En quoi l'IA générative représente-elle un enjeu dans la formation des citoyens ?

PIERRE-YVES OUDEYER
INSTITUT NATIONAL DE RECHERCHE EN SCIENCES
ET TECHNOLOGIES DU NUMÉRIQUE (INRIA)
UNIVERSITÉ DE BORDEAUX

Introduction

Alors qu'ils sont apparus très récemment, les grands systèmes d'IA générative ont déjà aujourd'hui un impact majeur sur la société, dans les domaines culturels, politiques, économiques, environnementaux et éducatifs. Leurs usages se développent en particulier très vite et massivement chez les jeunes, que ce soit dans le contexte scolaire ou en dehors. La vitesse de ce développement est telle que les études scientifiques permettant de mieux comprendre les usages et leurs impacts sont encore très rares car elles nécessitent un temps incompressible de mise en place et de vérification. On peut dire que globalement, d'un point de vue scientifique, c'est une *terra incognita :* on sait peu et nombreuses sont les questions ouvertes. Néanmoins, ce qui est sûr, c'est que la formation des futurs citoyens aux mécanismes et aux enjeux cognitifs, sociaux et culturels de l'IA générative est un enjeu majeur : comment ces systèmes fonctionnent-ils ? Comment bien les utiliser ? Quels sont les défis (par exemple : désinformation, uniformisation de la pensée, emplois, coûts environnementaux) et les opportunités (par exemple : aide aux apprentissages et à la création, accès et diffusion de cultures diverses et pour des personnes diverses, amélioration de la productivité) pour la société ?

I. Qu'est-ce-que l'IA générative (IAGen) ?

A. Une courte histoire de l'intelligence artificielle

Alors que nombreux sont les collégiens et lycéens à penser que l'intelligence artificielle (IA) désigne des systèmes qui sont apparus il y 5 ou 10 ans, c'est un terme et un domaine scientifique beaucoup plus anciens, et l'IA générative (IAGen) désigne seulement une forme d'IA parmi d'autres. L'expression « intelligence artificielle » a été inventée en 1955, peu de temps après l'invention des ordinateurs, pour désigner un domaine scientifique qui étudie les mécanismes de la cognition et de l'apprentissage des êtres vivants (en particulier des humains) en utilisant des modèles informatiques, c'est-à-dire en simulant certains aspects de ces mécanismes. Aujourd'hui, beaucoup de gens utilisent le mot « IA » comme un raccourci linguistique pour désigner les systèmes ou machines fabriquées par les chercheurs et ingénieurs de ce domaine, et nous ferons parfois aussi ce raccourci dans ce document.

Il y a plusieurs formes de systèmes d'IA, et certaines sont présentes depuis longtemps dans notre quotidien (Russell & Norvig, 2016). L'IA comportementale modélise le rôle de l'interaction entre certains réflexes sensorimoteurs, le corps et l'environnement : c'est le cas par exemple des robots Elmer et Elsie de Gray Walter, créés en 1949 (Walter, 1950 ; Cordeschi, 2002), et dont les réflexes d'attirance à la lumière permettaient de reproduire des comportements complexes très similaires à ceux de nombreux insectes. Une limite de cette approche est la difficulté à simuler des processus cognitifs abstraits comme le raisonnement logique ou le traitement du langage. L'IA symbolique modélise un domaine, comme celui du jeu d'échecs ou du jeu de dames, avec un ensemble de symboles et de règles logiques, et utilise des heuristiques programmées à la main par les ingénieurs pour calculer et déduire des plans d'actions optimaux pour résoudre un problème (Samuel, 1959). Une limite de cette approche est son incapacité à s'adapter à des situations nouvelles et non prévues par les ingénieurs. L'IA statistique désigne un ensemble de techniques permettant à une machine d'apprendre des savoir-faire nouveaux en identifiant des régularités dans des données, et au moyen de calculs statistiques (Lecun *et al.*, 2015). On appelle aussi ces techniques « apprentissage automatique ». Par exemple, avec l'apprentissage « supervisé », et si l'on a accès à une base de données associant des photos à des étiquettes indiquant ce qu'il y a sur les photos (par exemple « chat », « chien », « avion », etc.), il est possible de mettre au point un système logiciel qui va pouvoir prédire l'étiquette associée à une nouvelle image. Avec l'apprentissage par « renforcement », on peut permettre à un robot par exemple d'apprendre une nouvelle stratégie de mouvement pour attraper un objet : au départ il essaie des mouvements au hasard et mesure leur score pour attraper l'objet, et il va progressivement raffiner les paramètres des mouvements qui ont le meilleur score (par exemple en essayant des petites variations aléatoires des meilleurs mouvements). Ces différentes techniques d'IA sont déjà présentes depuis longtemps dans de nombreux contextes, par exemple sur internet dans les algorithmes de recommandation d'un restaurant ou d'un film, dans les logiciels d'aide à la conduite des voitures ou des avions, dans les logiciels de logistique, dans les logiciels de reconnaissance vocale ou faciale, ou pour programmer le comportement de personnages non joueurs dans les jeux vidéo. Cependant, jusqu'à il y a 4 ou 5 ans, toutes ces approches étaient encore très limitées dans leurs capacités, en particulier dans la maîtrise de la langue et des tâches qui s'expriment et se résolvent principalement en langue naturelle.

Le domaine de l'IA générative (IAgen) a apporté une évolution majeure qui a repoussé beaucoup de ces limites (Brown *et al.*, 2020), et explique l'impact sociétal dont nous parlerons ci-dessous. L'IA générative est une forme particulière d'IA statistique, qui s'est fait connaître du grand public par la sortie de ChatGPT 3.5 en 2022, puis d'autres logiciels comme Midjourney, DALL-E, Mistral, Claude ou Gemini plus récemment. Ce sont des logiciels capables de produire du texte, des images ou du son à partir de « prompts », c'est-à-dire de textes ou d'images qui sont souvent utilisés pour donner au logiciel un contexte et une instruction ou une question relative à ce contexte. Par exemple, on peut demander à une IAGen des questions telles que « Quels sont les monuments à visiter en priorité à Paris ? » ou « Qu'est-ce-que le big bang ? », et préciser à qui doit s'adresser la réponse, par exemple : « écris une réponse adaptée à des enfants à l'école primaire » ou « écris une réponse pour un adulte ayant une formation scientifique ». Il est possible aussi de donner à une IAGen un fichier pdf, par exemple un cours d'histoire, et de demander par exemple « peux-tu résumer ce cours ? » ou « poses moi des questions pour vérifier que j'ai bien appris ce cours ». Enfin, le résultat peut aussi être une image, par exemple si l'on demande « fait un dessin qui montre un pingouin qui skie sur une piste de neige artificielle dans Paris ».

Jusqu'à récemment, la plupart de ces tâches étaient réalisées d'une manière très médiocre par les logiciels d'IA. Mais leurs capacités ont considérablement et très rapidement augmenté en quelques années, d'une manière qui était imprévisible même pour la plupart des scientifiques travaillant dans ce domaine.

B. Comment ces IAGens fonctionnent-elles et qu'est-ce-qui a permis ces évolutions majeures ?

Les IAGens sont des logiciels qui, étant donné un texte/une image/un son en entrée, font des calculs pour produire un texte/une image/un son en sortie. Ces calculs sont réalisés par des milliards de petits modules élémentaires faisant chacun des calculs relativement simples et interagissant entre eux. Ces calculs sont déterminés par des paramètres internes (des nombres) à chaque petit module. Au départ, ces paramètres sont aléatoires. À partir de là, un algorithme va les faire évoluer au cours d'un « entraînement ». Cet entraînement consiste en deux phases. D'abord, dans une phase d'apprentissage supervisé, on donne au logiciel d'IAGen des textes/images à trous, fabriqués à partir de milliards de textes trouvés par exemple sur internet (et en cachant des bouts de texte ou d'image pour faire les trous). Les IAGens doivent alors prédire quel texte/image placer dans les trous. Chaque fois qu'ils font ce type d'exercice, les paramètres sont renforcés s'ils ont permis de bien deviner quoi mettre dans le trou, et sinon les paramètres sont modifiés un tout petit peu pour augmenter la probabilité de donner la bonne réponse la prochaine fois. Dans une seconde phase d'apprentissage par renforcement (Ouyang *et al.*, 2022), des millions de tâches sont données à faire à des IAGens (par exemple résumer un texte ou répondre à une question), et les productions des IAGens sont notées par des scoreurs humains. À partir de ces scores, les paramètres sont mis à jour pour augmenter la probabilité d'avoir un meilleur score la prochaine fois. En bref, les IAGens sont des logiciels entraînés pour produire les mots et les images les plus probables, et qui auront les meilleures notes des scoreurs humains, étant donnés des prompts sous forme de textes et d'images.

