ARCSI
Association des Réservistes du Chiffre et de la Sécurité de l'Information

Université Paris Cité

Compte rendu du Lundi de la cybersécurité du 16 mars 2026

Cyber et AI à l'horizon 2030 : la bataille pour la vitesse

Présenté par Guy-Philippe GOLDSTEIN

Organisé par Pr. Ahmed Mehaoua, Béatrice Laurent et Gérard Peliks
Rédigé par Rayan Al Mohaize, étudiant en Master 2 Cybersécurité

Table des matières

I Introduction

L'événement « Lundi de la Cybersécurité » est une série mensuelle de conférences en ligne créée et animée de manière indépendante par Gérard Peliks et Béatrice Laurent. Elle a pour vocation d'explorer chaque mois une thématique différente liée aux enjeux contemporains de la sécurité numérique et de la société.

1 Le thème de la course à l'IA entre les cyberattaquants et les cyberdéfenseurs

Le Lundi de la cybersécurité du 16 mars 2026 portait sur la bataille pour la vitesse entre les cyberattaquants et les cyberdéfenseurs concernant l'IA. Lors de ce webinaire, la réflexion s'est faite sur les différentes problématiques et difficultés auxquelles la cybersécurité est confrontée face à la croissance exponentielle de l'IA et à sa capacité d'amplifier les attaques, et comment nous pouvons nous adapter à cette croissance pour ne pas perdre la "course".

2 Les intervenants

Pour animer ce Lundi de la cybersécurité, **Guy-Philippe GOLDSTEIN**, professeur à l'École de guerre économique, nous a fait part de son expertise pour nous expliquer les enjeux de cette course à l'IA dans le monde de la cybersécurité, dans un contexte où l'intelligence artificielle et notamment l'IA agentique explosent et deviennent de plus en plus performantes et inquiétantes.

Par la suite, **Julien BREYAULT** a animé le quart d'heure des associations pour présenter CyberEdu, l'association dont il est le président.

II Cyber et AI à l'horizon 2030 : la bataille pour la vitesse (Par Guy-Philippe GOLDSTEIN)

1 Etat de l'art : "Winter is comming"

"Winter is comming" est une citation tirée de la série Game of Thrones qui est utilisée pour désigner que quelque chose de très dangereux va arriver très prochainement. C'est ainsi que Guy-Philippe GOLDSTEIN commence sa présentation avec une certaine inquiétude.

En 2025, des actions ont été menées, aux États-Unis et au sein de l'Union Européenne, dans le cadre de l'émergence des risques cyber liés à l'évolution du déchiffrement quantique devant être maîtrisée et balisée. Cette évolution de l'informatique quantique avait engendré un choc dans la sphère cyber à l'échelle mondiale. Cependant, comme nous l'a indiqué Guy-Philippe GOLDSTEIN, "un train peut en cacher un autre". Le risque est de ne pas voir le deuxième choc concernant la bascule liée à l'IA, dont l'accélération pourrait être bien plus dangereuse car celle-ci n'aura pas été balisée.

1.1 Ingénierie sociale amplifiée

L'ingénierie sociale amplifiée constitue la plus grande menace et le facteur de risque le plus craint dans le cadre de l'utilisation de l'IA dans des attaques cyber. Cela permet notemment une sophistication considérable du spearphishing (hameçonnage ciblé), une pratique ayant augmentée de plus de 4000% en 2022 d'après SlashNext. En effet, le spearphishing est notamment utilisé par la petite criminalité contre des personnes vulnérables (personnes âgées, personnes sans maturité cyber, etc.), notamment en procédant à du voice cloning, une pratique devenue courante et qui a augmenté de 400% en 2025, consistant à imiter la voix d'une personne avec l'IA.

