Un an après l'arrivée de ChatGPT:

Réflexions de l'Obvia sur les enjeux et pistes d'action possibles face à l'IA générative

Allison Marchildon
Claire Boine
Andréane Sabourin Laflamme
Dave Anctil
Antoine Boudreau LeBlanc
Sylvain Auclair
Christine Balagué
Félix-Arnaud Morin-Bertrand
Céline Castets-Renard
Philip Jackson
Lyse Langlois

Janvier 2024

obvia

Autrices et auteurs :

Allison Marchildon, Ph.D, professeure au Département de philosophie et d'éthique appliquée de l'Université de Sherbrooke et coresponsable de l'axe Éthique, gouvernance et démocratie de l'Obvia

Claire Boine, LL.M, M.A, candidate au doctorat en droit à l'Université d'Ottawa

Andréane Sabourin Laflamme, M.A, professeure de philosophie au Cégep André Laurendeau et candidate au doctorat en droit à l'Université de Sherbrooke

Dave Anctil, Ph.D, professeur de philosophie au Collège Jean-de-Brébeuf

Antoine Boudreau LeBlanc, Ph.D, scientifique en résidence au Ministère de la Cybersécurité et du Numérique et chercheur adjoint à l'Université Simon Fraser

Sylvain Auclair, M.A, professeur de philosophie au Cégep Sainte-Foy et candidat au doctorat en philosophie à l'Université Laval

Christine Balagué, Ph.D, professeure de management à l'Institut Mines Telecom Business School

Félix-Arnaud Morin-Bertrand, M.A, professionnel de recherche à l'Obvia

Céline Castets-Renard, Ph.D, professeure à la Faculté de droit (Section de droit civil) de l'Université d'Ottawa

Philip Jackson, Ph.D, professeur à l'École de psychologie de l'Université Laval et directeur scientifique Recherche de l'Obvia

Lyse Langlois, Ph.D, professeure au Département de relations industrielles de l'Université Laval et directrice générale de l'Obvia

Remerciements :

Richard Khoury, Ph.D, professeur au Département d'informatique et de génie logiciel de l'Université Laval

Produit avec le soutien financier des Fonds de recherche du Québec

Québec
Fonds de recherche – Nature et technologies
Fonds de recherche – Santé
Fonds de recherche – Société et culture

ISBN : 978-2-925138-29-7
DOI : 10.61737/ZKWZ3721

L'image de couverture a été produite sur Canva à l'aide d'un système d'intelligence artificielle.

Table des matières

Introduction 4

1. Les jalons du développement de l'IA générative 5

1.1 Les débuts 6
1.2 Les modèles de fondation 7
1.3 Le traitement du langage naturel et la naissance des LLMs (large language models) 7
1.4 L'IA générative multimodale – au-delà des mots 8
1.5 De l'évolution de l'IA à une révolution 9

2. Les enjeux de l'IA générative 11

2.1 Les enjeux transversaux de l'utilisation de l'IA générative 12
2.2 Les enjeux dans le monde du travail 14
La perte, le déplacement et la précarité d'emploi 15
La substitution et l'augmentation des travailleurs humains 16
La transparence et l'encadrement 17
2.3 Les enjeux en éducation 19
Des possibilités inédites 19
Un bouleversement des modes d'évaluation 20
Des risques significatifs : fracture, désinformation, biais, etc. 22
2.4 La commercialisation de l'IA générative 23
Une IA générative dans l'intérêt de qui? 24
Quel alignement avec quelles valeurs? 26
Les risques des discours de responsabilité des entreprises d'IA générative 30
2.5 En conclusion sur les enjeux de l'IA générative 32

3. L'encadrement juridique de l'IA générative à ce jour 33

3.1 Les initiatives canadiennes 34
3.2 Les initiatives au sein de l'Union européenne 37
3.3 Des difficultés sans frontières 38

4. Un appel à la vigilance et à l'action 39

4.1 Les initiatives normatives et législatives en matière d'IA 41
Piste d'action 1 : Mettre en place des normes de droit reconnues au niveau national et international 41
Piste d'action 2 : Mettre en place des structures de gouvernance adaptées à la nature évolutive de l'IA 42
Piste d'action 3 : Intégrer des obligations d'audit et de reddition de comptes pour favoriser la responsabilité des entreprises productrices d'IA 43
4.2 Les initiatives éthiques en matière d'IA 44
Piste d'action 4 : S'appuyer sur des initiatives et outils éthiques reconnus 44
Piste d'action 5 : Envisager les initiatives éthiques comme un processus plutôt que comme une liste de principes 44
4.3 Les initiatives de démocratisation de l'IA 46
Piste d'action 6 : Favoriser le développement de compétences pour une participation plus large et plus diversifiée 46
Piste d'action 7 : Favoriser la participation citoyenne dans l'évaluation de l'acceptabilité (sociale) et dans l'orientation des SIA 46

5. Lexique 47

6. Bibliographie 49

Introduction

Depuis l'automne 2022, avec l'arrivée abrupte de ChatGPT, la version gratuite du modèle de langage de l'entreprise OpenAI, les expérimentations à grande échelle du déploiement de ce système d'intelligence artificielle ont suscité un tsunami de réactions dont on perçoit les conséquences autant dans nos vies personnelles que professionnelles. En raison de l'ampleur des implications de cet événement, qui symbolise un changement hautement disruptif pour nos sociétés, il s'avérait incontournable qu'un groupe de chercheuses et de chercheurs de l'Obvia s'y penche afin de contribuer à la réflexion sur le sujet et dégager des pistes d'action.

Le but du présent document vise par conséquent à apporter des éléments de réflexion sur les différents enjeux que soulève l'émergence soudaine de ces technologies au potentiel révolutionnaire. Le présent document vise plus spécifiquement:

- à situer dans son contexte le développement des applications d'IA générative;
- à identifier ce qui nous semble être les principaux enjeux transversaux soulevés par l'arrivée de ces technologies dans nos vies et nos sociétés;
- à esquisser quelques pistes d'action pour tenter de composer adéquatement avec ces enjeux.

Ce document a été rédigé par une équipe multidisciplinaire de chercheuses et de chercheurs membres de l'Obvia. En effet, pour mener une réflexion sur une situation aussi large et complexe, il nous semblait essentiel de réunir des expertes et experts de plusieurs domaines, notamment en éthique appliquée, en neuropsychologie, en droit, en bioéthique et en philosophie. Les différentes thématiques qui émergent de cette réflexion ont été organisées comme suit, mais elles peuvent être découvertes dans l'ordre ou dans le désordre, selon les intérêts de la lectrice ou du lecteur.

1
Les jalons du développement de l'IA générative

1. Les jalons du développement de l'IA générative

1.1 Les débuts

L'histoire de l'**intelligence artificielle (IA)** est marquée par des moments de grandes avancées entrecoupées par des périodes que l'on qualifie d'hivers de l'IA (Russel et Norvig, 2010). Dès 1956, un groupe de dix éminents chercheurs se réunit à Dartmouth avec l'ambition de concevoir une machine dotée d'une intelligence équivalente à celle d'un être humain que l'on qualifierait aujourd'hui d'**intelligence artificielle générale** (***artifical general intelligence* (AGI)** en anglais). Bien que ce projet n'ait toujours pas abouti à un tel objectif, des systèmes spécialisés basés sur l'IA ont néanmoins vu le jour au cours des années suivantes dans des domaines spécifiques et pour réaliser des tâches bien précises telles que jouer aux échecs et réaliser des démonstrations mathématiques.

Au cours des années 1980, et après un premier hiver de l'IA, une forme d'IA nommée « systèmes experts » voit le jour. Ces systèmes étaient basés sur des règles formelles établies manuellement, et donc présentaient des limitations en termes de leurs capacités d'adaptation et d'évolution. Cette période correspond aussi à une remontée du connexionnisme, c'est-à-dire l'étude des processus cognitifs et du comportement humain à l'aide des mathématiques et de réseaux. Ces mouvements parallèles, qui se basent sur des réseaux à la fois pour tenter d'expliquer le cerveau humain et pour tenter de simuler son fonctionnement avec des systèmes, pavent la route au développement de **l'apprentissage automatique**.