L'entraînement des IAGens consiste donc à apprendre à faire des exercices relativement élémentaires. Si l'on entraîne de cette façon des IAGens qui ont peu de paramètres (quelques millions) et avec peu de textes à trous (quelques millions), alors aucune capacité notable n'apparaît. Si au contraire on entraine des IAGens qui contiennent des milliards de paramètres avec des milliards de textes/images à trous, alors apparaissent assez soudainement des capacités nouvelles et diverses. D'un point de vue scientifique, l'apparition de ces capacités dans ce contexte est encore largement un mystère. Pour mesurer ces capacités, les scientifiques utilisent des bancs d'essais (« *benchmarks* »), qui sont des tests de connaissance, de raisonnements et de savoir-faire dans de nombreux domaines (Srivasta *et al.*, 2023 ; Chang *et al.*, 2024). Certains de ces tests sont même directement des tests faits au départ pour des examens ou concours universitaires pour les étudiants humains. Ainsi, des expérimentations récentes sur des logiciels comme GPT-4, Claude Sonnet 3.5, Llama 3 70B ou Mistral Large ont montré qu'ils étaient capables d'obtenir d'excellentes notes pour des examens universitaires de droit, de mathématiques, ou d'informatique, et qui leur permettrait d'obtenir des diplômes dans ces matières. Si certains de ces examens ou concours comportent des tests de connaissances et nécessitent surtout un apprentissage par cœur, d'autres comportent des exercices qui n'ont jamais été vu par ces logiciels pendant leur entraînement, et nécessitent pour les humains des raisonnements avancés : la capacité des modèles de langage à réaliser ces tâches montre ainsi une forme de généralisation relativement puissante. Cependant, comme nous le verrons ci-dessous, cela peut aussi amener les IAGens à affirmer avec assurance des informations ou des raisonnements qui sont complètement faux.

C. Les IAGens sont des technologies culturelles

Ces processus permettant de fabriquer des logiciels d'IAGens reposent ainsi sur deux piliers : 1) l'apprentissage des régularités apparaissant dans les bases de données de texte/image/son qui permettent de les entraîner ; 2) le *scoring* d'évaluateurs humains. Ainsi, on peut voir les IAGens comme des systèmes qui compressent et encodent les régularités culturelles qui sont présentes dans les corpus de texte/images (et les *feedback* humains) sur lesquels ils sont entraînés (Hershcovich *et al.*, 2022 ; Bender *et al.*, 2021 ; Johnson *et al.*, 2022) : ils les reproduisent alors lorsqu'ils sont utilisés (dans le cadre de discussions directes avec un humain, ou par d'autres machines qui vont ensuite simuler des populations humains, par exemples les bots sur les réseaux sociaux), ce qui amène à leur amplification. Les grands modèles d'IAGen (de langage, visuels, multimodaux) sont donc fondamentalement des outils de transmission culturelle : ce sont en fait des modèles de culture.

Plus précisément, ils encodent des régularités culturelles selon de nombreuses dimensions, allant des valeurs, des connaissances usuelles, des normes socio-culturelles, des répertoires de concepts définissant ce qui est saillant ou pas, intéressant ou pas, vrai ou pas (Hershcovich *et al.*, 2022). Ces régularités incluent toutes formes de biais, c'est-à-dire de stéréotypes, qui peuvent être problématiques pour la société, en particulier pour des groupes de minorités (Bender *et al.*, 2021 ; Johnson *et al.*, 2022).

D. Hallucinations et biais

Alors que leurs comportements et leurs capacités sont souvent impressionnants (par exemple répondre correctement à un problème de mathématiques de niveau universitaire), les mêmes logiciels d'IAGen peuvent aussi faire des erreurs majeures sur des questions élémentaires (par exemple répondre faux à un problème de maths d'école primaire). On parle alors d'hallucinations ou de confabulations. Il y a plusieurs raisons à cela. D'abord, les systèmes d'IAGen sont entraînés sur des textes très divers provenant de sources très variées : beaucoup d'entre eux contiennent des erreurs, mais aussi des biais ou même des informations de propagande volontairement destinés à influencer la pensée de leurs lecteurs : ces erreurs, ces biais et ces propagandes sont ainsi encodés et restitués par les modèles de langage. Une autre raison majeure est que, comme expliqué ci-dessus, ils ne sont pas entraînés pour répondre "juste", mais pour donner les réponses les plus probables étant donné les bases de textes ou d'images utilisées pendant leur entraînement (ou à prédire les réponses qui recevraient probablement de bonnes notes des scoreurs humains). Ainsi, les IAGens n'ont pas de notion intrinsèque de vrai ou de faux (et n'ont pas de métacognition, Guilleray *et al.*, 2024), et par ailleurs sont faites pour essayer de deviner les réponses à des questions nouvelles, qui n'étaient pas exactement présentes telles quelles dans leur entraînement. C'est ce qui leur permet de généraliser, par exemple quand il s'agit de résumer un texte totalement nouveau, et dans ce cas c'est utile. C'est aussi ce qui les amène à inventer de toutes pièces des faits qui n'en sont pas (par exemple, à la question « Combien de buts le PSG a-t-il marqué en finale de la ligue des champions en 2024 ? », un logiciel récent a répondu « 4 buts » alors que le PSG n'était pas qualifié pour la finale).

Au-delà d'erreurs factuelles, ce sont ces mêmes raisons qui amènent les logiciels d'IAGen à reproduire des biais fréquemment présents dans les textes qui servent à les entraîner (par exemple liés au genre, à la race, à la religion, aux métiers, etc.), avec le risque d'en amplifier les conséquences néfastes auprès de populations diverses, en particulier auprès de minorités. De nombreuses approches sont aujourd'hui explorées pour mieux comprendre et limiter ces biais, par exemple en demandant aux scoreurs de mettre des mauvaises notes quand ils observent des productions qui les reflètent. C'est ce qu'on appelle l'« alignement », signifiant qu'on souhaite « aligner » les comportements des IAGens sur les valeurs et les préférences culturelles des groupes d'humains qui les utilisent (Ji *et al.*, 2023). Cela pose de nombreux défis. Un exemple récent permettant d'en comprendre la complexité, est celui d'un système d'IAGen a qui l'on avait demandé de générer une illustration de l'armée nazie pendant la Seconde Guerre mondiale : cette illustration a montré une troupe comprenant des femmes et des personnes dont les couleurs de peau étaient diverses. Dans ce cas, le logiciel d'IAGen a suivi des mécanismes qui le guidaient pour ne pas générer de "biais" dans des représentations (ici liées à l'armée) : ce-faisant, il a produit un résultat qui est une erreur factuelle qui fausse la représentation d'un événement réel. Il serait donc nécessaire que dans ce cas le logiciel puisse faire la différence entre une demande qui concerne un fait historique (notion dont l'objectivité est parfois débattue chez les historiens et les philosophes), et une demande qui n'est pas liée à un fait historique, ce qui est encore loin d'être le cas aujourd'hui. Cet exemple illustre aussi qu'au-delà de problème technique très compliqués (Ji *et al.*, 2023), la question de l'alignement pose aussi des enjeux

politiques forts : quels sont les régularités culturelles que l'on souhaite encoder dans tel ou tel logiciel d'IAGen, et quels sont les biais que l'on souhaite conserver ou au contraire éviter ?

