Exercice 1 : Risques d’utilisation d’un agent conversationnel

Consignes à l’enseignant : Constituer six groupes d’étudiants de 3 ou 4 étudiants et distribuer à chacun des groupes un cas d’étude.

Cas 1 — L'avocate[1]

Un juge de la Cour suprême de la Colombie-Britannique a réprimandé une avocate ayant inclus deux cas fictifs générés par l'intelligence artificielle dans une requête déposée en décembre dernier.

Les deux cas en question générés par ChatGPT n'ont, en fin de compte, jamais été utilisés dans les arguments de l'avocate, puisqu'ils ont été retirés lorsqu'elle a appris qu'ils n'avaient jamais existé.

Dans une décision rendue lundi, le juge David Masuhara a précisé qu'il ne pensait pas que l'avocate Chong Ke avait l'intention de tromper le tribunal. Il a néanmoins avoué avoir été troublé. […]

L'avis de demande citait deux affaires : l'une dans laquelle une mère avait emmené son « enfant, âgé de 7 ans, en Inde pendant six semaines, et l'autre dans laquelle avait été autorisée une demande d'une mère de voyager avec son enfant, âgé de 9 ans, en Chine pendant quatre semaines pour rendre visite à ses parents et à ses amis ».

« Ces affaires sont au cœur de la controverse, car on a découvert qu'elles n'existaient pas », écrit le juge Davi Masuhara. […]

Le juge Masuhara a expliqué que l'avocate avait ensuite fait une déclaration sous serment, soulignant son manque de connaissance des risques liés à l'utilisation de ChatGPT. Elle évoque le moment où elle a découvert que les cas étaient fictifs, le décrivant comme étant terriblement gênant.

Je n'avais pas l'intention de générer ou de faire référence à des cas fictifs dans cette affaire. C'est clairement une erreur et ce n'est pas quelque chose que je ferais en toute connaissance de cause, s'est défendue Chong Ke dans sa déposition.

Cet incident semble constituer une première au Canada. Un cas similaire avait fait les gros titres aux États-Unis l'année dernière. Un avocat de Manhattan avait demandé la clémence d'un juge fédéral après avoir déposé un dossier reposant uniquement sur des décisions dont il a appris par la suite qu'elles avaient été inventées par ChatGPT. […]

Le juge a toutefois estimé que Chong Ke devrait supporter les coûts des mesures entreprises par les avocats de Nina Zhang pour remédier à la confusion.

[1] Radio-Canada. (2024, 27 février). Une avocate réprimandée pour avoir cité des cas fictifs générés par ChatGPT. *INFO Radio Canada Colombie-Britannique*. https://ici.radio-canada.ca/nouvelle/2052522/ia-chatgpt-justice-avocate-reprimande

Cas 2 — L'ado[2]

La société californienne, fondée par d'anciens ingénieurs de Google, fait partie des entreprises qui proposent des compagnons IA — des robots conçus pour converser et divertir, capables d'interagir de façon similaire à des humains en ligne.

Dans un procès intenté en Floride en octobre, une mère a affirmé que la plateforme était responsable du suicide de son fils de 14 ans.

L'adolescent, Sewell Setzer III, avait noué une relation intime avec un robot conversationnel inspiré du personnage de *Game of Thrones* Daenerys Targaryen, et avait évoqué un désir de mettre fin à ses jours.

Selon la plainte, le robot l'a encouragé à passer à l'acte, répondant « Je vous en prie, mon doux roi » lorsqu'il a dit qu'il « partait au paradis » avant de se suicider avec l'arme de son beau-père.

L'entreprise « s'est donnée beaucoup de mal pour créer chez Sewell, 14 ans, une dépendance néfaste à ses produits, elle l'a abusé sexuellement et émotionnellement, et n'a finalement pas offert d'aide ni prévenu ses parents lorsqu'il a exprimé des idées suicidaires », accusent les avocats de la mère.