L'étude des processus cognitifs et du comportement humain à l'aide des mathématiques et de réseaux a pavé la route au développement de l'apprentissage automatique.

Au milieu des années 1990, grâce à l'augmentation de la puissance de calcul des ordinateurs permettant de gérer de plus grands jeux de données, à la création de jeux de données massifs, notamment grâce à Internet, et à l'amélioration des méthodes statistiques, l'apprentissage automatique se perfectionne et permet aux systèmes qui en sont dotés d'acquérir, par exemple, de nouvelles connaissances en analysant des tendances dans des ensembles de données. Ceci favorise le développement d'une variété d'applications allant de la régression à la classification d'images. Cependant, l'apprentissage automatique classique rencontre également des limites, notamment en matière de reconnaissance d'images complexes. L'**apprentissage profond**, une forme

avancée d'apprentissage automatique inspirée des multiples réseaux de neurones du cerveau, a contribué à surmonter ces défis en créant des couches de représentations simplifiées et interconnectées. Les domaines d'application les plus courants de

l'apprentissage profond sont la vision par ordinateur (p. ex., assistance à la conduite routière, reconnaissance de visages, détection de tumeurs), le traitement audio (p. ex., suppression de bruit), et le traitement du langage naturel (p.ex., traduction, génération du langage par la prédiction).

1.2 Les modèles de fondation

Les **systèmes d'intelligence artificielle (SIA)** qui dominent la recherche de pointe en IA depuis quelques années comme BERT (Google) et GPT (OpenAI) font partie d'une catégorie de systèmes appelée les « **modèles de fondation** » (***foundation models*** en anglais). Ces modèles, aussi appelés systèmes d'**IA à usage général** (***General Purpose AI*** *systems* ou **GPAI**) sont des modèles entraînés sur de larges quantités de données et qui peuvent être adaptés en aval à une grande variété d'applications (Bommasani et al., 2021). Ce type d'apprentissage se base sur une immense quantité de données non structurées, qui sont traitées sur un grand nombre de processeurs, la plupart du temps de type GPU1. Ces modèles de fondation demandent très peu d'intervention humaine, et peuvent apprendre à faire de nouvelles tâches par eux-mêmes (Bommasani et al., 2021). En revanche, ils occasionnent des coûts de calculs et énergétiques impressionnants, qui ne pouvaient, jusqu'à très récemment, être financés que par les laboratoires ou compagnies dotés de budgets substantiels (Amazon, Google, Microsoft ou Meta).

Les modèles de fondation, entraînés sur de larges quantités de données, demandent très peu d'intervention humaine et peuvent apprendre à faire de nouvelles tâches par eux-mêmes.

1.3 Le traitement du langage naturel et la naissance des LLMs (*large language models*)

Issus du perfectionnement des modèles de fondation, les **grands modèles de langage (*large language models* (LLMs)** ont permis à l'IA de faire des bonds de géants en traitement du langage naturel (*Natural Language Processing*, NLP). Les LLMs sont donc une sous-catégorie des modèles de fondation fonctionnant à partir du langage (Toner, 2023). C'est justement grâce aux modèles LLM que l'**IA générative** a connu un bond spectaculaire dans la génération de texte, à l'exemple de ChatGPT développé par OpenAI ou de BARD développé par Google. L'IA générative est ainsi un terme général utilisé pour décrire les SIA ayant comme principale fonction de produire du contenu (Toner, 2023). Ce contenu peut, par exemple, prendre la forme de textes, d'illustrations ou même de sons (Toner, 2023; Stanford HAI, 2023).

Les grands modèles de langage ont permis à l'IA de faire des bonds de géants en traitement du langage naturel, en particulier dans la génération de texte, à l'exemple de ChatGPT.

L'intérêt pour le traitement du langage naturel remonte aux années 1950, avec notamment la proposition du jeu de l'imitation (*Imitation Game)* d'Alan Turing (Turing, 1950), connu depuis sous le nom de test de Turing (*Turing Test)*. Par ce processus, on cherche à produire des systèmes capables de produire le langage comme le fait l'humain. Après plusieurs phases de développement et le perfectionnement de différents systèmes basés sur des règles précises et prédéterminées dont les nombreux robots conversationnels (*chatbots*), l'avènement des modèles de type ***transformers*** a marqué une étape importante dans le domaine du traitement du langage naturel. Ces modèles ont été introduits par Vaswani et al. (2017) de l'entreprise Google Brain et reposent sur un mécanisme d'attention qui permet de traiter les mots d'un texte de manière parallèle plutôt que séquentielle. Lorsqu'ils sont pré-entraînés sur de vastes corpus de textes, ces modèles sont capables de générer des réponses contextuellement appropriées et d'effectuer des tâches complexes liées au langage. Cela a conduit à une amélioration significative, et même surprenante, des performances des SIA dans la réalisation de diverses tâches liées au langage.

Les systèmes d'IA générative ont toutefois suscité un questionnement éthique important, en raison du fait qu'ils miment la parole et le langage, laissant ainsi planer le doute que la machine peut penser.

