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

Université Paris Cité

Compte-rendu du « Lundi de la cybersécurité » n°67
Lundi 18 Mars 2024

Intelligence artificielle générative : géniale et dangereuse

Organisé par Pr. Ahmed Mehaoua, Béatrice Laurent et Gérard Peliks

Rédigé par Clarisse Veron, étudiante en Master 1 Cybersécurité et E-santé

SOMMAIRE

Introduction

La 67ème session du "Lundi de la Cybersécurité" s'est tenue dans un contexte de transformation radicale apportée par l'intelligence artificielle générative (IAG), un domaine qui, depuis la sortie de ChatGPT par OpenAI en novembre 2022, marque non plus une continuité, mais une rupture fondamentale dans notre manière de vivre et de travailler. Cette session, orchestrée par le Professeur Jean-Paul Delahaye, marquée par l'intervention du général d'armée (2S) Marc Watin-Augouard, fondateur du désormais renommé Forum In-Cyber, a plongé dans l'univers fascinant de l'IAG, mettant en lumière les modèles massifs de langage, la capacité de générer des images via IA, et les implications profondes pour la cybersécurité.

Le général Watin-Augouard a partagé sa vision d'un forum de croissance, ancré à Lille, capable d'accueillir jusqu'à 20,000 participants, soulignant l'importance du Forum In-Cyber comme lieu d'échange et de décloisonnement. Il a articulé l'objectif du forum de faciliter le dialogue entre le public, le privé, les militaires, et le monde académique, afin de naviguer ensemble à travers les défis et les opportunités présentés par l'IAG.

Le général a également souligné l'importance du Forum comme espace de décloisonnement et d'échange, où le public, le privé, les militaires, les chercheurs et les étudiants peuvent interagir librement. C'est dans cet esprit que la session s'est concentrée sur l'IAG, envisagée comme un pivot crucial dans la convergence des sciences dures et humaines, et a abordé les défis juridiques émergents posés par l'usage de l'IA, notamment dans la représentation numérique des individus.

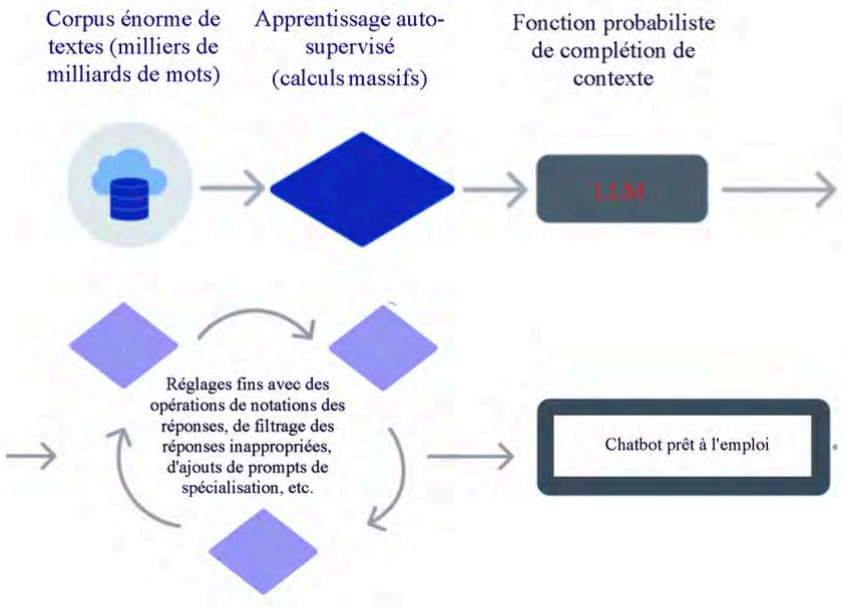
C'est dans ce contexte que le Professeur Jean-Paul Delahaye a pris la parole, offrant une plongée profonde dans l'univers de l'intelligence artificielle générative. Sa présentation, suivant l'introduction par le général Watin-Augouard, a mis en lumière l'impact transformateur des modèles massifs de langages, l'innovation dans la génération d'images par IA, et les défis inédits pour la cybersécurité.