Une autre plainte déposée lundi au Texas concerne deux familles qui affirment que le service a exposé leurs enfants à des contenus sexuels et les a encouragés à s'automutiler. […]

[2] Agence France-Presse. (2024, 12 décembre). Character.AI prend des mesures de sécurité. *La Presse*. https://www.lapresse.ca/affaires/techno/2024-12-12/suicide-d-un-utilisateur-mineur/character-ai-prend-des-mesures-de-securite.php

Cas 3 — L'animateur radio[3]

La presse polonaise rapporte que la radio en ligne Off Krakow a dû mettre fin prématurément à des émissions gérées par l'intelligence artificielle et qu'elle avait commencé à diffuser de manière expérimentale. En effet, la radio s'était lancée dans l'aventure de l'IA la semaine dernière, suscitant l'émoi dans le pays.

C'en est donc fini pour Emi, Kuba et Alex, les trois animateurs de Off Radio Krakow, qui avaient commencé leur aventure radiophonique le 22 octobre. Ces trois avatars de la génération Z, entièrement générés par l'intelligence artificielle, ont pendant une semaine principalement diffusé "de la musique à destination des jeunes auditeurs", remarque le média en ligne Press.pl, dédié au journalisme. […]

Sauf que l'expérimentation, qui voulait sonder, trois mois durant, "les effets que le développement de l'intelligence artificielle peut avoir sur la culture, les médias, le journalisme et la société", a dû s'arrêter net sous les feux de la critique. […]

En cause ? Le licenciement "quelques semaines plus tôt [...] des animateurs" de la radio, rapporte la chaîne d'information privée TVN24. […]

Autre polémique : la radio en ligne, issue du service public, a également diffusé un entretien avec Wyslawa Szymborska, prix Nobel de littérature en 1996, et décédée six ans plus tard. "La voix", comme les propos de la poète cracovienne, "ont été créés à l'aide de l'intelligence artificielle", précise le site d'information en ligne Wirtualna Polska. D'autres médias, comme Press.pl, avancent que des sons d'archives ont par ailleurs été utilisés. […]

"L'IA n'a pas été utilisée pour des activités simples et répétitives qui pourraient faciliter le travail à la radio, mais pour créer du contenu sur des sujets culturels et idéologiques très sensibles", estime Malgorzata Fraser, une autrice de podcast interrogée par Press.pl.

[3] Courrier international. (2024, 30 octobre). En Pologne, tollé autour de l'utilisation de l'intelligence artificielle à la radio. *Courrier international*. https://www.courrierinternational.com/article/medias-en-pologne-tolle-autour-de-l-utilisation-de-l-intelligence-artificielle-a-la-radio_223973

Cas 4 — L'auteur[4]

Se fier aveuglément aux conseils d'une intelligence artificielle peut réserver quelques mauvaises surprises. Mark Pollard, auteur australien en stratégie marketing, en a récemment fait les frais après avoir sollicité ChatGPT, le robot conversationnel d'OpenAI. Comme le rapporte Tech & Co , ce vendredi, l'homme n'a pas pu assister à une conférence au Chili à cause d'une mauvaise information fournie par l'IA.

L'homme s'est retrouvé bloqué à l'aéroport sans pouvoir entrer sur le territoire chilien. Il pensait en effet qu'il n'avait pas besoin de visa pour un court séjour, comme le lui avait indiqué ChatGPT, mais c'était faux. […]

De leur côté, nos confrères de Tech & Co ont formulé auprès de ChatGPT la même demande que Mark Pollard. Et le robot a une nouvelle fois assuré que le visa pour le Chili n'était pas indispensable pour un Australien s'il s'agit d'un court séjour. […]

L'Australien, lui, estime aujourd'hui qu'il aurait mieux valu passer par Google, mais déplore les « mauvais » résultats des moteurs de recherche, et notamment « les publications sponsorisées ». « Ce n'est vraiment pas agréable. ChatGPT, c'est plus simple », a-t-il déclaré, tout en assurant qu'il ne l'utiliserait plus que pour son planning désormais, évitant notamment d'y avoir recours pour « les choses très importantes ».