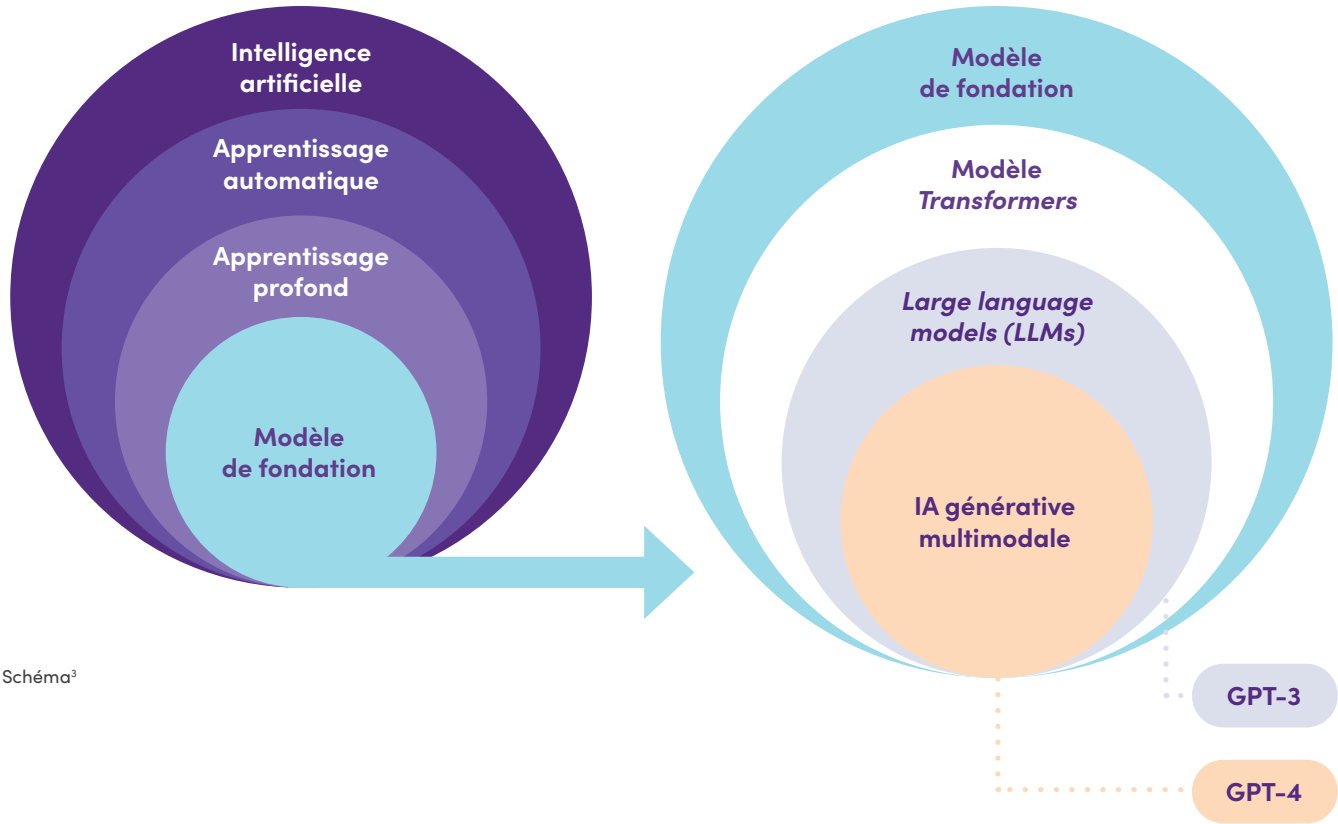
1 Graphics Processing Unit

1.4 L'IA générative multimodale – au-delà des mots

L'amélioration des performances des systèmes d'IA basés sur des modèles de plus en plus avancés dépend principalement de trois facteurs : la puissance de calcul, la taille des jeux de données utilisés et le nombre de paramètres (Kaplan et al., 2020). Lorsque l'on augmente ces trois facteurs, non seulement les performances du système entraîné augmentent de manière significative, mais le système acquiert également des capacités qui n'avaient pas été prévues par les développeurs. Ainsi, bien qu'entraînés pour produire du texte, ces modèles acquièrent toujours plus de capacités, et plusieurs ont appris à faire de l'arithmétique et de l'algèbre à partir de prédictions et de probabilités, et non sur la base d'une série de commandes précises2, un processus que l'on appelle l'émergence (Wei et al., 2022) et sur lequel nous reviendrons plus loin.

Ces SIA sont au cœur de la révolution dite de l'IA générative (Stanford HAI, 2023). Ils peuvent en effet être spécialisés durant la phase finale de réglage (*tuning*) pour accomplir la quasi-totalité des tâches autrefois effectuées par des modèles spécialisés développés par simple programmation ou par **apprentissage supervisé**. Ces nouveaux modèles sont dits **multimodaux** car ils peuvent désormais non seulement comprendre et générer du texte et des images, mais aussi produire du code informatique, résoudre des problèmes mathématiques et scientifiques, concevoir des contenus dans la plupart des domaines techniques ou non techniques, dont la génération de musique, d'image (p. ex., DALL-E (OpenAI) ; Midjourney (laboratoire indépendant)), et de vidéo (OpenAI, 2023a ; Agostinelli et al., 2023).

« ...bien qu'entraînés pour produire du texte, ces modèles acquièrent toujours plus de capacités, et plusieurs ont appris à faire de l'arithmétique et de l'algèbre à partir de prédictions et de probabilités, et non sur la base d'une série de commandes précises, un processus que l'on appelle l'émergence. »



Schéma[3]

2 Voir Elhage et al. (2021) au sujet d'une tentative « d'interprétabilité mécanistique » par ingénierie inversée d'un modèle simple et Ganguli et al. (2022) sur l'impact des caractéristiques imprévisibles.

3 Schéma inspiré par la figure 2.1 dans Boudreau LeBlanc, Monteferrante et Verreault (2021)

1.5 De l'évolution de l'IA à une révolution

Nous faisons donc face à des systèmes technologiques au potentiel révolutionnaire qui ont sans conteste des répercussions sociétales significatives. Il est important de rappeler que ces modèles d'IA générative ont été entraînés à prédire le mot ou groupe de mots le plus probable en fonction du contexte et des mots ou groupes de mots précédents. Il ne s'agit en aucun cas de bases de données ou de moteurs de recherche. En outre, lorsqu'OpenAI a mis ChatGPT sur le marché, ses concepteurs envisageaient des utilisations à visée fictive (par exemple, écrire une histoire sur un thème précis, rédiger un poème dans le style d'une poète ou d'un poète connu, etc.). Cependant, les utilisateurs ont fait usage de ChatGPT comme on utiliserait un moteur de recherche, ce qui a influencé la direction du développement de ces systèmes. En conséquence, OpenAI a dû travailler à améliorer la véracité des sorties du système et a donné accès à Internet à ChatGPT, afin que le système puisse effectuer des recherches factuelles.

« Il est important de rappeler que ces modèles [...] ont été entraînés à prédire le mot ou groupe de mots le plus probable en fonction du contexte et des mots ou groupes de mots précédents. Il ne s'agit en aucun cas de bases de données ou de moteurs de recherche. »

Pour suivre la progression de ces nouveaux modèles d'IA générative, 132 institutions de recherche ont joint leurs efforts pour développer un processus d'évaluation appelé *Beyond the Imitation Game Benchmark* (Srivastava et al., 2022). Aussi appelé BIG-bench, ce point de référence (*benchmark*) est constitué d'un ensemble de tests couvrant plus de 200 tâches permettant d'évaluer la performance des modèles au moyen d'une large gamme de problèmes à partir desquels des échelles de comparaison humaine sont établies. Les tests standardisés peuvent ainsi mesurer le niveau de compétence des modèles dans des tâches sollicitant différentes formes du raisonnement humain (linguistique, mathématique, création, programmation et sens commun), ainsi que la performance à de nombreux autres tests appliqués au développement cognitif, psychologique et social (Srivastava et al., 2022). Mesurés via de tels tests, les modèles d'IA générative d'aujourd'hui se révèlent en mesure de résoudre de plus en plus de problèmes qui ont été opérationnalisés par des tâches précises et validées. Cette progression fulgurante de la performance des modèles d'IA générative au cours des dernières années surprend au point où elle soulève des questions fondamentales, à savoir si nous pouvons prédire et opérationnaliser toutes les tâches effectuées par les êtres humains.

« Cette progression fulgurante de la performance des modèles d'IA générative au cours des dernières années surprend au point où elle soulève des questions fondamentales, à savoir si nous pouvons prédire et opérationnaliser toutes les tâches effectuées par les êtres humains. »

À cet égard, les recherches récentes suggèrent un phénomène d'***émergence***, qui pourrait expliquer le développement de l'intelligence instrumentale sophistiquée que nous observons. En physique et en biologie, on parle d'***émergence*** lorsque l'accumulation de changements locaux atteint un seuil (quantitatif et mécanistique) suffisant dans un système pour modifier les propriétés (qualitatives) de l'ensemble de l'organisation (Steinhardt, 2022). En apprentissage profond, qui est un autre type de système reposant sur la complexité, ce phénomène est observé lorsque la taille des modèles et la quantité de données, mesurées en milliards de paramètres et en milliards de *tokens*, dépassent certains seuils : ainsi, de nouvelles capacités qui n'ont pas été prédites à partir des techniques mécanistiques classiques émergent du fonctionnement des réseaux neuronaux (Wei et al. 2022).

Or, le développement de critères de référence (*benchmarks)* dépend de nos connaissances de la cognition et du cerveau humain ainsi que des techniques issues des statistiques classiques. Ainsi, les nouvelles possibilités technologiques des SIA deviennent peu explicables (Mitchell et Krakauer, 2023) puisque d'une part nos connaissances de l'intelligence humaine ne sont pas absolues et d'autre part ces benchmarks ne sont pas développés pour évaluer différentes formes d'intelligences.