E. *Deepfakes* : désinformation et vie privée.

Les erreurs (hallucinations) des IAgens sont souvent involontaires. Cependant, il est aussi possible d'utiliser ces logiciels pour volontairement créer des textes ou des images qui mettent en scène des personnes ou des lieux réels, mais dans des situations qui n'ont pas existé. Dans le cas des images ou des vidéos, la qualité des logiciels d'IAGens permet aujourd'hui de générer des scènes pour lesquels il est quasi-impossible de deviner s'il s'agit d'une vraie image ou d'une image inventée : on parle alors de *deepfakes*. La capacité à produire des *deepfakes* est malheureusement utilisée aujourd'hui à grande échelle à des fins malveillantes, que ce soit par des particuliers ou de grandes organisations. Certains états les utilisent pour des campagnes de désinformation massive. Par exemple, en 2022, pendant l'invasion de l'Ukraine par Russie, une vidéo truquée de Volodymyr Zelensky, président de l'Ukraine, le montrait en train de demander à la population ukrainienne de se rendre. Ce type d'usage est devenu courant pendant les élections, y compris dans les États « démocratiques » occidentaux, et est amplifié par l'usage de *bots*, des logiciels qui simulent des personnes réelles sur les réseaux sociaux afin de propager ces fausses informations. Cela illustre les enjeux citoyens majeurs qui y sont associés. D'autres usages malveillants incluent les escroqueries commerciales ou les fausses publicités. Enfin, les *deepfakes* sont aussi utilisés de manière grandissante par les individus, y compris les adolescents, pour mettre en scène, parfois de manière ridicule, des personnes publiques ou bien certaines de leurs connaissances. Un autre usage consiste à se mettre en scène ou à utiliser ces logiciels pour modifier son apparence et montrer une image de soi qui ne correspond pas à la réalité. Ces usages peuvent ainsi avoir de graves conséquences sur la vie des personnes concernées.

Les capacités nouvelles et très larges des technologies d'IAGen, mais aussi leur nature qui en fait des outils puissants de transmission culturelle, ce qui inclut tout un ensemble de mécanismes de désinformation, posent ainsi aujourd'hui à la société en général, et pour les jeunes et futurs citoyens en particulier, de grands défis et de grandes opportunités que nous discuterons ci-dessous.

II. Les enjeux sociétaux et éducatifs globaux de l'IA générative

A. L'usage massif et grandissant de l'IA générative.

Quelques mois après la sortie du modèle GPT-3 de l'entreprise OpenAI, qui a marqué un tournant technologique et scientifique, le modèle ChatGPT 3.5 a été lancé en novembre 2022 : il a atteint 1 million d'utilisateurs en 5 jours, 100 millions en deux mois, et il en compte aujourd'hui 180 millions. De nombreux autres logiciels sont apparus, permettant aussi de produire des images (par exemple MidJourney avec 20 millions d'utilisateurs, DALL-E ou Stable Diffusion) ou des vidéos (par exemple Sora ou Runway). Alors que les premiers logiciels d'IAGens étaient « privés », c'est à dire que leurs paramètres internes étaient secrets, un nombre grandissant

d'acteurs institutionnels et individuels ont commencé à développer et partager des modèles « *open-weights* », c'est-à-dire qu'ils partageaient le code de ces modèles sur internet pour un usage libre. Par exemple, la plateforme Hugging Face héberge aujourd'hui plus de 400 000 modèles d'IAGen *open-weights*, c'est-à-dire des logiciels dont tout le code et les paramètres sont accessibles et librement téléchargeables.

En France, selon une enquête Ifop/Talan (2024), près de 70 % des 18-24 ans utilisent personnellement les logiciels d'IAGen, contre seulement 47 % des 25-34 ans et 22 % des 35 ans et plus. Les utilisateurs estiment gagner 38 % de productivité et d'efficacité grâce aux IA génératives. En particulier, 46 % des 18-24 ans estiment cette hausse de la productivité à plus de 40 %. Enfin, 44 % des utilisateurs (et 61 % des 25-34 ans) reprennent les résultats des IA génératives tels quels sans les modifier et 35 % déclarent qu'ils auraient du mal à se passer des IA génératives. Un sondage pilote en Nouvelle Aquitaine auprès de lycéens, réalisée par l'équipe Flowers à Inria en juin 2024, indique que plus de 90 % des élèves de seconde ont déjà utilisé un logiciel d'IA générative pour les aider à faire leurs devoirs. Par ailleurs, une grande proportion d'entre eux, après avoir expérimenté ChatGPT, l'utilisent directement ensuite pour faire des recherches d'information, sans passer par des moteurs de recherche classique. Globalement, en France, l'augmentation du nombre d'utilisateurs a été de 60 % en un an, et c'est surtout un usage massif des jeunes générations.

Dans le monde professionnel, les usages de l'IA générative par les salariés sont très divers, allant de l'aide à la rédaction d'email ou de rapports, la génération de réponses aux demandes de clients, le résumé automatique de rapports ou de réunions, la traduction, l'aide au brainstorming ou à la résolution de problème, l'analyse sémantique de productions écrites, l'aide à la classification de CVs ou de dossier de financements, l'aide à la production d'illustration graphiques, ou encore l'aide à la programmation pour les développeurs informatiques.

Chez les particuliers en général, une grande proportion des utilisateurs utilise l'IA générative principalement pour des usages récréatifs. Cependant, chez les plus jeunes, qui sont aussi les plus grands utilisateurs, l'utilisation de ces logiciels pour l'aide aux devoirs est grandissante.

Globalement, les contenus produits avec des IA génératives sont aussi très largement partagés en ligne, à la fois parce que les individus partagent les contenus qu'ils ont produits avec des IAGens, mais aussi parce que l'on observe un usage massif et grandissant d'organisations qui déploient à grande échelle des logiciels d'IAGen sur les réseaux sociaux, simulant des utilisateurs réels, et produisant ainsi du contenu lu à la fois par des humains et par d'autres logiciels d'IAGen. Selon Europol, 90 % du contenu en ligne pourrait avoir été généré par des IA générative à la fin 2026 (Europol, 2022).

B. L'IA générative à l'école : les défis aujourd'hui

Plusieurs études convergentes indiquent qu'en Europe et en Amérique du Nord, plus de 80 % des 14-18 ans ont déjà utilisé ChatGPT pour faire leurs devoirs. Parmi eux, une étude américaine indique que 38 % d'entre eux l'ont fait sans le dire à leur enseignant (*Common Sense Media and Impact Research*, 2023). Ces usages peuvent consister à demander à ChatGPT de résoudre leurs

exercices de maths, proposer des rédactions en français ou en histoire, traduire en texte en LV1 ou préparer un exposé en sciences physiques. À l'université, les usages sont aussi devenus massifs : par exemple, l'étude 2024 De Vinci *Higher Education*, *RM conseil and Talan*, réalisée auprès de 1 600 étudiants de 4e année universitaire en management, en ingénierie et en informatique, montre que 92 % d'entre eux utilisent l'IAGen régulièrement, et 30 % d'entre eux paient un abonnement de 20 euros par mois pour avoir accès aux meilleures versions. Par ailleurs, 65 % d'entre eux disent que la présence d'IA générative sera un critère majeur dans le choix des entreprises dans lesquelles ils souhaiteront travailler.

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 incapable d'évaluer leurs propres incertitudes), peut amener de nombreux élèves à surestimer à la fois les compétences des IAGen 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 al.*, 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 et 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[1] (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

[1] Les biais cognitifs sont des raccourcis de pensée qui peuvent nous induire en erreur sans que l'on s'en rende compte. Par exemple, l'attribution d'agentivité est notre tendance à croire que des objets ou des événements ont une volonté, comme lorsqu'on pense qu'une voiture « refuse » de démarrer. L'attribution d'intention, c'est quand on imagine qu'un objet agit avec un but précis, comme si une machine « voulait » quelque chose.

particulièrement problématique étant donné que les IAGen encodent de nombreux stéréotypes (raciaux, de genre, religieux, etc) (Bender *et al.*, 2021).