I. Modèles massifs de langages (LLM)

La première partie de la conférence, présentée par le professeur Jean-Paul Delahaye, se focalise sur les modèles massifs de langage (LLM). Ces systèmes d'intelligence artificielle générative représentent un bond qualitatif dans le domaine, illustré notamment par l'arrivée de ChatGPT en novembre 2022. Avec 175 milliards de paramètres, ChatGPT offre un niveau de qualité de traitement du langage naturel supérieur à tout ce qui existait auparavant. Le terme "ChatGPT" combine l'idée de discussion («to chat») avec les notions de transformation générative et de pré-entraînement (Generative Pre-trained Transformer).

Les LLM se distinguent par leur capacité à générer des réponses à des questions couvrant un large éventail de sujets, à traduire ou résumer des textes, à composer des courriers, poèmes, contes pour enfants, et même à créer des images, de la musique, ou des vidéos. Leur impact économique est déjà considérable, touchant des professions variées telles que les enseignants, les traducteurs, les journalistes, les illustrateurs, les programmeurs, les écrivains, et d'autres types de rédacteurs.

La base de ces systèmes est l'utilisation de réseaux de neurones et de méthodes d'apprentissage profond pour construire une fonction probabiliste capable de proposer une suite crédible à tout « contexte » présenté. Le processus inclut un apprentissage non supervisé, où de gigantesques corpus textuels sont utilisés pour entraîner les modèles sans nécessité d'annotation manuelle par des humains. Ces corpus comprennent des milliers de milliards de mots, bien au-delà de ce qu'un humain peut rencontrer dans sa vie, illustrant le besoin d'une quantité massive d'informations pour former ces systèmes.



Ces modèles génératifs sont aussi soumis à des réglages fins ou "fine tuning", incluant des interventions humaines pour évaluer et classer les réponses générées, ajustant ainsi les systèmes dans la direction souhaitée. Cette phase de réglage fait appel à des techniques comme le "Renforcement de l'apprentissage par réactions humaines" (RLHF), mobilisant des milliers d'heures de travail humain et conduisant à des coûts de développement s'élevant à des millions de dollars.

Un aspect remarquable des LLM est leur capacité à créer des dialogues persuasifs et informatifs, se comportant comme s'ils possédaient une compréhension générale du monde, bien qu'ils n'apprennent ni grammaire ni règles d'accord de manière explicite. Cependant, malgré ces avancées, les LLM restent des boîtes noires dont le fonctionnement précis échappe à la compréhension humaine, soulevant des questions sur leur fiabilité et leur capacité à raisonner ou à comprendre la réalité de manière significative.

En conclusion, les modèles massifs de langage marquent une avancée importante dans le domaine de l'intelligence artificielle générative, avec un potentiel transformateur dans de nombreux secteurs. Cependant, leur développement complexe, coûteux et en grande partie opaque soulève des défis significatifs en termes de compréhension, de contrôle et d'éthique.

II. Exemples de dialogue

Dans la seconde partie de sa présentation, le Professeur Jean-Paul Delahaye a offert une exploration captivante des capacités des modèles massifs de langage à travers divers exemples de dialogue, illustrant la puissance et parfois les limites de ces technologies. À travers un éventail de chatbots, dont ChatGPT, il a démontré comment ces outils peuvent générer des réponses variées, mettant en lumière leur utilité, leur créativité, mais aussi leurs imperfections.

Catégorie 1 : Les Réponses Convaincantes

Le professeur a commencé par mettre en avant la capacité des chatbots à fournir des informations précises et pertinentes sur des sujets factuels. Par exemple, interrogés sur la ville la plus peuplée de France, les chatbots ont correctement identifié Paris et fourni des détails sur sa population, illustrant leur utilité comme outils de recherche rapide.

Catégorie 2 : La Créativité à l'Œuvre

Ensuite, le focus a été mis sur leur capacité à générer du contenu créatif. Le professeur a partagé l'exemple d'une histoire très courte impliquant un chat, un avion et une femme, démontrant comment le chatbot a tissé une narrative cohérente et engageante, révélant ainsi le potentiel de ces systèmes pour la création littéraire et artistique.

