Doerig & Partner | Stefan Frei, ETH Zürich | Roger Halbheer, Microsoft | Pascal Lamia, MELANI | Martin Leuthold, Switch | Hannes Lubich, Verwaltungsrat und Berater | Adrian Perrig, ETH Zürich | Raphael Reischuk, Zühlke Engineering AG | Riccardo Sibilia, VBS | Bernhard Tellenbach, ZHAW | Daniel Walther, Swatch Group Services | Andreas Wespi, IBM Research Lab
Rédaction et graphisme
Beatrice Huber; Claude Naville, Adrian Sulzer, Nicole Wettstein

www.satw.ch
Septembre 2020