Le rôle de ces biais est aussi illustré par une expérimentation récente réalisée auprès d'un groupe d'étudiants d'école de commerce en France pendant un cours d'économie comportementale (Hill, 2023). Chaque étudiant devait résoudre deux cas d'étude (choisis au hasard parmi 14) : l'un pour lequel il devait trouver la réponse par lui-même, et l'autre pour lequel une réponse était déjà fournie, et il devait la corriger ou l'améliorer. Dans le second cas, on donnait aux élèves des réponses à corriger qui étaient produites soit avec ChatGPT, soit par un élève de l'année précédente. Les élèves étaient informés de la procédure, mais ne savaient pas si une réponse donnée venait d'un autre élève ou de ChatGPT. Il leur était aussi dit que les réponses de ChatGPT étaient souvent de piètre qualité. Les résultats étaient très nets : les élèves ont eu en moyenne une note de 28 % supérieure dans le premier cas (sans proposition de réponse) que dans le second (réponse à corriger/améliorer). L'analyse qualitative de leurs productions a montré une grande difficulté à se départir de la proposition initiale, et en particulier de celles fournies par ChatGPT, qui néanmoins étaient de moins bonne qualité que s'ils avaient répondue sans cette proposition de réponse.

Les défis éducatifs liés aux limites dans la compréhension des logiciels d'IAGen sont aussi illustrées par une étude récente réalisée dans plusieurs collèges en Nouvelle Aquitaine (Abdelghani, 2024). Dans cette étude, des exercices en sciences ont été donnés à 72 élèves de 4e et 3e dans quatre collèges : chaque exercice comportait une illustration et un court texte présentant une observation d'un phénomène naturel, et ils devaient chercher et rédiger une courte explication. ChatGPT était l'outil de recherche qui était mis à leur disposition, et l'objectif de l'étude était de comprendre comment ils formulent leurs questions et dans quelle mesure ils arrivent à formuler un contexte permettant au logiciel de donner une réponse pertinente. En effet, la capacité à formuler des questions d'investigations précises et informatives est essentielle dans les processus d'apprentissage, et est associée aux capacités métacognitives des apprenants (Abdelghani, 2024). D'abord, les données récoltées montrent que 73 % de ces élèves ont déjà utilisé ChatGPT. Par ailleurs, ils font une grande confiance aux réponses de ChatGPT : 82 % d'entre eux pensent qu'elles sont fiables. En même temps, seulement 33 % d'entre eux disent ne pas connaître les limites de ChatGPT. Ensuite, l'expérimentation montre que leur capacité à formuler des questions pertinentes et bien contextualisées (ou choisir les bonnes questions parmi plusieurs proposées), est faible (dans 49 % des cas ils choisissent une question qui n'est pas adaptée, ce qui est l'équivalent d'un choix au hasard). Enfin, 79 % des participants ne posent qu'une seule question à ChatGPT et ne remettent pas en question l'exactitude de sa réponse, les menant à un taux de réussite aux exercices bas (43 %).

Les enjeux sont aussi majeurs pour les enseignants. D'abord, l'usage grandissant et massif des IAgen pour l'aide à la réalisation des devoirs à la maison rend très difficile leur évaluation par les enseignants. En particulier, si des outils logiciels d'IA sont apparus avec l'objectif de détecter automatiquement les textes ou images générés par IA, beaucoup de scientifiques s'accordent sur le fait que cet objectif est quasi impossible à atteindre (Oravec, 2023) : de nouveaux logiciels d'IAGen apparaissent en permanence, en particulier avec des fonctionnalités pour contourner la détection automatique de leur utilisation, et avec un risque élevé de détecter de faux positifs, c'est-à-dire d'attribuer à un logiciel d'IAGen des textes véritablement écrits par des humains.

Globalement, l'arrivée de l'IAGen a amené de nombreux enseignants à modifier leurs pratiques d'enseignement. Dans une étude de grande ampleur réalisée auprès de 908 enseignants du primaire et secondaire en Estonie (Laak & Aru, 2024), on observe que l'arrivée de l'IAGen a amené 49 % des enseignants à modifier leurs pratiques, en éliminant une grande partie des devoirs à la maison, et en incluant des activités favorisant la pensée critique. Cependant, cette étude, ainsi que d'autres études convergentes (par exemple l'étude *Impact Research* 2024 réalisée auprès de 1 003 enseignants aux États-Unis, et Klopfer *et al.*, 2023), montrent aussi des usages évalués comme positifs par les enseignants (74 % d'entre eux dans l'étude de Laak et Aru, 2024), en particulier pour les aider dans la préparation d'activités créatives et motivantes sur un sujet, ou la mise au point de plans de cours et d'exercices/quiz associés, ou enfin pour répondre aux emails envoyés par les parents. En revanche, l'utilisation de ces logiciels pour les aider à évaluer les travaux des élèves est jugée peu pertinente et peu efficace par les enseignants (*Impact Research*, 2024). Enfin, beaucoup d'entre eux souhaiteraient recevoir plus de formations sur l'IAGen et la manière de la prendre en compte et l'inclure dans les enseignements.

C. IA générative et éducation : des opportunités demain ?

Au-delà de ces grands défis associés au rôle de l'IA générative dans les apprentissages scolaires, ces évolutions technologiques pourraient aussi amener à des opportunités éducatives diverses.

Comme expliqué ci-dessus, l'IA générative n'est qu'une approche de l'IA parmi beaucoup d'autres, et il y a une longue tradition de travaux alliant IA, sciences cognitives et sciences de l'éducation, ayant mené à la mise au point de logiciels éducatifs proposant des activités d'apprentissage basés sur des principes cognitif et/ou pédagogiques forts et utilisant l'IA pour plus d'accessibilité et de personnalisation (Nkambou *et al.*, 2010). Par exemple, des travaux récents dans l'équipe Flowers à Inria, ont permis de développer un algorithme de personnalisation de séquences d'exercices basé sur les principes de notre compréhension des mécanismes de l'apprentissage dirigé par la curiosité chez l'enfant (Oudeyer *et al.*, 2016). Des expérimentations à l'école primaire auprès d'environ 1 000 enfants de 7/8 ans ont montré que cette approche permettait, en comparaison avec des séquences d'exercices faites à la main par un expert en didactique des maths, d'améliorer significativement l'efficacité d'apprentissage et la motivation, en particulier pour les élèves différant de l'élève « moyen » (Clement *et al.*, 2015, Clément *et al.*, 2024). Cette approche a ensuite été transposée dans le logiciel AdaptivMaths, développé dans le cadre du programme français P2IA, et supporté par le ministère de l'Éducation nationale[2].

De la même manière, une part grandissante de la communauté scientifique qui travaille sur la mise au point d'usages pédagogiquement pertinents de l'IA générative, a commencé à étudier les manières d'utiliser l'IA générative au bénéfice des élèves, des enseignants, et plus largement de l'écosystème éducatif. Prenons quelques exemples qui illustrent cette diversité et ces perspectives.

Dans une étude réalisée auprès de 8 762 étudiants de 146 pays, inscrits pour une formation en ligne d'introduction à la programmation informatique, les auteurs ont étudié l'effet de la mise à

[2] Logiciel accessible au lien suivant : https://www.adaptivmath.fr/

disposition gratuite d'une version de GPT-4 prompté de manière à fournir une aide pédagogique (ne pas donner la réponse directement) aux élèves quand ils leur posaient des questions sur le cours (Nie *et al.*, 2023). L'étude a d'abord montré que l'impact de la disponibilité de GPT-4 était très différent entre les pays à haut indice de développement humain (IDH), par exemple les États-Unis, la Norvège ou la Suisse, et les pays à bas développement humain (par exemple Sénégal, Zimbabwe, Pakistan) : dans les pays à haut IDH, l'engagement et la participation (optionnelle) des élèves à l'examen a diminué significativement par rapport à un groupe contrôle qui n'avait pas accès à GPT-4, tandis qu'il a augmenté significativement dans les pays à bas IDH (l'étude ne permet pas de comprendre pourquoi on observe cette différence, mais deux hypothèses sont avancées : soit les élèves des pays à haut IDH ont une vision plus négative et plus méfiante de l'IA générative, soit ils arrêtent plus facilement l'usage de cette version de ChatGPT pour aller interagir avec la version grand public qui peut leur donner des réponses directement). Par ailleurs, l'étude montre que le score à l'examen des participants qui ont adopté l'usage de GPT-4 était significativement supérieur par rapport à ceux qui ne l'avaient pas utilisé. Ces résultats, dont certains aspects restent encore difficiles à interpréter, montrent néanmoins le rôle positif que pourrait avoir l'IAGen dans les populations venant de pays dont le système éducatif est peu développé.