Ce que nous savons toutefois avec certitude, c'est que les récents modèles d'IA générative dépassent déjà les compétences des modèles antérieurs évalués par les équipes de BIG-bench et du laboratoire Google DeepMind en 2022 (Wei et al. 2022). Par exemple, le rapport technique d'OpenAI sur le modèle GPT-4, ainsi qu'une étude préliminaire effectuée sur ce même modèle par le laboratoire Microsoft Research, démontrent que GPT-4 peut résoudre les tâches du BIG-bench à un niveau humain, ainsi qu'un ensemble d'autres tâches dans des domaines aussi variés que la reconnaissance visuelle, la programmation, les mathématiques; il est même parvenu à exceller à certains types d'examens standardisés anglais et américains de médecine, de droit, de psychologie et bien d'autres encore (OpenAI, 2023b; Bubek et al., 2023). En plus de démontrer les capacités inédites des récents modèles d'IA générative, ces résultats viennent ainsi remettre en question ce type de tests qui semblent davantage évaluer une forme d'intelligence instrumentale que la singularité de l'intelligence et des connaissances humaines.

L'objet du présent document de réflexion n'est pas d'approfondir cette comparaison entre l'humain et la machine. En revanche, ce débat met en lumière le potentiel transformateur de l'IA face aux interactions entre humain et machine au niveau de l'acquisition et la consolidation des connaissances. Il interroge ce faisant la responsabilité éthique et sociale des développeurs d'IA et des entreprises à but lucratif qui en soutiennent le développement et qui sont engagés dans ce que l'on peut qualifier de course effrénée vers une **IA générale**. Les prochaines sections mettront par conséquent en évidence certains des enjeux reliés à l'utilisation et à la commercialisation de l'IA générative.

2
Les enjeux de l'IA générative

2. Les enjeux de l'IA générative

2.1 Les enjeux transversaux de l'utilisation de l'IA générative

Une récente étude menée par la firme KPMG permet d'observer que les Canadiens ont rapidement intégré l'IA générative à leurs pratiques et que tout indique que cette tendance va en s'accélérant. Selon le portrait dressé par cette étude, 20 % des Canadiens sondés affirment utiliser des systèmes d'IA générative dans le cadre de leur travail ou de leurs études (KPMG, 2023a).

Or, les enjeux liés à l'IA générative sont nombreux (Yip et Balagué, 2023) et sont loin d'être anodins. On déplore notamment que pour entraîner ces modèles de langage, plusieurs bases de données aient été utilisées, sans nécessairement respecter les droits de **propriété intellectuelle**. Des poursuites ont d'ailleurs déjà été intentées contre OpenAI et d'autres acteurs majeurs de l'IA à cet égard (Davis, 2023; Mangan, 2024).

D'autre part, ces modèles comportent plusieurs biais et ne sont dès lors pas neutres. En effet, les contenus générés varient par exemple en fonction du langage utilisé pour effectuer une requête (Zhuo et al., 2023), en plus de mettre de l'avant certaines inclinaisons politiques (Hartmann et al., 2023; Rozado, 2023). On soulève par ailleurs de plus en plus le fait que les données utilisées pour entraîner les modèles d'IA générative peuvent reproduire des préjugés et stéréotypes sexistes et raciaux et ainsi générer des résultats biaisés pouvant mener à différentes formes de **discrimination**, automatisant et amplifiant ce faisant les **inégalités** sociohistoriques existantes. En ce sens, il a été démontré qu'il était facile d'orienter les requêtes (*prompts*) vers des réponses stéréotypées et biaisées (Gallienne et Poibeau, 2023) et que les systèmes d'intelligence artificielle étaient susceptibles de générer des résultats potentiellement discriminatoires envers certains individus, groupes ou communautés, notamment envers les femmes (Conseil du statut de la femme, 2023).

Il importe par conséquent de travailler à identifier, anticiper et minimiser les risques de biais injustes et potentiellement discriminatoires associés à ces systèmes, mais cette tâche soulève néanmoins d'importants défis : comment et sur la base de quels critères ou normes faut-il identifier et corriger ces biais? Et qui peut ou doit décider quels correctifs devront être apportés?

De plus, les récents modèles d'IA générative, tels que ChatGPT, peuvent être facilement utilisés pour accomplir des **actions malveillantes**, et ce, malgré les filtres « éthiques » qui y ont été ajoutés, puisque ceux-ci peuvent être facilement contournés par les personnes utilisatrices pour générer du contenu considéré toxique (Borji, 2023; Zhuo et al., 2023) ou encore des incitations à commettre des actes criminels (Borji, 2023; Christian, 2023; Franceschi-Bicchierai, 2023; Murgia, 2023; Smalley, 2023). On entrevoit aussi d'importants risques de manipulation de l'information ou de l'opinion publique, qui entraîneraient à leur tour la désinformation. En effet, compte tenu de la capacité impressionnante de l'IA générative à produire de grandes quantités de contenus, que ce soit dans des textes courts (par exemple, des messages sur les médias sociaux) ou des textes plus longs (par exemple, des rapports ou des discours politiques), il en résulte que ces informations peuvent être utilisées à des fins nuisibles ou trompeuses. Or, ces informations trompeuses s'avèrent difficiles à détecter en temps utile, notamment sur les réseaux sociaux, où la modération des contenus (factchecking) ne peut parvenir à vérifier efficacement les centaines de millions de messages qui y sont véhiculés chaque minute (Masnick dans Fernández Gibaja, 2023).

Les conséquences d'une telle manipulation de l'information peuvent évidemment s'avérer dévastatrices, notamment au niveau individuel, pour la réputation des personnes au sujet desquelles de fausses informations peuvent être générées et véhiculées. La *Federal Trade Commission* américaine a d'ailleurs ouvert une enquête sur OpenAI afin de déterminer si les lois assurant la protection des consommateurs avaient été transgressées par les façons de faire d'OpenAI, notamment en matière de sécurité et de fausses informations pouvant entraîner des dommages réputationnels (Kang et Metz, 2023). Mais ces conséquences peuvent aussi s'étendre à une échelle beaucoup plus large, puisque l'IA générative peut, si elle prend la place des journalistes et autres humains chargés de vérifier et transmettre l'information à la population, comme cela a été observé notamment chez MSN (CNN, 2023), mener à l'exposition de centaines de millions de personnes à des contenus parfaitement faux, mensongers ou propagandistes. Elle peut par ailleurs devenir un véhicule permettant de multiplier de manière exponentielle la portée des opérations d'influence (partis politiques, candidats, gouvernements étrangers, groupes de pression, organisations sociales ou entreprises privées) qui chercheraient à manipuler et à tromper l'opinion publique à des fins électorales ou idéologiques (Fernández Gibaja, 2023).

On comprendra donc que les conséquences de l'IA générative sur la démocratie peuvent s'avérer considérables, que ce soit par leur rôle potentiel pour influencer de façon directe ou abusive l'opinion publique, ou pour décupler indûment la voix de certains acteurs, comme celle des entreprises ou de leurs lobbyistes, qui peuvent recourir à l'IA générative pour rédiger automatiquement des commentaires soumis dans le cadre de processus réglementaires ou des lettres à l'éditeur pour publication dans les journaux locaux, ou encore commenter des millions de fois par jour des articles de presse, des articles de blog et des messages sur les médias sociaux, en donnant l'impression qu'ils représentent l'opinion publique (Sanders et Schneider, 2023). Ce faisant, la flexibilité des outils d'IA générative pourrait permettre d'exercer une influence sur de nombreuses politiques et juridictions simultanément, avec toute l'érosion de la démocratie que cela entraîne.

« Les conséquences de l'IA générative sur la démocratie peuvent s'avérer considérables, que ce soit par leur rôle potentiel pour influencer de façon directe ou abusive l'opinion publique, ou pour décupler indûment la voix de certains acteurs, comme celle des entreprises ou de leurs lobbyistes. »

Ajoutons à ces enjeux celui des **impacts environnementaux** non négligeables de l'IA, qui commence à peine à faire partie des préoccupations liées à l'IA, et ce malgré son importance. En effet, de plus en plus de chercheuses et chercheurs mettent en lumière les impacts de l'IA sur l'environnement, qu'il s'agisse des quantités élevées d'eau ou d'énergie utilisées pour refroidir les serveurs (nombre de processeurs, énergie pour les faire fonctionner et les refroidir, etc.), ou leur empreinte carbone plus générale4.