Le spearphishing et le voice cloning peuvent aussi atteindre les grandes organisations et les dirigeants. Guy-Philippe GOLDSTEIN nous donne un cas réel de fraude au directeur qui utilise du voice cloning dans des outils de collaboration. En effet, un attaquant s'est fait passer pour un collaborateur lors d'une visioconférence dans la filiale hongkongaise de la société d'ingénierie Arup, et a demandé d'effectuer une transaction de 25 millions de dollars, que l'attaquant a pu voler. Ainsi, on pourrait s'attendre dans le futur à une utilisation du deepfake combiné au voice cloning pour imiter la voix mais également l'image d'une personne de confiance, ce qui pourrait rendre les attaques encore plus dangereuses et difficiles à détecter par un humain.

D'autres exemples concrets nous ont été présentés par Guy-Philippe GOLDSTEIN, qui se sont déroulés notamment dans la zone indopacifique, où ce genre d'attaque est largement répandu :

- **Octobre 2024**. Des gangs cybercriminels en Asie ont été arrêtés par la police de Hong Kong et auraient récupéré 46 millions de dollars en utilisant des jeunes femmes, créées par deepfake, pour escroquer des dirigeants d'entreprise de Taïwan, de Singapour ou de l'Inde.
- **Avril 2024**. De fausses communications de grands dirigeants ou de personnes ayant une forte influence sont diffusées au grand public avec des vidéos deepfake dans le but de manipuler les cours de bourse.

1.2 L'avantage des attaquants

Dans cette partie, Guy-Philippe GOLDSTEIN fait une analogie avec le monde médiéval pour montrer que, dans la protection d'une forteresse, c'est le défenseur qui a l'avantage sur l'assaillant : on dit qu'il faudrait trois assaillants pour un défenseur. Cet avantage est dû à la parfaite connaissance des lignes de communications au sein de la forteresse et à la position haute des défenseurs pour avoir une vue globale sur la forteresse et tout ce qui l'entoure, mais aussi pour une anticipation optimale des attaques.

Or, dans notre paradigme cyber actuel et dans un contexte de protection d'une organisation, l'avantage passe du côté des attaquants. En effet, les réseaux de communication corporate sont extrêmement complexes (il est très compliqué d'avoir un visuel sur plus de 95% de l'ensemble des activités du réseau), et il est pratiquement impossible de détecter un attaquant au moment de sa pénétration (la détection se fait en moyenne au bout de 5 à 10 jours après la pénétration). De plus, les hackers sont très peu condamnés, voire pas du tout dans certains cas. Plus la taille de l'organisation est importante, plus ces complexités

deviennent importantes et plus cet avantage pour les attaquants augmente.

D'après Palo Alto et Google, en moyenne il faut 5 à 10 jours pour arriver à détecter une tentative de compromission de données et en moyenne 2 jours pour passer de la compromission initiale à l'exfiltration des données. Cet écart de vitesse et de réactivité entre les attaquants et les défenseurs risque de s'agrandir en faveur des attaquants avec l'évolution de l'IA, tel qu'on l'observe dans les attaques cyber, et les défenseurs risquent de ne plus avoir les capacités nécessaires pour pouvoir assurer une protection efficace.

1.3 L'IA qui contribue à l'accélération du tempo des attaques

Guy-Philippe GOLDSTEIN nous présente un graphique qui montre le pourcentage de CVE exploités avant ou le jour même de leur publication. On est passé de 39% en 2023 à 67% en 2026 des CVE exploités avant ou le jour même de leur publication. Or, cette période entre 2023 jusqu'à maintenant correspond bien à l'explosion de l'IA, notamment avec l'apparition de GPT-3.5 fin 2022 et de l'apparition plus tard des IA agentiques. Nous pouvons donc en conclure qu'en effet, l'IA contribue aujourd'hui à l'exploitation des CVE.

À partir de 2023, nous avons commencé à évaluer les capacités "intellectuelles" des LLM en les faisant participer à des concours type olympiade de mathématiques. En 2023, aucun LLM n'avait pu gagner de médaille, en 2024 DeepThink de Google gagne une médaille d'argent, et en 2025 DeepThink et OpenAI gagnent une médaille d'or. On observe très clairement ici des capacités qui évoluent de manière relativement rapide avec des effets relativement surprenants.