Les systèmes d'IA générative sont aussi étudiés pour aider à générer des contenus personnalisés dans les logiciels éducatifs : exercices, indices et feedback personnalisés, et même explications. Dans une extension de l'étude ci-dessus, réalisée auprès des 8 762 élèves de 146 pays, les auteurs ont comparé, pour des exercices de programmation, l'efficacité de messages d'erreurs générés par des méthodes classiques avec des messages d'erreurs générés avec GPT (Wang *et al.*, 2024). Ils ont montré que les messages d'erreurs générés avec GPT amenaient les élèves à apprendre à résoudre les exercices significativement plus vite.

Un autre exemple est l'étude présentée dans un article d'Abdelghani *et al.* (2023), qui étudie l'utilisation de GPT-3 pour entrainer les enfants à poser des questions curieuses (c'est-à-dire des questions en rapport avec un sujet qu'on leur a introduit et dont ils ne connaissent pas déjà la réponse), ainsi que pour entrainer leurs compétences métacognitives (compétences permettant de réfléchir à ses propres processus d'apprentissage). Pour réaliser cela, GPT-3 est utilisé pour pré-générer des exercices et des indices permettant de réaliser cet entraînement. On sait en effet que ces compétences jouent un rôle clé dans les apprentissages en général. Ici, le logiciel a été testé auprès d'un groupe de 75 enfants de 9-10 ans dans des écoles primaires en Nouvelle-Aquitaine. À partir de courts textes sur des sujets de sciences, les élèves avaient pour objectif de formuler des questions complexes en lien avec le texte mais dont la réponse n'était pas dans le texte. L'objectif de l'étude était de comparer deux types d'indices (des aides ou des repères pour aider les élèves à formuler leurs questions) : d'une part, des indices créés manuellement par des experts humains, et d'autre part, des indices générés automatiquement par GPT-3. L'étude a montré que les indices créés par GPT-3 étaient au moins aussi bons, voire meilleurs dans certaines catégories, que ceux créés par des experts humains. GPT-3 a donc eu un impact aussi positif, voire plus, sur la qualité des questions formulées par les élèves.

Un autre usage de l'IA générative est expérimenté dans le projet GPTeach (Markel *et al.*, 2023) : simuler le comportement d'élèves divers afin d'entraîner des apprentis enseignants à utiliser des stratégies pédagogiques. Plus précisément, le système étudié simule des étudiants avec des

personnalités, des besoins et des objectifs d'apprentissage variés, et scénarise des situations d'enseignement qui peuvent inclure un groupe d'élèves simulés. Bien que cette étude comporte un petit nombre de participants (24), les résultats montrent une appréciation très positive de ces apprentis enseignants : ces situations d'entraînement leur permettent d'essayer et de répéter des interventions pédagogiques sans la pression qu'ils pourraient avoir de se retrouver tout de suite devant des élèves dans une classe réelle.

D. L'IA générative : une grande variété d'usages, des sciences aux arts.

Comme mentionné ci-dessus, l'IA générative est aujourd'hui utilisée de manière grandissante dans le monde professionnel. Dans les entreprises, un usage majeur est celui de l'aide à la rédaction. Une étude randomisée, réalisée auprès de 453 professionnels occupants des fonctions variées (consultants, ressources humaines, analystes de données, managers), a par exemple montré que l'usage de ChatGPT pour des tâches courantes (emails, rapports de synthèse, communiqués de presse) permettait de réaliser un gain de temps de 40 %, tout en augmentant la qualité de 18 % en moyenne (Noy and Zhang, 2023). Avec d'autres usages comme l'aide à la décision, ou l'accès aux bases de connaissances d'une entreprise, l'IA générative est ainsi en train de provoquer des transformations majeures du monde du travail, impliquant de nombreux défis pour les employés et les organisations elles-mêmes (GPAI, 2023).

Au-delà de l'entreprise, les usages sont aussi nombreux dans divers secteurs. Dans le domaine des sciences, l'IA générative commence à ouvrir des perspectives extraordinaires pour aider physiciens, chimistes, biologistes, ou mathématiciens à faire de nouvelles découvertes. Par exemple, de nombreux laboratoires étudient aujourd'hui l'usage de systèmes d'IA générative pour générer efficacement de nouvelles structures chimiques ou physiques (Park *et al.*, 2024), ouvrant la possibilité de découvrir par exemple des protéines nouvelles et pertinentes pour des applications en santé ou en agriculture (Zambaldi *et al.*, 2024), ou de nouveaux matériaux (Merchant *et al.*, 2023). En mathématiques, des projets récents ont montré comment des modèles de langages pouvaient permettre d'explorer et de trouver de nouvelles solutions à des problèmes ouverts (Romera-Paredes *et al.*, 2024), comment l'IA générative pouvait aussi être utilisée pour aider les mathématiciens à formaliser et développer des preuves de théorèmes (Wu *et al.*, 2022), et des recherches actuelles s'intéressent à la manière dont ces systèmes pourraient générer de nouvelles conjectures intéressantes (Bengio et Malkin, 2023).

L'IA générative peut aussi servir d'autres sciences comme l'archéologie. Dans le cadre du « Vesuvius Challenge »[3], une équipe a en effet réussi à déchiffrer les restes de textes sur des rouleaux de papyrus très endommagés lors de l'éruption du Vésuve. En éthologie, le projet *Earth Species*[4] explore l'utilisation de l'IA générative pour aider à décoder les signaux de communication animale, ce qui n'est pas sans poser des difficultés qui restent encore largement à résoudre (Yovel & Rechavi, 2023).

[3] Accessible au lien suivant : https://scrollprize.org/

[4] Accessible au lien suivant : https://www.earthspecies.org/

D'autres domaines dans lesquels les usages de l'IA générative se développent incluent l'aide à la création artistique, en particulier la création d'images, de musiques, de voix et de films avec des systèmes comme MidJourney, DALL-E, HeyGen Suno AI ou Sora - ce qui pose de nombreuses questions éthiques et juridiques – ou l'aide à la mise au point de jeux vidéo (Bruce *et al.*, 2024).

E. Les enjeux culturels et démocratiques.

Ainsi, les opportunités et les défis sociétaux de l'IA générative, en particulier vue comme un ensemble d'outils de transmission et d'amplification culturelle, sont aujourd'hui très grands, et commencent à être bien décrits et caractérisés dans la littérature. Dans cette perspective, la mise au point de corpus d'entraînement (et de guidage par le *scoring* humain) peut être vue comme une forme d'« éducation » des modèles d'IAGen, dans le sens où ces corpus vont définir les orientations culturelles qui vont y être encodées. Il y a donc un enjeu émergent qui est très similaire à celui de l'éducation des humains pour les organisations, notamment les États : quelles sont les connaissances et les valeurs, et plus généralement la culture, que nous souhaitons voir encoder/apprise par les modèles d'IAGen ? Dans quels buts et pour quels usages et usagers ? C'est un enjeu principalement politique et culturel, plus que technologique. À court terme, une première étape consiste à comprendre comment les données utilisées pour entraîner les grands logiciels d'IA générative pourraient être rendues accessibles, au moins à des institutions tierces de confiance, afin de pouvoir vérifier un certain nombre de dimensions culturelles et juridiques associées à ces données : c'est l'un des enjeux de l'application du récent *European Digital AI Act*[5].

