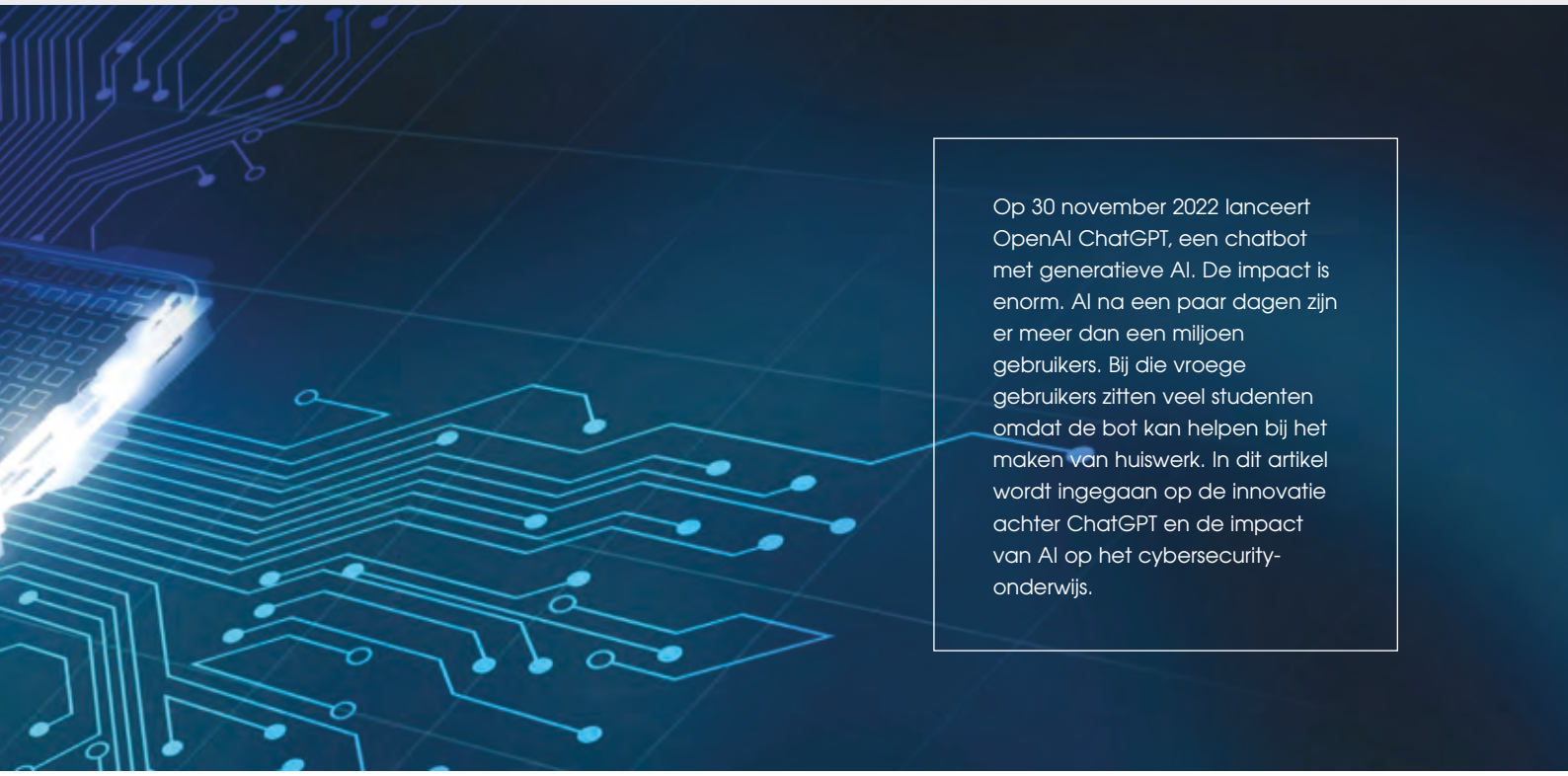
Auteurs: Pieter Burghouwt, Emiel Kerpershoek en Marcel Spruit werken bij de Haagse Hogeschool en verzorgen cybersecurity-onderwijs zowel op bachelor- als masterniveau. Pieter Burghouwt e.a. zijn te bereiken via: P.Burghouwt@hhs.nl.

Kansen en bedreigingen van AI in het onderwijs

Bijna dagelijks is er berichtgeving over nieuwe AI-ontwikkelingen. Veel banen zijn ingrijpend aan het veranderen of gaan ingrijpend veranderen en bedrijven als OpenAI en NVIDIA groeien spectaculair, waarbij laatstgenoemde meer waard werd dan Google 's Alfabet (1). Het is meer dan alleen een hype; achter de soms overdreven verwachtingen liggen belangrijke technische ontwikkelingen, met name op het gebied generatieve AI.