Catégorie 3 : Les Limites de la Compréhension

Toutefois, ces systèmes ne sont pas sans failles. Le Professeur Delahaye a exposé des situations où les chatbots ont fourni des réponses incorrectes ou déroutantes à des questions logiques ou techniques, comme des problèmes de logique élémentaire ou des questions de physique simple. Ces erreurs mettent en évidence les limites de leur compréhension du monde et la prudence nécessaire lors de leur utilisation pour des tâches nécessitant une précision absolue.

Le Professeur a également souligné l'aspect aléatoire des réponses générées par ces systèmes, mentionnant que des tentatives répétées peuvent aboutir à des résultats différents. Cette caractéristique soulève des questions intéressantes sur la fiabilité et la prévisibilité des IA génératives.

En conclusion de cette section, le Professeur Delahaye a partagé des observations sur l'évolution constante de ces outils. Malgré leurs imperfections actuelles, les avancées continues dans le domaine laissent présager une amélioration significative de leur précision et de leur utilité.

En somme, cette partie de la présentation a offert un aperçu fascinant des capacités et des limites des chatbots et des modèles massifs de langage, soulignant à la fois leur potentiel transformateur et les défis qui restent à surmonter.

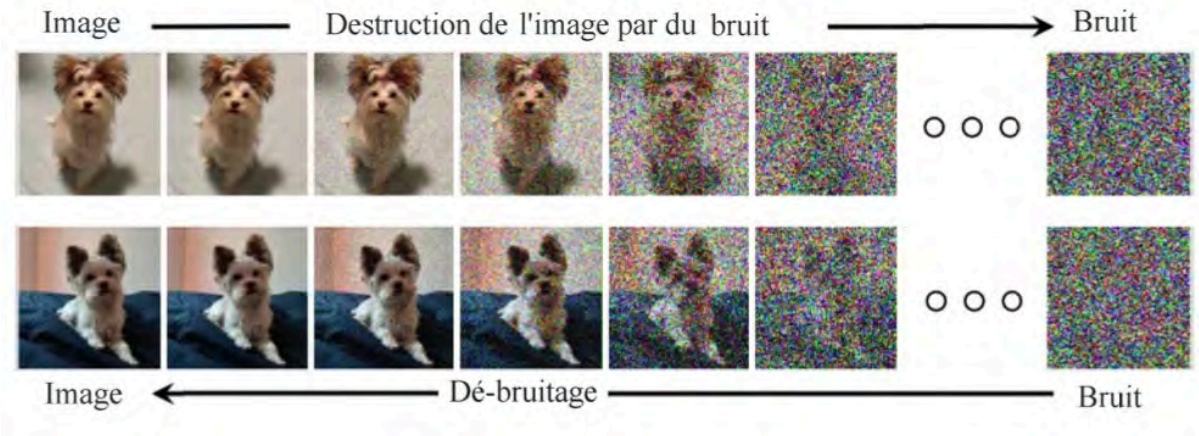
III. IA génératrice d’images

La troisième partie de la présentation du Professeur Jean-Paul Delahaye, intitulée "L'IA génératrice d'images", met en lumière l'une des avancées les plus remarquables et surprenantes dans le domaine de l'intelligence artificielle générative. Cette section commence par exprimer l'étonnement du professeur face à la capacité de ces IA à produire des images de haute qualité à partir de simples descriptions textuelles, soulignant l'aspect presque "miraculeux" de ces technologies, comme Dall-E, Midjourney, et Stable Diffusion.

Le principe de fonctionnement de ces IA repose sur l'alimentation de vastes quantités d'images annotées, permettant à l'IA de déduire les associations que les humains établissent entre les mots et les images. Ce processus d'apprentissage conduit à une forme de "compréhension" par la machine, lui permettant de transformer les descriptions textuelles en visuels impressionnants.