Alors qu'il y a aujourd'hui, dans la plupart des États, une longue tradition (théorique et pratique) de méthodes de mise au point de programmes scolaires (utilisées y compris par des États totalitaires qui souhaitent contrôler la pensée de leur population), il n'y a aujourd'hui dans l'écosystème de l'IA aucune approche globale et organisée de cette problématique. Elle est même quasiment absente de la plupart des débats actuels, qui se focalisent sur des enjeux technologiques, alors que les plus grandes questions au sujet des grands modèles d'IA générative sont des questions politiques et culturelles.

En pratique, aujourd'hui les grandes organisations privées comme *OpenAI* constituent des corpus très grands, très hétérogènes, à partir de sources très variées qu'ils captent sur internet, et filtrés avec un mélange d'automatisation massive et de travail humain (*micro-workers*). Pour le *scoring* humain, il est fait à très grande échelle en faisant appel à du micro-travail : par exemple OpenAI fait appel à des travailleurs kenyans, dans des conditions humaines qui sont fort critiquables (Gray *et al.*, 2019). La constitution de ces corpus, de ces modes de *feedbacks*, et des filtres pour « corriger »/« éliminer » les textes ou images générés qui sont évaluées comme « inacceptables », le sont à partir de valeurs et modèles de cultures qui ne sont pas explicités en détails par ces organisations. En pratique, elles reflètent un mélange entre une vision des valeurs et de la culture portée par l'industrie « tech » californienne (Soleiman & Dennison, 2021 ; Rozado, 2022), et une superposition de valeurs et de cultures présentes sur internet (Kovac *et al.*,

[5] Accessible au lien suivant : https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

2023), sans qu'il n'y ait de bonne compréhension de ces valeurs, de ces cultures et de la manière dont elles sont encodées.

Plusieurs actions et projets ambitieux ont été développés récemment pour mettre au point des méthodologies plus solides, et surtout sociétalement plus positives, inclusives et acceptables, de la constitution de corpus et de modes de *feedbacks*. En particulier, on peut noter les travaux du consortium BigScience qui a mis au point une approche pour penser la gouvernance de la constitution de ces corpus (Jernite *et al.*, 2022), a conçu un corpus (ROOTS, Laurençon *et al.,* 2022) avec cette méthode et l'a partagé avec la communauté scientifique. Cependant, ce corpus n'a été conçu et analysé que superficiellement d'un point de vue culturel : seule la langue et quelques biais relativement simples ont été pris en compte. Une exception notable est le travail présenté par Johnson *et al.* (2022), qui présente une étude avancée des dimensions culturelles des corpus utilisés pour entraîner GPT-3 (et les valeurs encodées par le modèle).

Globalement, à peu près tout reste à faire, tant du point de vue scientifique que du point de vue politique et culturel, pour comprendre comment constituer des corpus et des approches de *feedback* humains qui visent à aligner des modèles sur des ensembles de régularités culturelles bien comprises. Cela pose même des questions fondamentales en sciences humaines pour caractériser ces régularités culturelles, qui sont posées de manière nouvelle avec ces technologies de transmission et d'amplification culturelle.

Il est aussi essentiel de garder à l'esprit que tous les concepts et les techniques qui visent à aligner les grands modèles de langage sur des valeurs humaines particulières sont des approches et techniques qui permettent de contrôler les valeurs culturelles qui seront transmises. En ce sens, ces approches peuvent être autant utiles pour transmettre des valeurs et des connaissances alignées par exemple avec celles de démocraties occidentales, mais aussi par les organisations, petites ou grandes, étatiques ou privées, qui souhaitent influencer, voire contrôler, ce que pensent des populations ciblées.

F. IA et impact environnemental

L'usage massif des systèmes d'IA générative pose aussi des enjeux environnementaux majeurs (Trystram *et al.*, 2021). En effet, chaque logiciel d'IA générative requiert des ressources de calculs très grandes à la fois au moment où il est entraîné, mais aussi chaque fois qu'il est utilisé pour donner une réponse à un « prompt ». Les centres de calculs qui permettent cela et hébergent des super-ordinateurs, peuvent avoir une taille gigantesque (plusieurs hectares) : ils occupent des sols à la place d'espaces verts ou de logements, et consomment des quantités considérables d'électricité. Dans de nombreux pays où sont hébergés ces centres de calculs, l'électricité est produite au moyen d'énergie fossile, ce qui provoque ainsi le rejet de grandes quantités de CO2 et de nombreux polluants. On estime que ces six dernières années, la consommation électrique mondiale liée à l'IA a été multipliée par un million, et certaines études estiment qu'elle pourrait atteindre 10 % de la consommation électrique mondiale en 2030. Pour refroidir les processeurs, ces centres de calculs consomment aussi de grandes quantités d'eau (Li *et al.*, 2023). Certaines estimations évaluent qu'une trentaine de requêtes à ChatGPT consomment environ un demi-litre d'eau, ce qui permet d'avoir un ordre de grandeur de la consommation associée aux requêtes de millions d'utilisateurs chaque jour.

Aux côtés de ces impacts négatifs, des projets étudient aussi différentes manières dont l'IA générative pourrait contribuer à nous aider à gérer le changement climatique (Cowles *et al.*, 2023). Certains projets visent à simuler et mieux prédire l'évolution du climat et les évènements extrêmes, quand d'autres visent à utiliser l'IAGen pour optimiser la gestion de l'énergie dans les bâtiments ou les transports, ou enfin analyser des données satellites pour aider les agriculteurs à mettre au point des stratégies de culture plus respectueuses de l'environnement.

Par ailleurs, de nombreux projets scientifiques travaillent aujourd'hui à comprendre comment il serait possible de mettre au point des logiciels d'IA générative beaucoup moins coûteux en énergie (on parle alors d'IA frugale). Il s'agit par exemple de développer, au lieu d'un grand modèle généraliste, un ensemble de petits modèles d'IA (Touvron *et al.*, 2023), qui peuvent être spécialisés pour les besoins particuliers d'un groupe d'utilisateurs (Fu *et al.*, 2023). De manière générale, un certain nombre de chercheurs et d'organisations travaillent à augmenter la transparence des travaux dans ce domaine, en amenant les concepteurs à mesurer et partager le coût énergétique de la production et de l'utilisation des modèles d'IA générative.

III. L'éducation à la littératie en IA générative : pistes et outils

Étant donné ces enjeux sociétaux, il apparaît ainsi aujourd'hui fondamental de développer l'éducation à l'IA générative, c'est-à-dire de permettre aux enfants et adolescents (et au-delà) d'acquérir progressivement une littératie en intelligence artificielle générative, c'est-à-dire une compréhension générale qui permet d'interpréter les informations sur ce sujet avec un esprit critique et éclairé. Il ne s'agit pas ici d'une éducation technique qui vise à les préparer à un métier dans ce domaine : il s'agit au contraire de leur permettre d'avoir une culture suffisamment développée sur les mécanismes, les usages et les enjeux sociétaux, leur permettant d'une part d'utiliser ces outils à bon escient, et d'autre part d'exprimer de manière éclairée leur opinion de citoyen pour contribuer aux grandes orientations collectives en rapport avec l'IA générative (et donc en rapport avec des grands enjeux culturels et démocratiques). Dans cette perspective, il semble aussi fondamental de développer l'acculturation et la littératie en IA générative chez les enseignants, et chez les parents en général, à la fois pour équiper conceptuellement les adultes dans leur rôle de citoyen, mais aussi afin qu'ils puissent contribuer à construire l'écosystème dans lequel les plus jeunes pourront développer cette littératie. En bref, le développement de la littératie en IA générative est un enjeu pour tout le monde.