[4] Ouest-France. (2025, 28 mars). ChatGPT lui donne une fausse information, il se retrouve bloqué à l'aéroport et rate sa conférence. *Ouest-France*. https://www.ouest-france.fr/high-tech/intelligence-artificielle/chatgpt-lui-donne-une-fausse-information-il-se-retrouve-bloque-a-laeroport-et-rate-sa-conference-b11741de-0bc9-11f0-ae86-6d5913869474

Partie 1 (travail individuel et en groupe, 15 minutes)

Consignes aux étudiants : Lire de manière individuelle le cas d'étude affecté par votre enseignant, puis, en groupe, réfléchir et répondre aux questions suivantes.

- Question 1 : Quelles sont les limites de l'agent conversationnel exposées dans cette situation ?
- Question 2 : Le contexte d'utilisation de l'agent conversationnel vous semble-t-il approprié ? Pourquoi ou pourquoi pas ?
- Question 3 : Quel aurait été un usage plus responsable ou plus sécuritaire de l'agent conversationnel dans le contexte présenté ?

Partie 2 (restitution en plénière, 15 minutes par affaire soient 60 min)

Consignes à l'enseignant : Guider la restitution et animer la discussion au travers des questions. Chaque groupe résume l'affaire pour les autres avant de répondre aux questions à l'oral. Noter les réponses au tableau pour une vue globale.

Exercice 2 : Observer les biais des LLMs que vous utilisez

Consignes à l'enseignant : Avant de commencer, présenter les limites (biais...) des agents conversationnels et des humains et les risques qui peuvent en découler.

On parle de biais dans un LLM lorsqu'il manifeste une tendance à traiter différemment des individus selon leur appartenance à un groupe défini par des critères tels que la culture, la nationalité, le genre, l'âge, la langue, la religion, l'origine ethnique, la localisation, le niveau de revenu ou la catégorie socio-professionnelle... Ces biais peuvent se traduire par la propagation de normes implicites, le renforcement de stéréotypes sociaux ou encore la génération de contenus discriminants.

La détection des biais repose sur l'analyse des réponses fournies par un agent conversationnel à des requêtes liées à une norme implicite ou à un stéréotype présumé. Il s'agit

1) d'observer ses réponses par rapport à une demande formulée dans un contexte neutre,

2) puis de comparer les réponses par rapport à une demande formulée pour des individus de groupes différents selon un critère.

3) La comparaison de réponses produites par des agents différents aident aussi à relever la présence de biais dans les réponses de certains agents.

Partie 1 (15 minutes, travail en individuel puis en groupe)

Consignes aux étudiants : Le cas d'étude est présent dans un tableur numérique. Chaque cas se compose de 3 amorces (une neutre et deux avec des contextes spécifiques) et de réponses produites par 3 agents conversationnels différents (Qwen, Grok et ChatGPT) ; un lien accompagne un copier-coller de la réponse. Confronter vos observations individuelles au sein de votre groupe.

- Question 1 : Dans les réponses neutres produites par les 3 agents, les conseils prodigués conviendraient-ils à tout le monde dans votre classe ? Pourquoi certaines personnes auraient-elles plus de difficulté à les suivre ?

- Question 2 : Est-ce qu'il y a des choses, des personnes ou des façons de vivre qui n'apparaissent jamais dans les réponses neutres ? Pourquoi à votre avis ?

- Question 3 : En considérant les réponses produites aux trois configurations par les 3 agents, quels comportements spécifiques observez-vous chez chaque agent ?

- Question 4 : Êtes-vous en mesure d'identifier les valeurs ou les normes véhiculées dans les réponses produites par certains agents ? Identifiez-vous un discours en faveur/défaveur d'une norme sociale ? D'un stéréotype de groupe d'individu ? Lesquels ? Le discours est-il nocif/toxique ?

Partie 2 (60 minutes, restitution en plénière)

Consignes à l'enseignant : Inviter chaque groupe à présenter d'abord tous leurs prompts pour que tout le monde les ait en tête puis chacun des groupes répond aux questions. Noter les réponses au tableau pour offrir une vue globale. Animer la discussion.