Bien que difficile à calculer avec précision, l'empreinte carbone se révèle considérable : pour l'entraînement de certains modèles d'IA générative, elle est estimée jusqu'à plus de cinq fois les émissions produites dans l'ensemble du cycle de vie d'une voiture moyenne (Borji, 2023).

Si ce sont là les enjeux les plus fréquemment soulevés en ce qui concerne l'IA générative, cette liste pourrait se poursuivre sur plusieurs pages, et elle s'allongera par ailleurs au fur et à mesure du développement et du perfectionnement de ces modèles. Dans ce document, l'objectif n'est ainsi pas de dresser une liste exhaustive des enjeux liés à l'IA générative, mais plutôt d'en souligner la multitude et l'importance d'y être attentif afin d'informer nos actions et décisions face à l'utilisation de l'IA générative. Dans cet esprit, nous mettrons dans ce qui suit en évidence certains enjeux qui apparaissent particulièrement saillants en raison des changements fondamentaux qu'ils entraînent relativement à nos modes de fonctionnement, à notre rapport au monde, à notre accès à l'information et au savoir ainsi qu'à nos modes d'interaction. Ils s'avèrent dans certains cas moins visibles et pourraient mettre plus de temps à se matérialiser, mais engendreront néanmoins dans leur foulée nombre d'autres enjeux. Nous nous attarderons ainsi plus particulièrement aux enjeux de l'utilisation de l'IA générative dans deux sphères d'activité qui sont susceptibles d'être particulièrement impactées par les progrès récents de l'IA générative, soit le monde du travail et de l'éducation, et nous envisagerons les enjeux liés à la commercialisation de l'IA, qui sont moins discutés, mais dont les implications sont tout aussi importantes.

« De plus en plus de chercheuses et chercheurs mettent en lumière les impacts de l'IA sur l'environnement, qu'il s'agisse des quantités élevées d'eau ou d'énergie utilisées pour refroidir les serveurs [...] ou leur empreinte carbone plus générale. »

2.2 Les enjeux dans le monde du travail

Selon plusieurs expertes et experts, le recours croissant à l'IA générative dans les différents milieux de travail provoquera une transformation accélérée des modes de fonctionnement des organisations et des tâches des personnes qui y travaillent (Eloundou et al., 2023). En effet, compte tenu du fait que les développements récents en matière d'IA générative permettent dorénavant d'automatiser des tâches qui jusqu'à tout récemment étaient l'apanage des humains, nous pouvons prévoir que le marché du travail connaîtra au cours des prochaines années de profonds changements. Si de nombreuses études estiment que le recours à l'IA générative permettra d'augmenter la productivité dans certains contextes (Briggs et Kodnani, 2023), les chercheurs estiment que les impacts de l'IA générative seront assurément à géométrie variable selon les différents milieux de travail et auront des effets différents sur les travailleurs en fonction de leur champ et de leur niveau d'expertise (Chui et al., 2023).

4 Voir notamment Luccioni, Viguier, S., et Ligozat. (2022) sur l'empreinte carbone et Li et al. (2023) sur les dépenses en eau des modèles de langage.

À la lumière de l'intégration rapide des outils d'IA générative dans les milieux d'emploi (Goldberg, 2023) et face à la multiplication de la mise sur le marché de produits qui reposent sur l'IA générative et qui sont destinés aux différents milieux de pratique (Lu, 2023), il est possible d'anticiper que le recours à ces dispositifs technologiques transformera la manière dont nous travaillons ainsi que les compétences professionnelles qui seront recherchées au cours des prochaines années. Bien qu'il soit impossible de prédire avec précision les effets de l'intégration de l'IA générative dans les pratiques professionnelles, il est possible d'anticiper certains impacts, notamment en termes de nombre d'emplois et du besoin d'encadrement de ces systèmes, ainsi que les risques qui sont susceptibles de leur être associés.

La perte, le déplacement et la précarité d'emploi

On estime que l'IA générative pourrait remplacer jusqu'à un quart des tâches professionnelles effectuées actuellement : cette automatisation équivaudrait à 300 millions d'emplois à temps plein dans le monde (Briggs et Kodnani, 2023; Brynjolfsson et al., 2023). Plusieurs études mettent l'accent sur l'augmentation de la productivité et des économies que l'IA permettra en matière de coûts de main-d'œuvre (Noy et Zhang, 2023). Mais il importe également de rendre compte des risques d'instabilité pour les travailleurs touchés par cette disruption du marché du travail ainsi que d'éventuelles pertes et déplacements d'emplois. Même si l'intégration de l'IA générative dans les différents milieux de pratique est susceptible d'être la source de nouveaux emplois, qui à terme pourraient remplacer les emplois initialement déplacés ou supprimés, plusieurs travailleurs risquent d'être touchés par la transformation du marché.

Cela dit, les études divergent sur les emplois qui seraient les plus menacés. Selon les recherches effectuées par l'Institut McKinsey, les emplois les plus vulnérables sont ceux liés au soutien administratif, au service à la clientèle, à la vente, aux services de restauration et à la production (Gmyrek et al., 2023). Ces postes étant majoritairement occupés par des femmes et des minorités visibles, ces groupes risquent d'être davantage précarisés. L'étude de McKinsey conclut que ce sont les personnes qui occupent des emplois moins bien rémunérés ou exigeant un niveau d'éducation plus faible qui risquent d'être les plus touchées par les éventuelles pertes et déplacements d'emplois provoqués par la délégation de tâches à l'IA générative (Ellingrud et al., 2023).

Par ailleurs, compte tenu de l'évolution constante et imprévisible des modèles de fondation et des produits qui y sont associés, ainsi que de la possible dépendance des organisations qui intègrent ces produits à leurs pratiques, il est probable que les travailleuses et travailleurs devront continuellement s'ajuster à ces évolutions dans les années à venir, ce qui pourrait accentuer leur précarité (Williams, 2023). Notamment, selon Felten et al. (2023), ce sont les emplois nécessitant un certain niveau de compétences intellectuelles, par exemple des emplois en éducation, qui sont présentement les plus exposés à l'IA générative et qui pourraient être le plus affectés dans un avenir rapproché.

« … ce sont les personnes qui occupent des emplois moins bien rémunérés ou exigeant un niveau d'éducation plus faible qui risquent d'être les plus touchées par les éventuelles pertes et déplacements d'emplois provoqués par la délégation de tâches à l'IA générative (Ellingrud et al., 2023). »

Il faut aussi souligner que la production des systèmes d'IA repose sur du travail humain. Aussi, les conditions de travail des personnes qui conçoivent et entraînent ces systèmes, de même que celles qui classent et étiquettent les données utilisées à cet effet sont un élément important de la réflexion sur les effets de l'IA générative sur le monde du travail. Or, il a récemment été révélé (Perrigo, 2023) que trop souvent, ce sont des personnes sous-payées qui sont engagées pour identifier le contenu indésirable (violent, sexuel ou haineux) issu des modèles d'IA générative et en rendre le produit plus acceptable, mettant ainsi en évidence des formes d'exploitation derrière les modèles d'IA générative. Face à ce paradoxe, on peut à juste titre se questionner sur les véritables bénéficiaires de ces technologies et sur les risques que certains groupes en deviennent les victimes collatérales. Il est donc nécessaire d'envisager des mesures pour protéger les travailleurs les plus vulnérables aux changements imminents du marché du travail. Cela appelle ainsi la nécessité de définir les responsabilités des entreprises qui produisent les modèles d'IA générative, celles qui les adoptent ainsi que des obligations des gouvernements à cet égard (Goldberg, 2023).