Une autre étude a été menée pour évaluer les capacités des LLM à gérer des tâches complexes. Ce qui a été observé est que jusqu'aux modèles GPT-3.5 et GPT-4, les LLM pouvaient gérer des tâches qu'un humain prendrait jusqu'à 4 heures à réaliser. Puis en 2026, une envolée de ces capacités a été observée, notamment avec le modèle Opus4.6 pouvant réaliser des tâches complexes qu'un humain prendrait jusqu'à 12h à réaliser, par exemple exploiter une vulnérabilité dans un contrat Ethereum. De plus, on peut observer d'après les données empiriques que cette capacité à mener des projets de grande envergure avec un chaînage complexe double tous les 3 mois. On peut donc nous interroger sur leur capacité à mettre en place des chaînages complexes pour des projets de hacking.

En 2024, l'idée du hacker IA autonome relevait auparavant du domaine de la science-fiction et est devenue une réalité depuis 2025. Par exemple :

- En février 2025, le MITRE a démontré que Deepseek était capable de fournir de bonnes approches de cyberoffensive et que Llama 3 pouvait mener et réussir des mouvements latéraux et des approches living off the land.
- En mai 2025, Palisades Research, un labo de recherche californien, a utilisé des IA autonomes pour participer à des CTF et les IA sont arrivées dans le top 5% des meilleurs compétiteurs.
- En juin 2025, le système IA XBOW arrive en première place du classement Hackerone dans la détection de bug bounty
- En août 2025, plus de 54 vulnérabilités ont été découvertes en 4 heures par des IA lors de la DARPA AI Cyber-Challenge
- En septembre 2025, une équipe de chercheurs de Boston U. and UC Santa Barbara

crée un LLM qui, sur la base de CVE, crée un exploit vérifiable avec un taux de succès de 51%.
— En janvier 2026 est arrivée l'industrialisation de la génération d'exploits en utilisant des modèles LLM.

D'autres révélations soulignent les capacités surprenantes et inquiétantes de l'IA, notamment dans le domaine cyberoffensif, comme l'utilisation de Claud par les Chinois dans le cadre d'une campagne de cyberespionnage.

Pour finir, Guy-Philippe GOLDSTEIN nous raporte une reponse interessantes à une question qu'il avait posé lors de la conférence "AI action summit" :
Question de Guy-Philippe GOLDSTEIN : Quelle est la projection sur les risques émergents de l'IA dans le cadre des cybermenaces ?
Réponse de Nicolas Batz (patron de la cybersécurité de microsoft) : D'ici 2027, quasiment 100% des attaques seront augmentées par l'IA.

Un autre point de vue intéressant est celui de Heather Adkins, VP, Security Eng., Google, Bruce Schneier, Harvard Kennedy School et Gadi Evron, Fmr Chairman IL-CERT qui annoncent dans leur article "Autonomous AI Hacking and the Future of Cybersecurity", publié en octobre 2025 dans le magazine CSO, que nous sommes proches d'un point de bascule dans lequel les IA agentiques pourraient égaler, voire surpasser les meilleurs des hackers humains.

2 Horizon 2030 : Nouveau champ de batille et aspect offensif

2.1 L'ingénierie sociale en 2030

Avec les capacités actuelles de l'IA et les observations sur son évolution rapide, à l'horizon 2023, l'ingénierie sociale pourrait considérablement se transformer, notamment en utilisant des capacités d'analyse comportementale, d'identification des traits de personnalité d'un individu et de prédiction des intentions d'un individu (theory of mind). D'après Michał Kamiński, chercheur en psychologie à Stanford, en termes de théorie of mind, GPT-3 avait les capacités d'un enfant de 7 ans, GPT-3.5 avait les capacités d'un enfant de 9 ans et GPT-4 a aujourd'hui les capacités d'un adulte. Par cette évolution, les IA pourraient maîtriser des modèles psychosociaux avancés et développer par exemple des capacités de persuasion des êtres humains, ce qui engendrerait des risques cyber considérables.