- *D'abord tous les groupes restituent leur réflexion pour la 1 et la 2 (20 min.).*
- *Puis tous les groupes rapportent sur la 3 (10 minutes).*
- *Enfin la question 4, agent par agent, chaque groupe s'exprime (20 minutes).*

Exercice 3 : De l'anthropomorphisme ?

Les agents conversationnels ou chatbot sont conçus pour converser avec un humain en langage naturel en simulant le comportement d'un être humain en tant qu'interlocuteur.

L'exercice suivant permet de s'exercer à repérer des caractéristiques propres à l'humain dans les productions de l'agent ainsi que celles spécifiques à la machine.

Partie 1 (10 minutes, travail en individuel)

Consignes à l'étudiant : chaque groupe travaille sur les conversations sur lesquelles il a déjà travaillé dans l'exercice précédent. Répondre aux questions.

- Question 1: Avez-vous eu l'impression que l'agent conversationnel « parlait comme un humain » ? Pourquoi ?

- Question 2 : Relevez les caractéristiques de comportement propre à l'humain au sein des réponses produites. Relevez aussi les caractéristiques propre à la machine.

- Question 3 : Qu'est ce qui conduit l'agent conversationnel à générer ces marques ? Pourquoi est-il important de reconnaître ces marques d'imitation dans le contexte scolaire ou professionnel ?

Partie 2 (5 minutes, restitution en plénière)

Consignes à l'enseignant :
Pour la première question, sonder à main levée. Demander pourquoi permet de glisser à la seconde question (qqs minutes)

Pour la seconde question, guider la restitution des caractéristiques. L'idée n'est pas d'être exhaustif mais d'avoir une vision globale avec des exemples attestés.
Demander aussi les caractéristiques propres à la machine. (15 minutes)

La question 3 est une question d'ouverture. Faire le point sur les risques.

Une diapositive est disponible pour faire une synthèse.

Exercice 4 : Application de la démarche CRISTAL

Extrait de texte proposé[5] :

« Cela pose plusieurs enjeux majeurs pour les élèves et les enseignants. D'abord pour les élèves, quand ils utilisent par exemple l'IA générative pour les devoirs à la maison, le risque est que ces outils soient utilisés d'une manière qui court-circuite l'effort cognitif nécessaire à un apprentissage efficace (Kasneci et al., 2023 ; Abdelghani et al., 2023). Plus précisément, l'image de "super- intelligence" véhiculée dans de nombreux médias, combinée au ton assuré des logiciels d'IAGen (alors qu'ils sont incapables de métacognition, c'est-à-dire incapables d'évaluer leurs propres incertitudes), peut amener de nombreux élèves à surestimer à la fois les compétences des IAGens et leurs propres compétences, limitant le développement et l'expression de leur curiosité, de leur esprit critique et de leur métacognition qui sont pourtant essentiels à des apprentissages efficaces et motivants (Abdeghani et al., 2023 ; Oudeyer et la., 2016). Ces effets sont amplifiés par l'absence de posture pédagogique dans le comportement des IAGen : en effet, ils ont été entraînés à prédire les mots et les images les plus probables, ainsi qu'à répondre le plus directement et efficacement aux questions des utilisateurs. En conséquence, quand un élève leur pose une question ou leur donne un exercice, ils vont avoir une très forte tendance à donner tout de suite la réponse, au lieu de donner des indices pédagogiques pour aider et permettre à l'apprenant de faire l'effort de trouver par lui-même la réponse (Macina et al, 2023 ; Jurenka et al., 2024).