« … on peut à juste titre se questionner sur les véritables bénéficiaires de ces technologies et sur les risques que certains groupes en deviennent les victimes collatérales. Il est donc nécessaire d'envisager des mesures pour protéger les travailleurs les plus vulnérables aux changements imminents du marché du travail. »

La substitution et l'augmentation des travailleurs humains

En matière d'utilisation de l'IA générative dans le monde du travail, le déploiement du vaste programme Copilot, de Microsoft, le 1er novembre 2023, fait en sorte que le modèle GPT-4 sera dorénavant intégré à la suite Office 365 pour les grandes entreprises achetant un minimum de 300 licences. L'implantation de Copilot signifie que les outils les plus utilisés par les professionnels de ces grandes organisations seront dorénavant « assisté » par la puissance de GPT-4, et ce, sans même qu'ils aient à changer d'application.

Microsoft présente Copilot comme un simple « outil d'augmentation du travail » qui s'inscrirait dans un continuum naturel consistant à augmenter la performance des outils disponibles et donc la productivité et la qualité du travail. Cependant, cette augmentation par l'IA générative risque de se traduire en réalité par une substitution accélérée de l'intelligence humaine. Cette substitution soulève une foule d'enjeux critiques en matière de droit du travail, de sécurité industrielle, d'intégrité et de responsabilité professionnelle.

Comme nous l'avons montré en première partie, l'IA générative est capable de se substituer à la cognition humaine pour accomplir un nombre croissant de tâches, en partie ou en totalité, à un niveau humain ou supérieur; il ne s'agit donc pas d'un simple outil comme des logiciels ou une calculatrice, qui suppose le contrôle et la compétence approfondie des utilisateurs dans l'accomplissement de tâches spécifiques. Un acteur – une organisation ou un professionnel – choisissant de confier des tâches au « copilote » choisit également de faire confiance aux capacités du SIA, et ce, possiblement pour certaines de ses tâches le plus fondamentales ou spécialisées. Or, si la relation du professionnel avec le SIA « copilote » lui permet d'augmenter sa productivité, et donc sa valeur utile et marchande, cela engendre en contrepartie plusieurs enjeux. En effet, le professionnel sera incité à confier un nombre proportionnel de tâches à son « copilote », mais tendra aussi à ne plus veiller à la supervision de ces dernières. À long terme, les professionnels risquent de se déresponsabiliser face à des aspects essentiels de leur travail, mais sans toutefois pouvoir déléguer leur responsabilité pour les erreurs coûteuses et dangereuses à leur « copilote », ce qui peut évidemment entraîner de graves conséquences systémiques.

Les professionnels pourraient payer cher l'augmentation de leur productivité en perdant graduellement des compétences durement acquises, par la formation et l'expérience de travail, et ainsi contribuer à la diminution de valeur de ces compétences sur le marché du travail. Dans un contexte où les organisations sont déjà incitées à désinvestir de la formation continue et la requalification de leurs professionnels; la compétition les incitant plutôt à chercher des gains en productivité rapide, sans considérer les risques à long terme pour les travailleurs et les effets imprévus sur la société, cela s'avère être un autre risque systémique important.

Par conséquent, en sus des questions techniquement complexes de cybersécurité soulevées par les modifications en profondeur des fonctionnements des programmes Microsoft qu'entraîne le déploiement de Copilot, une telle intégration de l'IA générative aux outils informatiques que nous utilisons au quotidien risque d'affecter en profondeur la vaste majorité des écosystèmes de travail et des modes de fonctionnement organisationnels.

Or, si ces dynamiques sont prévisibles, elles sont cependant encore très mal connues. En particulier, l'augmentation-substitution du travail humain par l'IA générative risque fortement de porter atteinte à l'autonomie et l'intégrité professionnelles. La transformation annoncée du travail remet aussi profondément en question la structure légale des responsabilités et des droits du travail. Conséquemment, les organisations professionnelles et syndicales sont fortement interpellées par ces enjeux. Si elles sont à l'heure actuelle encore peu invitées à participer aux discussions avec les entreprises à cet égard, les organisations professionnelles et syndicales doivent néanmoins se pencher sérieusement sur ces risques et conséquences incertaines, comme le fait notamment le Conseil interprofessionnel du Québec, avant d'autoriser la substitution des tâches et des responsabilités de leurs membres.

Les organisations professionnelles et syndicales doivent se pencher sérieusement sur les risques et conséquences incertaines liés à l'augmentation-substitution du travail humain par l'IA générative, avant d'autoriser la substitution des tâches et des responsabilités de leurs membres.

« ... il apparaît important que les organisations mettent en place des mesures qui permettent davantage de transparence et un encadrement adéquat de l'IA générative au travail. »

La transparence et l'encadrement

De nombreuses entreprises, y compris des grandes banques, des hôpitaux et des géants du numérique tels qu'Apple, Samsung et Amazon, ont pour leur part interdit à leurs équipes d'utiliser les outils d'IA générative (Yu, 2023), principalement pour des raisons de confidentialité et de protection des secrets commerciaux. Cependant, d'autres organisations ont déjà intégré ces outils à leurs pratiques et les emploient pour effectuer diverses tâches, comme la création de contenu, la programmation, le service client ou en ressources humaines (Lu, 2023). De plus, de nombreux employés utilisent aussi des outils d'IA générative dans leurs tâches quotidiennes, et plusieurs le font d'ailleurs sans en aviser leurs supérieurs (Cardon et al., 2023). Malgré cet usage répandu, plusieurs entreprises n'ont toujours pas adopté de directives claires concernant l'utilisation de l'IA générative en milieu professionnel (Cardon et al., 2023). Compte tenu des limites et des failles de ces systèmes, mais aussi des enjeux éthiques et juridiques spécifiques associés à leur utilisation dans différents contextes, il apparaît important que les organisations mettent en place des mesures qui permettent davantage de transparence et un encadrement adéquat de l'IA générative au travail.

Par ailleurs, l'utilisation de l'IA générative à des fins de recrutement est de plus en plus fréquente, notamment pour trier des curriculums vitae des candidats à l'emploi, pour planifier et même pour conduire des entrevues d'embauche. L'IA est utilisée depuis plusieurs années en ressources humaines, et les avancées récentes en IA générative font que de plus en plus d'entreprises utilisent par exemple des agents conversationnels ou des assistants virtuels pour optimiser leurs processus de recrutement (Shrivastava, 2023). Les risques de biais et de discrimination relatifs au recours à l'IA en matière de recrutement sont désormais bien connus (Legros et Balagué, 2023), et malgré les efforts pour les éliminer au sein des nouveaux modèles d'IA générative, force est de constater que le problème demeure entier (Nicoletti et Bass, 2023) et que l'utilisation croissante de ces outils à des fins d'embauche pourrait amplifier les risques existants. Par exemple, le recours à des robots conversationnels en contexte d'embauche pourrait s'avérer problématique pour les personnes en situation de handicap, les personnes qui ne maîtrisent pas parfaitement la langue de conversation du robot, les personnes plus âgées ou encore celle en marge de la société qui ne répondent pas à la norme et donc à ce qui est attendu par l'agent conversationnel, même si cela n'est pas lié à leurs compétences professionnelles (Shrivastava, 2023). Ainsi, les risques de discrimination, qui sont de plus en plus connus en IA, pourraient se reproduire et s'amplifier dans le contexte de l'utilisation de l'IA générative à des fins de recrutement.

Bref, en ce qui a trait au domaine du travail, nous devons nous questionner sur la répartition des pouvoirs pour déterminer l'avenir de l'organisation du travail. À l'heure actuelle, une grande partie de ces pouvoirs sont dans les mains des grandes entreprises technologiques au profit des acteurs qui subissent les effets du développement de l'IA sur le travail humain et la transformation de leurs compétences. Il est donc impératif d'envisager comment nous pouvons rééquilibrer ces pouvoirs, afin notamment de mieux protéger les travailleurs les plus vulnérables dans ce contexte de transformation rapide et profonde du marché du travail. À ce titre, il faut s'intéresser, au-delà des promesses d'augmentation de productivité et d'optimisation des processus, aux risques de reproduction ou d'accentuation des dynamiques asymétriques de pouvoir existantes, de précarisation, de marginalisation, d'exploitation et de discrimination en emploi, mais aussi aux responsabilités qui doivent être assumées à cet égard, notamment par nos gouvernements.