Le professeur Delahaye illustre ce principe avec le réseau principal de Dall-E, CLIP (Contrastive Language-Image Pre-training), développé par OpenAI. CLIP apprend en associant des images à leurs descriptions, ajustant ses paramètres pour valider ces associations. Avec un corpus de 400 millions de couples texte-image utilisés pour son entraînement, CLIP crée un espace sémantique commun pour le texte et l'image, permettant de générer des images qui correspondent étroitement aux descriptions textuelles données.

Pour la génération d'images à partir de textes complètement bruités ou de nouvelles instructions, le système utilise la méthode de diffusion probabiliste ou de débruitage. Cette approche, qui a émergé récemment, permet à l'IA de partir d'une image entièrement bruitée et, guidée par le texte, de produire une image détaillée et affinée qui correspond à la description.



Le professeur fournit plusieurs exemples saisissants de ce que ces technologies peuvent accomplir, démontrant leur capacité à créer des images sur des sujets variés et avec des styles spécifiques sur demande. Ces exemples montrent non seulement l'incroyable capacité des IA à comprendre et à créer à partir de descriptions textuelles mais aussi leurs limites, notamment dans la représentation précise des détails anatomiques comme les mains.

Exemples de résultats obtenus avec le site https: //www.bing.com/images/create/ qui utilise Dall-E :

Une course de Pères noël à vélo dans un désert

L'explosion d'un avion de ligne au-dessus d'une plage

En conclusion, cette partie de la présentation souligne l'impact profond de l'IA génératrice d'images sur notre perception de la créativité et de l'intelligence artificielle, tout en mettant en avant les défis éthiques et techniques que ces avancées représentent.

IV. IAG et cyber-sécurité

La quatrième et dernière partie de la conférence sur l'Intelligence Artificielle Générative (IAG) et la cybersécurité, présentée par le Professeur Jean-Paul Delahaye, aborde les implications profondes des technologies d'IA générative dans le domaine de la cybersécurité. Cette section souligne la dualité de l'IAG, capable à la fois d'innovations bénéfiques et de nouveaux vecteurs d'attaque pour les cybercriminels.

Le professeur Delahaye commence par souligner l'aisance avec laquelle les IA génératives peuvent être utilisées pour créer des contenus faux ou trompeurs, tels que des images ou des vidéos manipulées, facilitant ainsi les opérations de phishing et d'autres formes de cyberattaques sophistiquées. Ces technologies ouvrent la voie à une automatisation accrue des attaques, permettant aux cybercriminels de cibler leurs victimes de manière plus personnalisée et élaborée, rendant les menaces plus difficiles à détecter et à contrer.

Un point clé de cette partie est la référence à une étude qui a examiné 212 cas d'utilisation d'IA générative pour des attaques cybernétiques. Cette étude révèle que les IAG peuvent considérablement faciliter la conception et l'exécution d'attaques, telles que la création de malwares, de logiciels espions, ou de campagnes d'hameçonnage personnalisées, en rendant ces attaques plus efficaces et plus convaincantes pour les victimes potentielles.

Le professeur Delahaye aborde également les mesures prises par les développeurs d'IA pour limiter l'usage malveillant de ces technologies, notamment par l'introduction de filtres et de garde-fous dans les systèmes pour empêcher la génération de contenus nocifs ou illégaux. Cependant, il met en évidence l'ingéniosité des cybercriminels qui trouvent des moyens de contourner ces mesures de sécurité, comme le montre un exemple où des instructions codées sont utilisées pour inciter l'IA à produire des réponses qui auraient normalement été bloquées.

En conclusion, la présentation met en lumière l'importance cruciale de développer des stratégies de cybersécurité robustes et adaptatives pour faire face aux défis posés par l'IA générative. Il est essentiel que la communauté de la cybersécurité, ainsi que les développeurs d'IA, travaillent ensemble pour anticiper et neutraliser les menaces potentielles, garantissant ainsi que les avantages de l'IAG puissent être exploités de manière sécuritaire et éthique.