Ces défis sont à la fois liés à la nature des logiciels d'IA générative, mais aussi aux biais cognitifs des élèves apprenants et à leur compréhension limitée de ces systèmes (Kidd and Birhane, 2023). D'abord, le cerveau humain a tendance à utiliser son estimation du niveau de compétence des autres individus pour décider quelles informations et croyances adopter ou questionner quand ces individus les expriment (Orticio et al., 2023). Par ailleurs, les humains ont aussi des biais cognitifs favorisant l'attribution d'agentivité et d'intention à des objets (Heider and Simmel, 1944), et cela s'applique en particulier aux IAGen. Enfin, le cerveau humain est aussi biaisé de telle manière qu'une fois que certaines connaissances ou croyances ont été acquises à partir de sources qu'il croyait solides, il est ensuite difficile de corriger ces croyances (Thompson and Griffiths, 2021). Ces trois biais combinés amènent ainsi les humains à avoir tendance à attribuer de l'agentivité aux IAGen, à penser que leurs connaissances sont solides étant donné leur ton assuré et affirmatif, et à ainsi potentiellement apprendre des informations fausses, ce qui est particulièrement problématique étant donné que les IAGens encodent de nombreux stéréotypes (e.g. raciaux, de genre, religieux, etc.) (Bender et al., 2021). »

Partie 1 (5 minutes, travail en individuel)

A) Lisez l'extrait de texte remis.

---

[5] Oudeyer, P.-Y. (2024, octobre). *IA générative, société et éducation: En quoi l'IA générative représente-elle un enjeu dans la formation des citoyens ?* Conférence de consensus « Nouveaux savoirs et nouvelles compétences des jeunes » du Cnesco. https://inria.hal.science/hal-04945894v1

B) Rédigez un résumé personnel de 3 à 5 phrases.

Partie 2 (5 minutes, travail en individuel)

C) Demandez ensuite à l'agent conversationnel de votre choix de le résumer.
D) Relancez jusqu'à obtenir le résultat que vous désirez.
E) Comparez les deux résumés et identifiez les différences notables. Quels éléments souhaitez-vous conserver de votre résumé personnel ? Quels éléments souhaitez-vous conserver du résumé généré par l'IAg ?

Partie 3 (5 minutes, travail en groupe)

F) Appliquez la méthode CRISTAL à l'aide du tableau suivant. Pour chaque principe, notez les questions concrètes que vous vous êtes posées et confrontez les avec celles du tableau. Pour chaque question, précisez les vérifications effectuées.

Principes	Les questions que je me pose…	Les vérifications que j'ai faites…
Critique	Ex. : *Le résumé de l'IAg reflète-t-il fidèlement le texte original ? Y a-t-il des erreurs factuelles, des généralisations ou des oublis ? L'IAg a-t-elle ajouté des éléments inexistants ?*	
Responsable	Ex. : *Ai-je utilisé un outil d'IAg sécuritaire ? Ma requête a-t-elle influencé le résultat ? Le résumé est-il neutre ? Certaines idées sont-elles absentes ou surreprésentées ?*	
Intègre	Ex. : *Est-ce que j'ai respecté les règles du cours sur l'IAg ? Le texte utilisé pouvait-il être téléversé ? Ai-je utilisé l'IAg pour m'aider, ou pour produire à ma place ? Ai-je appris ?*	
Sobre	Ex. : *L'IAg était-elle réellement utile ? Ai-je choisi un outil d'IAg adapté pour résumer ? Ai-je limité les interactions avec l'outil ?*	
Transparent	Ex. : *Si j'ai repris des éléments du résumé de l'IAg, ai-je indiqué son usage clairement ? Le résumé final permet-il d'évaluer mes compétences ? Ai-je bien conservé les traces de mon processus de rédaction ?*	
Autonome	Ex. : *Ai-je gardé le contrôle sur mes choix et sur le contenu final du résumé ? Ai-je évité de parler à l'IAg comme à une personne réelle ?*	
Libre et créatif	Ex. : *Est-ce que j'ai pris le temps de réfléchir par moi-même ? Ai-je modifié ou rejeté les éléments qui ne reflétaient pas ma compréhension ?*	

Exercice 5 : Application de la démarche CRISTAL

1. réfléchir librement en groupe sur comment construire une mesure quantifiable/qualifiable CRISTAL (par dimension et/ou global)
   (⅔-¾ du temps restant)

2. restituer et discuter les idées (⅓-¼ du temps restant)