Op 30 november 2022 lanceert OpenAI ChatGPT, een chatbot met generatieve AI. De impact is enorm. Al na een paar dagen zijn er meer dan een miljoen gebruikers. Bij die vroege gebruikers zitten veel studenten omdat de bot kan helpen bij het maken van huiswerk. In dit artikel wordt ingegaan op de innovatie achter ChatGPT en de impact van AI op het cybersecurity-onderwijs.

Cybersecurity en AI hebben belangrijke raakvlakken. Niet alleen wordt AI ingezet in het cybersecuritydomein voor bijvoorbeeld detectie van aanvallen; AI is zelf ook een IT-systeem en moet beschermd worden en veilig zijn.

Cybersecurityonderwijs en Machine Learning

Cybersecurity is een onderdeel of specialisatie van veel IT-opleidingen van het bacheloronderwijs, zoals bijvoorbeeld hbo-ICT en Computer Science. Daarnaast zijn er cybersecuritymasteropleidingen, die zowel vanuit hbo-instellingen als universiteiten worden aangeboden (2).

Al vele jaren introduceren de meer technisch-gerichte cybersecuritycurricula ML (Machine Learning) als middel om aanvallen, zoals intrusions, anomalies, spam en malware, te kunnen detecteren. De diepgang loopt daarbij sterk uiteen, afhankelijk van het opleidingsniveau en de focus van de studie. Uiteraard worden daarbij bekende ML-algoritmes geïntroduceerd, zoals bijvoorbeeld: neurale netwerken, beslisbomen en Naive Bayesian Filters maar er zijn wel verschillen:

- Op universiteiten wordt doorgaans een meer theoretische benadering gehanteerd, waarbij een sterke nadruk ligt op wiskundige fundamenten en theoretische aspecten van AI. Vaak is Machine Learning een onderdeel van datascience. Daarnaast heeft uiteraard de focus van onderzoeksgroepen invloed op de inhoud van het curriculum.
- Het onderwijs op hogescholen is meer gericht op praktische toepassingen en het gebruiken bij het oplossen van concrete problemen. De studenten maken bijvoorbeeld kennis met Naive Bayesian Classifier om op praktische wijze SPAM van normale e-mails te onderscheiden.

Voor projecten en practica met ML wordt vaak Python gebruikt, met name vanwege de toegankelijkheid en de uitgebreide ondersteuning voor machine learning bibliotheken. Studenten maken daarbij meestal gebruik van publiekelijk beschikbare datasets om modellen te trainen en te testen. Door de focus op detectie wordt ML vaak ingezet voor binaire classificatie (bijv. wel of geen intrusion). Ook bij digitale forensische analyse speelt ML een belangrijke rol, zoals predictive coding bij E-discovery.

Beperkingen van ML in cybersecurity-onderwijs

Bij het aanbieden van onderwijs op het gebied van ML spelen er twee belangrijke beperkingen:

1. Effectieve toepassing van ML-algoritmes vereist domeinspecifieke en wiskundige kennis. Dit is met name voor hogescholen een belangrijke beperking omdat studenten daar minder op getraind worden dan studenten van universiteiten. Het gebruik van ML-bibliotheken en GUI-gebaseerde

datascience-applicaties kunnen helpen, maar het blijft voor studenten lastig om te overzien welk algoritme het beste kan worden toegepast, welke features (input) gebruikt moeten worden en hoe training en testen het beste kunnen worden aangepakt.
2. Succesvolle toepassing van ML vereist gespecialiseerde grote datasets, die vaak nog gelabeld moeten zijn voor training. Het verkrijgen van dergelijke datasets is ingewikkeld en tijdrovend, wat de praktische mogelijkheden voor educatieve doeleinden beperkt.

Foundation modellen

AI tientallen jaren speelt ML een belangrijke rol in het AI-domein. In plaats van een machine te vullen met regels, leert de machine uit data om daarmee te kunnen voorspellen, classificeren of genereren.