À l'heure actuelle, une grande partie des pouvoirs sont dans les mains des grandes entreprises technologiques. Comment rééquilibrer ces pouvoirs, afin notamment de mieux protéger les travailleurs les plus vulnérables dans ce contexte de transformation rapide et profonde du marché du travail?

2.3 Les enjeux en éducation

L'éducation est l'un des secteurs dans lesquels les impacts des récents progrès en IA générative se font le plus sentir. Ces avancées permettent d'envisager des opportunités inédites, comme la personnalisation des apprentissages, le suivi ciblé des étudiants en difficulté et l'optimisation des processus institutionnels. Or, l'intégration de l'IA générative en éducation soulève également de nombreux enjeux, notamment en ce qui a trait à l'intégrité académique et intellectuelle, à la fracture numérique, à la désinformation, aux biais, à l'uniformisation des contenus et à la dépendance aux géants du numérique. La performance remarquable de modèles de fondation comme GPT 3.5, et a fortiori son successeur GPT-4, pour la réalisation de tâches complexes en contexte éducatif (allant de la synthèse, au résumé, à la génération de contenu textuel, visuel, audio et vidéo jusqu'à la programmation) a suscité des réactions mitigées dans le milieu de l'éducation. Il y a, d'une part, un engouement considérable autour des possibilités pédagogiques offertes par ces technologies, mais il existe d'autre part beaucoup de craintes et d'appréhensions quant aux risques associés à l'intégration de l'IA générative dans le milieu académique et scolaire. Quelle que soit la posture adoptée quant à la valeur des outils d'IA générative en enseignement, l'émergence de ces systèmes exige une réflexion et une adaptation des pratiques, en particulier en ce qui a trait à l'évaluation des apprentissages et des compétences.

« Quelle que soit la posture adoptée quant à la valeur des outils d'IA générative en enseignement, l'émergence de ces systèmes exige une réflexion et une adaptation des pratiques, en particulier en ce qui a trait à l'évaluation des apprentissages et des compétences. »

Des possibilités inédites

L'IA s'est imposée de manière croissante dans le secteur de l'éducation au cours de la dernière décennie. Comme en témoignent plusieurs études (Zawacki-Richter et al., 2019), des établissements d'enseignement utilisent déjà l'IA depuis un certain temps, par exemple dans les processus d'admission ou pour estimer les chances de réussite ou de décrochage des étudiants. Le recours à des systèmes de tutorat intelligent et à des systèmes d'IA pour automatiser l'évaluation et la rétroaction ou encore pour personnaliser ou adapter l'enseignement existait dans les milieux éducatifs bien avant la mise en marché de ChatGPT et des autres modèles d'IA générative au cours des deux dernières années. Or, avec les avancées récentes en IA générative, le champ des possibilités s'élargit d'autant plus et offre de nouvelles opportunités. D'ailleurs, l'UNESCO a récemment publié un guide sur les différents usages que l'on peut faire de ChatGPT en enseignement supérieur (Sabzalieva et Valentini, 2023). À titre d'exemple, les enseignants peuvent utiliser ChatGPT pour les soutenir dans l'élaboration de leurs cours, que ce soit pour créer des questions de discussion, des évaluations ou des exercices interactifs. Ils peuvent aussi l'utiliser pour générer des idées novatrices en matière de stratégies pédagogiques. Les étudiants peuvent pour leur part utiliser ChatGPT comme « opposant socratique » afin de développer et de structurer leurs arguments en vue d'un débat ou d'une discussion, comme tuteur personnel pour les aider à comprendre des notions complexes ou encore afin d'évaluer leur niveau de compréhension de la matière préalablement à une évaluation.

En août dernier, OpenAI a publié son propre guide pour l'utilisation de ChatGPT en classe (2023c). Parmi les usages proposés, il y a la génération de plans de leçons, d'exemples ou d'analogies, de jeu-questionnaire et de jeux de rôle. OpenAI propose même de recourir à ChatGPT comme outil d'inclusion, facilitant la traduction ou la conversation pour les étudiants dont la langue maternelle n'est pas celle de l'enseignement. Ces exemples, qui illustrent la polyvalence des grands modèles de langage en pédagogie, ne représentent qu'une fraction des différents usages que l'on peut faire de l'IA générative en enseignement, et rappelons que les nombreux autres modèles basés sur la génération de contenu visuel, audio et vidéo offrent aussi d'innombrables possibilités. Ces modèles de fondation sont aussi utilisés pour développer des applications destinées à des usages spécifiques en éducation, laissant présager une vague de nouveaux produits d'IA générative pour le secteur éducatif. Malgré l'engouement autour des possibilités inédites qu'offre l'IA générative en éducation, il convient de rappeler que les études sur les impacts du recours à l'IA générative en pédagogie demeurent assez peu nombreuses et donc que l'efficacité de ces outils pour soutenir les apprentissages et la réussite reste à démontrer.

« … les études sur les impacts du recours à l'IA générative en pédagogie demeurent assez peu nombreuses et […] l'efficacité de ces outils pour soutenir les apprentissages et la réussite reste à démontrer. »

Un bouleversement des modes d'évaluation

Nonobstant cet enthousiasme, il est indéniable que la mise en marché de ChatGPT a secoué le milieu de l'éducation. Il faut sans doute rappeler que de nombreux étudiants, bien avant le lancement commercial du modèle d'OpenAI en 2022, utilisaient déjà l'IA dans le cadre de leurs travaux scolaires, par exemple pour l'édition et la correction de leurs textes (Anctil, 2023a), mais la démocratisation de ChatGPT, son succès commercial retentissant ainsi que ses performances impressionnantes dans la réalisation de tâches académiques font qu'il est désormais impossible d'ignorer les enjeux et problèmes éthiques qu'engendrent l'utilisation de l'IA générative en contexte éducatif. Des expériences rudimentaires avec ChatGPT révèlent sa capacité à produire en quelques secondes des travaux entiers de qualité allant d'acceptable à très bonne, selon la matière, le niveau ou le type d'évaluation. GPT-4, le nouveau modèle d'OpenAI, disponible sur abonnement payant seulement, est encore plus performant. Par exemple, il a non seulement excellé aux examens d'entrée des écoles de droit américaines, mais il a en plus passé haut la main les examens du Barreau de divers États, sans compter les examens standardisés de différentes disciplines universitaires, qui vont de la médecine à la psychologie. Ces résultats stupéfiants exigent que nous repensions à la manière dont nous évaluons l'acquisition des compétences académiques.

Selon un récent sondage effectué par la firme KPMG, 52 % des étudiants canadiens âgés de 18 ans et plus ont fait appel à l'IA générative pour effectuer leurs travaux académiques, et ce même si 60 % d'entre eux considèrent que son utilisation constitue une forme de plagiat ou de tricherie. Selon cette étude, ceux qui utilisent l'IA dans le cadre de leurs études le font principalement pour générer des idées (63 %), pour faire de la recherche (53 %), pour rédiger des essais ou des rapports (36 %) pour la production de résumés et pour l'analyse d'informations (KPMG, 2023a). Il règne par ailleurs actuellement une grande confusion quant à la perception et l'utilisation des outils d'IA générative par les étudiants dans le cadre de leurs travaux scolaires : « à peine le tiers (36 %) des étudiants disent à leurs éducateurs qu'ils utilisent des outils d'IA générative, et la plupart d'entre eux ne savent pas quelles sont les politiques de leur école ou s'il y a même des répercussions à les utiliser » (KPMG, 2023b). Afin d'éviter que les étudiants ne confient massivement à l'IA générative la réalisation de leurs travaux académiques, certains établissements d'enseignement ont mis à jour leurs politiques institutionnelles, associant automatiquement l'usage de l'IA générative à des actes de fraude ou de plagiat, et prévoyant des sanctions en conséquence. Or, face à la difficulté, voire à l'impossibilité, de détecter de manière efficace et certaine le contenu généré par l'IA (Sadasivan et al., 2023) – les logiciels de détection du contenu généré par l'IA s'avérant aussi inefficaces que la lecture du professeur le plus aguerri (Sabourin Laflamme, 2023) – la tendance semble s'inverser. Compte tenu du nombre important de faux positifs générés par ces modèles, ainsi que des biais qu'ils semblent présenter envers les étudiants qui écrivent dans une autre langue que leur langue maternelle (Liang et al., 2023), les établissements d'enseignement tendent à passer d'une approche répressive à une approche préventive (Bussières McNicoll, 2023) en matière d'intégrité intellectuelle.