Nous pouvons aussi imaginer des IA organisées en un ensemble d'agents qui pourraient s'introduire et s'accrocher discrètement dans des organisations, via du spear phishing, pour créer des jumeaux numériques de l'organisation qui permettraient de savoir quelles sont les failles techniques, organisationnelles et humaines, pour mettre en place des plans de side channel attack. Ce type d'attaque consiste à inférer de tout un ensemble d'informations utiles pour mener une attaque. Comme avec l'exemple de la visioconférence chez Arup, l'IA pourrait récolter toutes les données possibles (image, son, etc.) et traduire ces signaux en information.

2.2 Les nouvelles menaces

Dans cette partie, Guy-Philippe GOLDSTEIN nous présente toutes les nouvelles menaces apparues notamment depuis 2025, susceptibles d'évoluer et de créer une menace cyber importantes. Aujourd'hui on peut observer que des IA peuvent maîtriser des technologies anciennes et pas bien connues dans lesquelles il n'y a pas beaucoup d'experts, elles peuvent devenir des expertes de Cobol, être capables de comprendre de l'OT ou les BIOS Firmware par exemple. Tous ces éléments pourraient mener à des attaques difficiles à gérer par des êtres humains.

L'IA peut être aussi capable de mener des attaques basées sur des masses de type DDoS combinées avec de la superintelligence, on appelle cela le DDAS (Distributed Deployment of Agent Swarms). Ce type d'attaque consisterait à déployer plusieurs dizaines, voire centaines d'agents hackers autonomes pour mener des attaques complexes. Cette menace est apparue notamment avec l'apparition de OpenClaw pouvant gérer 10 agents max et de ClaudFlow pouvant gérer une centaine d'agents. Or, cette capacité multi-agent est largement susceptible d'évoluer dans le futur : qu'en sera-t-il quand des systèmes seront capables de gérer plusieurs milliers, voire plusieurs millions d'agents ?

3 Horizon 2030 : La nouvelle défense

Dans cette partie, Guy-Philippe GOLDSTEIN nous donne un peu d'espoir en expliquant que si toutes ses évolutions sont possibles du côté attaquant, alors des évolutions sont aussi possibles du côté défenseur, et qu'il est possible pour les défenseurs de gagner cette course ! En effet, aujourd'hui, les systèmes agentiques permettent d'aller beaucoup plus vite dans la détection. Voici quelques statistiques intéressantes rapportées par Stellar Cyber, Torq, ReliaQuest, Prophet Security et Omdia 2025-2026 sur les capacités défensives de l'IA agentique :

- 20 fois plus rapide pour la détéction ;
- -60% MTTR ;
- 3 min pour l'investigation IAM (vs 1 jour aujourd'hui).

L'IA pourrait aussi être utilisée pour détecter et régler des bugs dans du code, et pour analyser des langages de preuve formelle en les traduisant en langage plus solide, tel que Rust, pour limiter les vulnérabilités du code. Google prétend atteindre en 2030 l'objectif zéro vulnérabilité dans leurs softwares. Dans ce contexte, ils ont mis en place des modèles d'IA innovants tels que BIG SLEEP, pour détecter les 0-days et confirmer leur exploitabilité, et CODEMENDER pour les remédiations de ces vulnérabilités. Ces modèles pourraient aussi enrichir le LLM de Google Gemini, une initiative s'inscrivant dans la compétition des LLM entre Gemini, Claude et ChatGPT. Depuis 12 mois, BIG SLEEP a pu détecter près de 25 vulnérabilités critiques et, depuis octobre, CODEMENDER a pu réparer plus de 60 exploits, dont un exploit 0 click sur iOS.