Vanaf ongeveer 2010 gaat dit een stap verder door ontwikkelingen in Deep Learning (3). Met name bij beeldverwerking worden dan grote stappen gezet. De machine leert van ruwe data, zoals pixels, zonder 'handmatige' feature-engineering. Vanaf 2017 zijn er veel nieuwe ontwikkelingen op het gebied van generatieve AI. Met generatieve AI is het mogelijk om op basis van input tekst, beeld en geluid te genereren. Op zichzelf is dat niet nieuw, maar nieuwe architecturen, zoals met name transformers maken het mogelijk om AI te trainen op basis van zeer grote hoeveelheden data, door middel van self supervised learning (4). Hierdoor ontstaat een nieuw soort modellen, aangeduid als foundation modellen of GPAI (General Purpose AI) (5). Bekende voorbeelden van deze foundation modellen zijn GPT4, Gemini en Llama. Eén van de kenmerkende eigenschappen van foundation modellen is de hoge mate van 'transfer learning', waarbij het model ook problemen kan oplossen in domeinen waarvoor het niet of slechts minimaal getraind is.

Hierdoor is er sprake van een paradigma-shift in het ontwerpen van AI-systemen van smalle gespecialiseerde modellen, die gerealiseerd worden door intensieve training op een bepaalde taak, naar de meer universele pre-trained foundation modellen, die met beperkte aanpassingen op een nieuwe taak kunnen worden ingezet. Deze eigenschap droeg ook bij aan het succes van ChatGPT; de assistent die niet alleen welbespraakt is, maar ook zeer uiteenlopende taken kan uitvoeren, waaronder gespecialiseerde taken op basis van beperkte instructies en voorbeelden.

Generatieve AI in het cybersecurity-onderwijs

De komst van generatieve AI geeft in het cybersecuritydomein nieuwe mogelijkheden, zoals:

1. Het semantisch zoeken in documenten op basis van een vraag. Aan het AI-model wordt een vraag gesteld, waarbij een document(deel) wordt meegenomen. Door het meenemen van een document(deel) is het antwoord veel preciezer en terug te leiden naar de specifieke bron. Als gewerkt wordt met grote documenten of een hele bibliotheek, dan wordt door middel van speciale vectoren, 'embeddings', eerst het deel in de documentatie geselecteerd dat het meest relevant is. Men noemt deze techniek RAG (Retrieval Augmented Generation) (6).
2. Agents die bijvoorbeeld via API's, complexe applicaties, diensten of tools aanroepen. Hiermee kan op basis van een vraag een ingewikkelde handeling worden uitgevoerd. Het resultaat kan weer geïnterpreteerd en eenvoudig uitgelegd worden door het AI-systeem.
3. Het genereren van software en configuraties op basis van prompts. Zowel red teams als blue teams kunnen het ontwikkelen van speciale software of configuraties vereenvoudigen met GenAI. Github Copilot is een voorbeeld van een ontwikkeltool, die op basis van instructies in mensentaal, software kan schrijven, verbeteren en uitleggen (7).
4. Datascience op basis van een dataset met daarbij een vraag. Het AI-systeem voert automatisch data-analyse uit en gebruikt daarbij gegenereerde software voor analyses op de dataset. De uitvoer wordt vertaald naar resultaten in de vorm van een begrijpelijk antwoord (8).
5. Detectie van bijvoorbeeld threats en anomalies waarbij alleen finetuning nodig is in plaats van een volledige training (9).

De toepassingen zijn nog volop in ontwikkeling en het ligt voor de hand dat ze vaker teruggevonden gaan worden in de cybersecuritycurricula; waarschijnlijk eerst in projecten en later structureel in andere onderwijsvormen.

Een andere belangrijke ontwikkeling is dat kwaadwillenden steeds vaker AI zullen gebruiken om aanvallen te ondersteunen, bijvoorbeeld door te helpen met het programmeren van malware, het opstellen van geloofwaardige phishingmails of het genereren van deep fake beeld en geluid. Ook die kennis zal zijn weg moeten vinden in de cybersecuritycurricula (10).

Ten slotte vormen AI-systemen zelf ook een doelwit van cyberaanvallen (11). Voorbeelden hiervan zijn:

- Tijdens training het model beïnvloeden door aanpassing van training-data, pretraining-data, finetuning-data of modelparameters (poisoning)
- Tijdens gebruik het model misleiden door bepaalde features in de data aan te passen (evasion)

- De vertrouwelijkheid van het systeem aanvallen, bijvoorbeeld door middel van speciale input, die een preprompt-tekst laat zien (prompt injection)
- Aanvallen die te maken hebben met de implementatie, bijvoorbeeld een datalek via een AI-provider of ondersteunende software en bibliotheken met daarin Trojaanse paarden etc.

Ook dit zal aandacht moeten krijgen in de cybersecuritycurricula.