Plusieurs approches et outils pédagogiques commencent à être développés, mais les efforts en sont encore au commencement : un travail d'évaluation et d'adaptation de ces outils pédagogiques sera nécessaire au cours du temps. L'équipe Flowers, au Centre Inria de l'université de Bordeaux, a récemment développé plusieurs ressources visant à développer la littératie en IA générative pour le plus grand nombre, et en particulier pour les collégiens, lycéens et leurs enseignants. Tout d'abord, la série pédagogique « ChatGPT expliqué en 5 min[6] » est constituée d'un ensemble de vidéos courtes (entre 5 et 10 min) visant à introduire les

[6] https://developmentalsystems.org/chatgpt_5_minutes/fr/

mécanismes, les usages et les enjeux sociétaux de l'IA générative (Torres-Leguet *et al.*, 2024). Par exemple, la première vidéo introduit la notion de modèle de langage (ce que ça fait, comment on les fabrique), et introduit un résumé de l'ensemble des limites de ces systèmes (connaissances figées, hallucinations, acteurs économiques privés). La seconde vidéo se focalise sur les limites, et explique plus en détails les notions de biais et d'alignement, les connaissances statiques et non sourçables, la quantité de données nécessaire à l'entraînement des modèles, les hallucinations, ou le problème du manque d'ancrage des IA génératives dans le monde physique. Une autre vidéo présente les forces des modèles de langage, par exemple la grande maîtrise des langues, l'utilisation d'outils externes pour faire des calculs, ou la manière dont ils peuvent aider à brainstormer pour des projets créatifs. Enfin, plusieurs vidéos expliquent les différentes manières de construire des prompts et d'amener les IA génératives à simuler et expliquer des raisonnements, permettant aux apprenants de poser des questions en apportant un contexte qui leur permettra d'obtenir des réponses pertinentes et de mieux interpréter ces réponses. Enfin, une vidéo explique les applications dans différents domaines (quotidien, travail, santé, éducation, sciences, diversité culturelle, etc.). Ces vidéos sont toutes en licence Creative Commons CC-BY, ce qui autorise de manière gratuite tous les usages de ces vidéos, y compris les usages commerciaux ou leur intégration dans des outils pédagogiques plus large, et sont librement téléchargeables sur le site web[7].

Au-delà de la découverte des mécanismes et des enjeux sociétaux de l'IA générative, un enjeu essentiel est de développer des outils et approches permettant aux adolescents d'entraîner leur pensée critique, leur capacité à poser des questions curieuses et à remettre en question des informations qu'ils reçoivent, à apprendre de manière active et par eux-mêmes. Ces capacités sont très liées aux mécanismes de la curiosité (Oudeyer *et al.*, 2016) ainsi qu'à ceux de la métacognition (Proust, 2019), qui sont essentiels en général pour les apprentissages, et particulièrement importants quand ceux-ci sont réalisés en interaction avec des systèmes d'IA générative (Abdelghani *et al.*, 2023). Bien que ces compétences transverses soient encore très peu prises en compte dans les enseignements scolaires, plusieurs travaux ont commencé à montrer comment certains exercices et gestes pédagogiques peuvent être mis à profit afin d'entraîner la curiosité et la métacognition des enfants dans le contexte scolaire (Abdelghani *et al.*, 2023b ; Guilleray *et al.*, 2024). Cette voie ouvre ainsi non seulement des perspectives complémentaires et prometteuses pour le développement de la littératie en IA générative, mais pour l'éducation des futurs citoyens en général.

[7] https://developmentalsystems.org/chatgpt_5_minutes/fr/

Références

Abdelghani, R. (2024). *Guiding the minds of tomorrow: Conversational Agents to Train Curiosity and Metacognition in Young Learners*. Thèse de l'université de Bordeaux.

Abdelghani, R., Sauzéon, H., & Oudeyer, P. Y. (2023). Generative AI in the Classroom: Can Students Remain Active Learners? In *Workshop on Generative AI in Education at the Conference on Neural Information Processing Systems*.

Abdelghani, R., Law, E., Desvaux, C., Oudeyer, P. Y. & Sauzéon, H. (2023b). Interactive environments for training children's curiosity through the practice of metacognitive skills: a pilot study. In *Proceedings of the 22nd Annual ACM Interaction Design and Children Conference* (pp. 495-501).

Abdelghani, R., Wang, Y. H., Yuan, X., Wang, T., Lucas, P., Sauzéon, H. & Oudeyer, P. Y. (2024). GPT-3-driven pedagogical agents to train children's curious question-asking skills. *International Journal of Artificial Intelligence in Education*, 34(2), 483-518.

Bender, E. M., Gebru, T., McMillan-Major, A. & Shmitchell, S. (2021, March). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency* (pp. 610-623).

Bengio, Y. & Malkin, N. (2024). Machine learning and information theory concepts towards an AI Mathematician. *Bulletin of the American Mathematical Society*, 61(3), 457-469.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems* (pp. 1877-1901).

Bruce, J., Dennis, M. D., Edwards, A., Parker-Holder, J., Shi, Y., Hughes, E., ... & Rocktäschel, T. (2024, February). Genie: Generative interactive environments. In *Forty-first International Conference on Machine Learning*.

Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., ... & Xie, X. (2024). A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3), 1-45

Cordeschi, R. (2002). *The discovery of the artificial: Behavior, mind and machines before and beyond cybernetics* (Vol. 28). Springer Science & Business Media

Common Sense Media and Impact Research (2023) https://www.commonsensemedia.org/sites/default/files/featured-content/files/common-sense-ai-polling-memo-may-10-2023-final.pdf

Clement, B., Roy, D., Oudeyer, P. Y. & Lopes, M. (2015). Multi-Armed Bandits for Intelligent Tutoring Systems. *Journal of Educational Data Mining*, 7(2).

Clément, B., Sauzéon, H., Roy, D. & Oudeyer, P. Y. (2024). I*mproved Performances and Motivation in Intelligent Tutoring Systems: Combining Machine Learning and Learner Choice*. arXiv preprint arXiv:2402.01669.

Cowls, J., Tsamados, A., Taddeo, M. & Floridi, L. (2023). The AI gambit: leveraging artificial intelligence to combat climate change—opportunities, challenges, and recommendations. *AI & Society,* 1-25.

De Vinci Higher Education, RM conseil and Talan (2024) Study 2024 “The impact of generative AI on students”

Europol (2022). *Facing reality? Law enforcement and the challenge of deepfakes, an observatory report from the Europol Innovation La*b. Publications Office of the European Union, Luxembourg. https://www.europol.europa.eu/cms/sites/default/files/documents/Europol_Innovation_Lab_Facing_Reality_Law_Enforcement_And_The_Challenge_Of_Deepfakes.pdf

Fu, Y., Peng, H., Ou, L., Sabharwal, A. & Khot, T. (2023, July). Specializing smaller language models towards multi-step reasoning. In *International Conference on Machine Learning* (pp. 10421-10430). PMLR.

GPAI (2023) *Future of Work Working Group Repor*t. https://gpai.ai/projects/future-of-work/Future%20of%20Work%20Working%20Group%20Report%20v2%20(November%202023).pdf

Gray, M. L. and Suri, S. (2019) *Ghost work: how to stop Silicon Valley from building a new global underclass*. Eamon Dolan Books.

Guilleray, F., Proust, J. & Fernandez, J. (2024) *Glossaire de la métacognition*. Conseil scientifique de l’Éducation nationale/Réseau Canopé.

Heider, F. & Simmel, M. (1944). An experimental study of apparent behavior. T*he American journal of psychology*, 57(2), 243-259.

Hershcovich, D., Frank, S., Lent, H., de Lhoneux, M., Abdou, M., Brandl, S., ... & Søgaard, A. (2022). *Challenges and strategies in cross-cultural NLP*. arXiv preprint arXiv:2203.10020.

Hill, B. (2023). Taking the help or going alone: ChatGPT and class assignments. *HEC Paris Research Paper Forthcoming*.