En outre l'IA dans un contexte de défense pourrait être utilisée pour :

- Mettre en place des pièges pour ralentir les attaquants.
- Augmenter les segmentations, voir la micro segmentation.
- Mise en place de jumeaux numériques de l'organisation, y compris de l'organisation humaine, pour mieux appréhender les attaques, piéger et absorber les attaquants.

4 Horizon 2030 : Les mondes nouveaux

Guy-Philippe GOLDSTEIN nous explique que pour gagner la course contre la mauvaise IA il faudrait comprendre le tempo des attaques, être capable d'évoluer en interne en termes de maturité sur l'IA et la cyber et mettre en place des mesures pour ralentir l'attaquant (cybersécurité décéptive). En effet, la cybersécurité déceptive est estimée à 5% du marché mondial des produits de cyber en 2024 et atteindra 50% en 2030. Cette problématique de l'IA devient centrale dans la cybersécurité aujourd'hui. En 2026, soit on est capable de s'adapter et on gagne la course, soit on est trop lent et on perd la course.

III Présentation de CyberEdu (Par Julien BREYAULT)

Comme toujours, la seconde partie de ce Lundi de la cybersécurité est dédiée à un représentant d'une association pour lui donner l'opportunité de représenter ses objectifs et ses valeurs. Julien BREYAULT, président de CyberEdu, nous a donc fait l'honneur d'animer ce quart d'heure des associations pour nous présenter CyberEdu.

1 Présentation du label

La démarche derrière CyberEdu a été initiée par la stratégie nationale pour la sécurité du numérique, présentée le 16 octobre 2015 par le Premier ministre, dans laquelle l'aspect formation cyber a été fortement mis en avant. Par la suite, s'en est suivie une démarche portée par France Université, où a été mis comme priorité de sensibiliser à la cybersécurité les étudiants des formations du supérieur. Le postulat de départ est qu'une grande partie de la cyber est entre les mains de non-experts qui sont malgré eux des relais d'attaques qui réussissent. L'enjeu est de faire que ses non-experts puissent se sentir impliqués et être plus vigilants aux risques cyber.

Le partenariat qui en a résulté avec l'Agence Nationale de la Sécurité des Systèmes d'Information (ANSSI) porte sur deux objectifs principaux :

- Intégrer la sensibilisation à la cybersécurité dans toute formation supérieure et dans les formations continues.
- Intégrer la formation à la cybersécurité dans toute formation supérieure intégrant une part informatique.

Dans ce contexte, un référentiel de compétences a été créé au niveau de l'ANSSI pour se mettre d'accord sur le niveau de compétence cyber dont les industriels et les services éthiques ont besoin, pour ensuite labelliser ceux ayant une démarche active de sensibilisation et de formation à ces enjeux.

2 Présentation de l'association

L'association réunit des enseignants en informatique, spécialistes en sécurité et non spécialistes afin de contribuer à l'objectif commun, qui est d'améliorer la compréhension des enjeux cyber, notamment pour les non spécialistes. Pour atteindre cet objectif, l'association à participer à l'évolution des BTS, des BUT et autres diplômes pour y intégrer plus de cybersécurité.

3 Critères d'éligibilité

Pour qu'une formation soit éligible, il faut qu'elle soit reconnue, longue et aboutisse à un diplôme (BUT, BTS, Licence, Master, etc...) et il ne faut pas qu'elle soit spécialisée dans la cyber. Un autre label (SecNumEdu) est dédié aux formations spécialisées en cyber. L'objectif du label CyberEdu est de labelliser les formations qui ne rentrent pas dans la catégorie expert cyber, c'est à dire qui comportent moins de 70% de contenu cyber, ce qui permettrait aux étudiants de mettre en avant la plus-value de ce label dans leur CV, par exemple des formations en droit ou en comptabilité.

4 Groupe de travail

Un financement initial en 2016 a permis de produire des ressources pédagogiques publiées par l'association. Mais elle nécessite des mises à jour et mériterait aussi des compléments et des adaptations par secteur/métier. Avec le soutien du CFSSI et du coordinateur national cyber France 2030, Florent KIRCHNER, l'association cyberEdu aura l'opportunité de créer des ressources « communs cyber », sur la base de ceux existants, pour faire vivre cette « mallette pédagogique » de l'association et la rendre accessible à tous.