Ongeoorloofd gebruik AI

Sinds de komst van ChatGPT hebben studenten generatieve AI massaal omarmd. Schattingen over het gebruik van ChatGPT door studenten lopen uiteen. Uit onderzoek in het VK van oktober 2023 bleek dat 32% van de studenten meerdere keren per week gebruik maakt van ChatGPT maar bij informatica en technische studies lag dit op 66% (12). Het is natuurlijk zorgelijk wanneer generatieve AI door studenten ongecontroleerd wordt ingezet voor het maken van studieopdrachten. Bovendien worden in het hoger onderwijs de leerprestaties beoordeeld op basis van opdrachten die buiten het zicht van de docenten worden gemaakt. De validiteit en betrouwbaarheid van die beoordeling kan ernstig worden verstoord als onduidelijk is in hoeverre AI heeft geholpen bij het maken van een opdracht. Weliswaar zijn hiervoor verschillende detectietools ontwikkeld, maar ze blijken in de praktijk vaak foutgevoelig (13). Vaak is het gebruik van generatieve AI, zoals ChatGPT, zelfs zonder detectietools al duidelijk te herkennen. Zo zal de AI-applicatie in sommige gevallen literatuurreferenties verzinnen of een stuk hoogdravende tekst genereren zonder echte inhoud. Maar het gebruik van generatieve AI is niet altijd te bewijzen, zeker niet als generatieve AI verder evolueert.

In plaats van detectie is preventie van ongeoorloofd AI-gebruik in de studie een betere oplossing. Dat kan enerzijds door te werken in een gecontroleerde omgeving, die tijdens toetsing de toegang tot AI blokkeert. Anderzijds kan dat door het gebruik van AI juist te omarmen en zo veel mogelijk in te bedden in het onderwijs. In dat laatste geval moeten studenten bij gebruik goed de mogelijkheden en risico's weten en verantwoordelijkheid nemen. Belangrijke verantwoordelijkheden in deze context zijn:

1. Werken met toestemming (van de opdrachtgever en/of de school en binnen wet- en regelgeving)
2. Transparant AI-gebruik (door precies aan te geven waar en hoe AI gebruikt is)
3. Correctheid en aanvaardbaarheid van AI output (overgenomen AI-resultaten moeten correct en veilig zijn en mogen bijvoorbeeld dus niet gevaarlijk of beledigend zijn)
4. Geen hinder bij beoordeling leeruitkomsten (AI mag niet gebruikt worden voor zaken die de student zelf moet doen om de beoogde leeruitkomsten aan te tonen)

Dergelijke verantwoordelijkheden zijn alleen te nemen met voldoende kennis over hoe AI op de juiste manier gebruikt moet worden. Zaken als correctheid, ethiek, verklaarbaarheid en nieuwe regelgeving, zoals de AI-Act (14), spelen daarbij een belangrijke rol en horen dus ook in een modern cybersecuritycurriculum.

Ten slotte bieden de nieuwe mogelijkheden van AI ook allerlei didactische kansen in het onderwijs. Zo bieden OpenAI en Microsoft customizable AI-agents met aanvullende documenten, API-calls en prompting. Hiermee kan een docent een AI-applicatie maken die bijvoorbeeld de slides van een les uitlegt of oefenvragen stelt en die na beantwoording uitleg geeft of een inhoudelijke discussie aangaat met een student.

Kortom, we zitten in een zeer interessante periode, waarin AI flinke veranderingen brengt in het onderwijs en zeker ook in het cybersecurity-onderwijs.

Referenties

(1) https://www.forbes.com/sites/dereksaul/2024/02/12/nvidia-is-now-more-valuable-than-amazon-and-google/
(2) https://communities.surf.nl/cybersecurity/artikel/cybersecurity-opleidingen-bij-mbos-hogescholen-en-universiteiten-in-nederland
(3) https://www.nature.com/articles/nature14539
(4) https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee-243547dee91fbd053c1c4a845aa-Abstract.html
(5) https://arxiv.org/abs/2108.07258
(6) https://proceedings.neurips.cc/paper/2020/hash/-6b493230205f780e1bc26945df7481e5-Abstract.html
(7) https://docs.github.com/en/copilot
(8) https://platform.openai.com/docs/assistants/tools/code-interpreter
(9) https://sciendo.com/article/10.2478/kbo-2023-0072?content-tab=abstract
(10) https://ieeexplore.ieee.org/abstract/document/10198233
(11) https://csrc.nist.gov/pubs/ai/100/2/e2023/final
(12) https://www.thehrdirector.com/business-news/ai/chatgpt-32-university-students-admit-using-weekly/
(13) https://hai.stanford.edu/news/ai-detectors-biased-against-non-native-english-writers?ref=gptzero.ghost.io
(14) https://artificialintelligenceact.eu/the-act/