Ifop/Talan (2024) Baromètre 2024 Ifop pour Talan. Les Français et les IA génératives. https://www.talan.com/actualites/detail-actualites/news/barometre-2024-ifop-pour-talan-les-francais-et-les-ia-generatives/

Impact Research (2024). *AI Chatbots in Schools.* https://www.waltonfamilyfoundation.org/ai-in-the-classroom

Jernite, Y., Nguyen, H., Biderman, S., Rogers, A., Masoud, M., Danchev, V., ... & Mitchell, M. (2022). Data governance in the age of large-scale data-driven language technology. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (pp. 2206-2222)

Ji, J., Qiu, T., Chen, B., Zhang, B., Lou, H., Wang, K., ... & Gao, W. (2023). A*I alignment: A comprehensive survey*. arXiv preprint arXiv:2310.19852.

Jurenka, I., Kunesch, M., McKee, K. R., Gillick, D., Zhu, S., Wiltberger, S., ... & Ibrahim, L. (2024). *Towards responsible development of generative AI for education: An evaluation-driven approach*. arXiv preprint arXiv:2407.12687.

Johnson, R. L., Pistilli, G., Menédez-González, N., Duran, L. D. D., Panai, E., Kalpokiene, J. & Bertulfo, D. J. (2022). *The Ghost in the Machine has an American accent: value conflict in GPT-3*. arXiv preprint arXiv:2203.07785

Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., ... & Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and individual differences,* 103, 102274.

Kidd, C. & Birhane, A. (2023). How AI can distort human beliefs. *Science*, 380(6651), 1222-1223.

Klopfer, E., Reich, J., Abelson, H. & Breazeal, C. (2024). *Generative AI and K-12 Education: An MIT Perspective*.

Kovač, G., Sawayama, M., Portelas, R., Colas, C., Dominey, P. F. & Oudeyer, P. Y. (2023). *Large language models as superpositions of cultural perspectives*. arXiv preprint arXiv:2307.07870

Laak, K. J., & Aru, J. (2024, July). Generative AI in K-12: Opportunities for Learning and Utility for Teachers. In *International Conference on Artificial Intelligence in Education* (pp. 502-509). Cham: Springer Nature Switzerland

Laurençon, H., Saulnier, L., Wang, T., Akiki, C., Villanova del Moral, A., Le Scao, T., ... & Jernite, Y. (2022). The BigScience roots corpus: A 1.6 tb composite multilingual dataset. *Advances in Neural Information Processing Systems*, 35, 31809-31826.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. nature, 521(7553), 436-444.

Li, P., Yang, J., Islam, M. A., & Ren, S. (2023). Making ai less" thirsty": Uncovering and addressing the secret water footprint of ai models. arXiv preprint arXiv:2304.03271.

Macina, J., Daheim, N., Chowdhury, S., Sinha, T., Kapur, M., Gurevych, I. & Sachan, M. (2023). MathDial: A Dialogue Tutoring Dataset with Rich Pedagogical Properties Grounded in Math Reasoning Problems. In F*indings of the Association for Computational Linguistics: EMNLP 2023* (pp. 5602-5621).

Markel, J. M., Opferman, S. G., Landay, J. A. & Piech, C. (2023). Gpteach: Interactive ta training with gpt-based students. In *Proceedings of the tenth acm conference on learning@ scale* (pp. 226-236).

Merchant, A., Batzner, S., Schoenholz, S. S., Aykol, M., Cheon, G. & Cubuk, E. D. (2023). Scaling deep learning for materials discovery. *Nature*, 624(7990), 80-85.

Nie, A., Chandak, Y., Suzara, M., Malik, A., Woodrow, J., Peng, M., ... & Piech, C. (2024). *The GPT Surprise: Offering Large Language Model Chat in a Massive Coding Class Reduced Engagement but Increased Adopters' Exam Performances* (No. qy8zd). Center for Open Science.

Nkambou, R., Mizoguchi, R. & Bourdeau, J. (Eds.). (2010). *Advances in intelligent tutoring systems* (Vol. 308). Springer Science & Business Media.

Noy, S. & Zhang, W. (2023). Experimental evidence on the productivity effects of generative artificial intelligence. *Science*, 381(6654), 187-192.

Oravec, J. A. (2023). Artificial intelligence implications for academic cheating: Expanding the dimensions of responsible human-AI collaboration with ChatGPT. J*ournal of Interactive Learning Research*, 34(2), 213-237.

Orticio, E., Meyer, M. & Kidd, C. (2023). Children flexibly adapt their evidentiary standards to their informational environment. In *Proceedings of the Annual Meeting of the Cognitive Science Society (*Vol. 45, No. 45).

Oudeyer, P. Y., Gottlieb, J. & Lopes, M. (2016). Intrinsic motivation, curiosity, and learning: Theory and applications in educational technologies. *Progress in brain research*, 229, 257-284.

Oudeyer, P. Y., Gottlieb, J. & Lopes, M. (2016). Intrinsic motivation, curiosity, and learning: Theory and applications in educational technologies. *Progress in brain research*, 229, 257-284.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., ... & Lowe, R. (2022). Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35, 27730-27744.

Proust, J. (2019). La métacognition : les enjeux pédagogiques de la recherche, in: S. Dehaene (ed.), *Les sciences au service de l'école*. Odile Jacob.

Park, H., Li, Z. & Walsh, A. (2024). Has generative artificial intelligence solved inverse materials design? *Matter*, 7(7), 2355-2367.

Romera-Paredes, B., Barekatain, M., Novikov, A., Balog, M., Kumar, M. P., Dupont, E., ... & Fawzi, A. (2024). Mathematical discoveries from program search with large language models. *Nature*, 625(7995), 468-475.

Rozado, D. (2022) *The political orientation of the ChatGPT AI system Applying Political Typology Quizzes to a state-of-the-art AI Language model.* https://davidrozado.substack.com/p/the-political-orientation-of-the

Russell, S. J. & Norvig, P. (2016). *Artificial intelligence: a modern approach*. Pearson.

Salesforce Generative AI statistics (2024) https://www.salesforce.com/news/stories/generative-ai-statistics/

Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. *IBM Journal of research and development*, 3(3), 210-229.

Solaiman, I. & Dennison, C. (2021). Process for adapting language models to society (palms) with values-targeted datasets. *Advances in Neural Information Processing Systems*, 34, 5861-5873.

Srivastava, A., Rastogi, A., Rao, A., Shoeb, A. A. M., Abid, A., Fisch, A., ... & Wang, G. X. (2023). *Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models*. Transactions on Machine Learning Research.

Thompson, B. & Griffiths, T. L. (2021). Human biases limit cumulative innovation. *Proceedings of the Royal Society* B, 288(1946), 20202752.

Torres-Leguet A., Romac, C., Carta, T. & Oudeyer, P-Y. (2024) ChatGPT en 5 min : une série pédagogique pour le grand public. https://developmentalsystems.org/chatgpt_5_minutes/fr/

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M. A., Lacroix, T., ... & Lample, G. (2023). *Llama: Open and efficient foundation language models*.arXiv:2302.13971.

Trystram, D., Couillet, R. & Ménissier, T. (2021) *Apprentissage profond et consommation énergétique : la partie immergée de l'IA-ceberg*. The Conversation, https://theconversation.com/apprentissage-profond-et-consommation-energetique-la-partie-immergee-de-lia-ceberg-172341

Walter, W. G. (1950). An electro-mechanical «animal». *Dialectica*, 206-213.

Wang, S., Mitchell, J. & Piech, C. (2024, March). A large scale RCT on effective error messages in CS1. In *Proceedings of the 55th ACM Technical Symposium on Computer Science Education* V. 1 (pp. 1395-1401).

Wu, Y., Jiang, A. Q., Li, W., Rabe, M., Staats, C., Jamnik, M. & Szegedy, C. (2022). Autoformalization with large language models. Advances in Neural Information *Processing Systems*, 35, 32353-32368.

Yovel, Y. & Rechavi, O. (2023). AI and the Doctor Dolittle challenge. *Current Biology,* 33(15), R783-R787.

Zambaldi, V., La, D., Chu, A. E., Patani, H., Danson, A. E., Kwan, T. O., ... & Wang, J. (2024). *De novo design of high-affinity protein binders with AlphaProteo*. arXiv preprint arXiv:2409.08022.