IV Questions / Reponses

La session de questions/réponses a été entièrement menée par Guy-Philippe GOLDSTEIN, portant toute sur sa présentation.

Les agents IA agentiques sont ils les chevaux de Troie du 21e siècle ?

Les agents sont très clairement des vecteurs de menace très forts. Des règles pourront tout de même être mises en place pour encadrer l'utilisation de ces agents et pour qu'ils ne soient pas utilisés de manière malveillante, mais ces règles seront probablement inefficaces. Il faudra donc trouver des moyens de défense contre ces agents, plutôt que des règlementations.

Est ce que l'IA peut être capable de différencier un deepfake du réel ?

Oui, il existe déjà plusieurs solutions pour différencier les deepfakes du réel, mais cela s'inscrit aussi dans la course. Jusqu'où pourra aller l'IA dans la création de deepfakes et jusqu'où pourrons-nous mettre en place des systèmes permettant de les détecter ?

Comment vont faire les petites entreprises pour se protéger des attaquants utilisant l'IA avec de grosses ressources ?

Effectivement les petites entreprises auront du mal à se protéger tandis que les grands groupes pourront se préparer et acquérir les ressources nécessaires pour leur défense. Cependant, ce qui est inquiétant dans cette problématique est le fait que la vulnérabilité des ETI et des PME puisse mener à des attaques assez conséquentes pour créer un effet rebond qui pourrait atteindre les grands groupes et les rendre vulnérables à leur tour.

L'IA peut elle trouver réellement des bugs dans le code ?

Les données empiriques démontrent qu'il est largement possible de trouver les bugs dans le code, notamment depuis 2025.

L'architecture d'internet n'est-elle pas devenue caduque face à l'explosion de l'IA ? Ne faudrait-il pas imaginer une autre architecture fondée sur un contrôle plus vigoureux ?
***REMARQUE** : Gerard Peliks et la personne ayant posé la question ont rebondi sur ce sujet en évoquant RINA, une architecture qui intègre plus nativement la sécurité que l'architecture actuelle basée sur TCP/IP.*

D'après Guy-Philippe GOLDSTEIN, le fait de faire évoluer l'architecture et d'ajouter des protocoles complémentaires pour plus de sécurité peut être une bonne initiative, l'essentielle est que cela puisse s'intégrer facilement dans l'architecture existante en prenant bien en compte l'aspect économique : "Si ça coûte trop cher, personne n'en voudra !".

Le portefeuille d'identité numérique, mis en place par le nouveau règlement européen eIDAS version 2, est-il un élément de confiance déterminant dans le cadre de la sophistication et des dangers des IA ?

Tout élément supplémentaire de meilleure authentification est bénéfique. La difficulté se trouve dans l'intégration dans les workflows professionnels où l'on va retrouver le problème du facteur humain. Par exemple, si un banquier est appelé par un client, doit-il lui imposer des procédures d'authentification avant d'échanger avec lui, ou doit-il se fier à son propre jugement et se baser sur le numéro de téléphone et la voix qui correspondent bien au client ? Avec la sophistication des deepfakes et des clonages des voix par l'IA, il va falloir mettre en place des outils d'authentification. Or, le design de ces outils se doit d'être simple à intégrer, de la façon la plus naturelle possible, dans les workflows professionnels.

Comment pouvons-nous nous protéger face aux nombreuses failles reconnues présentes sur les systèmes d'IA étrangers ?

En effet, les outils peuvent contenir des failles de base mais aussi des failles installées volontairement par les concepteurs étrangers. Par exemple des outils qui permettent de détecter des bugs dans du code mais qui, by design, ne vont pas détecter certains bugs. La souveraineté est indispensable pour maîtriser ces technologies.