V. Séance de Questions - Réponses

La session de questions-réponses a exploré divers aspects de l'intelligence artificielle générative (IAG), notamment ses capacités et ses limites, ainsi que son impact potentiel sur le futur de la prise de décision humaine et la cybersécurité. Les intervenants, dont le professeur Jean-Paul Delahaye, ont discuté de la nature de l'intelligence et de la distinction entre les capacités humaines et ce que les IA peuvent réaliser.

Question 1 : La première question a souligné l'importance de ne pas confondre la maîtrise de connaissances disponibles par l'IA avec les processus décisionnels et émotionnels humains. Le futurologue a posé la question de la place de l'IA dans la prise de décision, soulignant que, bien que l'IA puisse traiter de vastes quantités de données, elle ne se substitue pas à la complexité des processus décisionnels humains.

Le professeur Delahaye a répondu en soulignant l'évolution surprenante de l'IA ces dernières années, notamment avec le développement de systèmes de dialogue comme GPT. Il a conservé une position ouverte quant à la possibilité future de créer une IA vraiment équivalente à l'intelligence humaine, rappelant que l'impossible d'hier pourrait devenir le possible de demain.

Question 2 : Cette question a porté sur le seuil de compréhension du sens derrière les phrases construites par l'IA, et si l'IA pourrait un jour franchir ce seuil pour approcher une compréhension semblable à celle des humains.

Le professeur Delahaye a exprimé une vision similaire, indiquant que l'ajout de modèles et de concepts pourrait permettre aux systèmes d'IA de réaliser un raisonnement plus complet, bien qu'il reste sceptique quant à la capacité actuelle de l'IA de remplacer complètement l'intelligence humaine.

Question 3 : La dernière question a exploré l'idée d'une intelligence artificielle non humaine capable de comprendre et de prédire des phénomènes naturels ou sociétaux que nous considérons actuellement comme aléatoires ou incompréhensibles.

Le professeur a mentionné des recherches où l'IA a été utilisée pour des pronostics politiques, montrant que dans certains domaines, l'IA peut déjà surpasser l'intelligence humaine dans la capacité de traitement et de raisonnement.

Conclusion

En conclusion, la 67ème session du "Lundi de la Cybersécurité" a mis en exergue l'impact monumental et les défis inédits introduits par l'intelligence artificielle générative. Sous la direction éclairée du Professeur Jean-Paul Delahaye, avec une ouverture remarquable apportée par le général d'armée (2S) Marc Watin-Augouard, les discussions ont traversé les horizons de l'IAG, de ses prouesses techniques aux questionnements philosophiques et éthiques qu'elle soulève.

La session a révélé l'étendue et la profondeur des modèles massifs de langages, illustrant comment ces outils peuvent non seulement imiter la conversation humaine avec une précision stupéfiante mais aussi générer des œuvres créatives qui défient notre compréhension traditionnelle de l'intelligence artificielle. La capacité des IA génératrices d'images à créer à partir de simples descriptions textuelles ouvre des perspectives fascinantes pour les créateurs de contenu, tout en posant des questions sur l'originalité et la propriété intellectuelle.

La discussion sur les implications de l'IAG en matière de cybersécurité a souligné une arène de double tranchant où les mêmes technologies qui promettent de révolutionner nos vies posent également de nouveaux risques en termes de cyberattaques plus sophistiquées et personnalisées. La séance de questions-réponses a ensuite brillamment capturé le dialogue entre l'enthousiasme pour les possibilités qu'offre l'IAG et une prise de conscience prudente de ses limites et de ses dangers.

Dans l'ensemble, cette session a non seulement mis en lumière la puissance transformatrice de l'IAG mais a également invité à une réflexion profonde sur l'avenir de l'intelligence, qu'elle soit artificielle ou humaine. En naviguant entre l'admiration pour les avancées technologiques et la vigilance face aux défis éthiques qu'elles posent, cette session rappelle l'importance cruciale d'un dialogue continu et d'une exploration collaborative entre les scientifiques, les penseurs, et la société dans son ensemble pour façonner un avenir où l'intelligence artificielle générative enrichit l'humanité sans compromettre ses valeurs fondamentales.