Face à la difficulté, voire à l'impossibilité, de détecter de manière efficace et certaine le contenu généré par l'IA, les établissements d'enseignement tendent à passer d'une approche répressive à une approche préventive (Bussières McNicoll, 2023) en matière d'intégrité intellectuelle.

Dans certains cours, on a choisi de ramener les évaluations en classe, mais cette solution n'est évidemment pas possible dans tous les contextes, en plus de requérir des aménagements importants, tant sur le plan pédagogique que matériel. Une autre approche consiste à exiger une transparence absolue de la part des étudiants quant au recours à l'IA générative dans le cadre de travaux, en leur demandant par exemple de citer les passages générés par les modèles (McAdoo, 2023) ou encore en indiquant de manière très précise toutes les interactions avec des outils d'IA générative qui ont eu lieu dans le cadre de la réalisation d'un travail académique. Évidemment, les enseignants de tous les niveaux d'enseignement sont aux prises avec ces difficultés et doivent ajouter à leur emploi du temps déjà surchargé la transformation de leurs évaluations et de leurs stratégies pédagogiques de manière à assurer une évaluation juste et efficace des compétences des étudiants dans ce nouveau contexte. Des journées d'étude et de réflexion se tiennent d'ailleurs dans les diverses régions du Québec et les différents établissements d'enseignement dans le but d'adopter des pratiques permettant de surmonter ces difficultés.

Compte tenu du fait que dans certains contextes, il peut être souhaitable de permettre et même d'encourager le recours à l'IA générative dans le cadre de travaux étudiants, mais que dans d'autres contextes, il convient plutôt de le proscrire ou de l'empêcher, il apparaît évident que les solutions ne pourront être unilatérales et devront être adaptées aux différentes disciplines, types d'évaluation et niveaux d'enseignement.

Il faut aussi noter que bon nombre d'établissements d'enseignement n'ont toujours pas pris position sur la question. Dans le contexte de la préparation d'une table ronde réunissant plus de 40 ministres de l'éducation de partout à travers le monde, l'UNESCO a publié une étude qui démontre que parmi les 450 écoles et universités sondées, moins de 10 % avaient adopté des politiques institutionnelles ou des orientations formelles concernant l'utilisation d'applications d'IA générative (2023). Face à cette situation inédite, aux problèmes qu'elle entraîne et à la nécessité d'adapter les pratiques à cette nouvelle réalité, il apparaît également essentiel de former tous les membres de la communauté académique – étudiants, professeurs, professionnels, cadres et administrateurs – à l'utilisation responsable de l'IA générative en éducation.

Des risques significatifs : fracture, désinformation, biais, etc.

Si la question de l'intégrité académique et du plagiat tend à monopoliser les débats qui portent sur les enjeux éthiques liés à l'utilisation du recours à l'IA générative dans l'enseignement, ces préoccupations ne doivent pas faire ombrage à d'autres risques tout aussi significatifs, mais souvent relégués à l'arrière-plan. D'abord, le recours à l'IA générative pour évaluer les apprentissages pourrait, au lieu de réduire les inégalités entre les étudiants, les exacerber. En effet, compte tenu de la performance des outils d'IA générative, ceux qui en ont un meilleur accès et une meilleure maîtrise pourraient être d'autant plus avantagés alors que les étudiants moins privilégiés ou qui ont une littératie numérique moins développée pourraient ne pas être en mesure de tirer profit de ces avantages, accentuant ainsi la fracture numérique existante (Commission de l'éthique en science et en technologie, 2023).

De plus, malgré leurs capacités impressionnantes, les modèles de traitement du langage naturel ont une fâcheuse tendance à générer du contenu erroné, trompeur ou qui n'est pas issu de leur données d'entraînement, ce que les spécialistes qualifient parfois d'« hallucinations ». Ce problème, qui ne semble pas sur le point d'être résolu, est particulièrement préoccupant lorsque l'on utilise l'A générative dans un contexte éducatif, où la véracité et l'exactitude des contenus sont cruciales. La situation est d'autant plus inquiétante vu la vulnérabilité des personnes, bien souvent des mineurs, qui sont en situation d'apprentissage et la confiance avec laquelle les modèles de langage peuvent générer du contenu erroné ou inventé de toute pièce, accentuant ainsi le phénomène connu sous le nom d'effet ELIZA[5] – qui consiste à prêter des traits humains et amicaux à des systèmes d'IA– et, dans le cas qui nous occupe, à leur attribuer une autorité sur un sujet donné.

Or, faire croire que plus les informations sont répétées par les SIA, plus elles ont de chances d'être vraies « constitue une heuristique habile mais dangereuse et, disons-le, inadaptée pour établir une valeur de vérité.» (Poibeau, 2023). Il apparaît donc essentiel de se pencher sur le risque de désinformation et sur les impacts spécifiques de ce risque en éducation. En fait, ce problème pourrait possiblement constituer un des obstacles les plus importants à l'utilisation éthique et responsable de l'IA générative en contexte éducatif.

« ...Les étudiants moins privilégiés ou qui ont une littératie numérique moins développée pourraient ne pas être en mesure de tirer profit de ces avantages, accentuant ainsi la fracture numérique existante (Commission de l'éthique en science et en technologie, 2023). »

Les modèles de traitement du langage naturel ont une fâcheuse tendance à générer du contenu erroné.

Ce problème est particulièrement préoccupant dans un contexte éducatif, où la véracité et l'exactitude des contenus sont cruciales considérant la vulnérabilité des personnes, bien souvent des mineurs, qui sont en situation d'apprentissage.

5 L'effet ELIZA tire son nom du programme informatique homonyme, l'un des premiers agents conversationnels (*chatbot*), développé par Joseph Weizenbaum (1966). ELIZA simulait notamment les réponses d'un psychothérapeute, l'effet ELIZA est donc la tendance à anthropomorphiser un SIA, voire à penser qu'un SIA est humain.

La production de contenu biaisé par les modèles d'IA générative constitue également un enjeu majeur pour le réseau de l'éducation. Bien que les entreprises qui conçoivent ces modèles soient conscientes de ces risques et consacrent beaucoup de ressources à l'élimination de contenu sexiste, haineux ou violent, il reste que les résultats générés par ces systèmes reflètent les biais qui sont présents dans leurs données d'entraînement (Buolamwini et Gebru, 2018). Malgré les efforts déployés, ces problèmes, bien connus dans certains domaines de l'IA, comme les systèmes de recommandation, se manifestent également en IA générative (Nicoletti et Bass, 2023). La reproduction des stéréotypes est particulièrement alarmante en contexte éducatif, où l'objectif est au contraire de promouvoir la pensée critique, la diversité et l'équité et de lutter contre les préjugés. Par ailleurs, la diversité des contenus est fondamentale pour assurer une éducation riche et nuancée. Le fait que les modèles d'IA générative tendent à générer du contenu similaire à leurs données d'entraînement comporte le risque de réduire la diversité des idées qui sont présentées aux étudiants, par exemple en privilégiant des perspectives dominantes au détriment de voix plus marginales. Enfin, il convient de rappeler que les entreprises qui produisent les modèles d'IA générative et celles qui mettent en marché des produits dérivés qui sont destinés au milieu de l'éducation ont avant tout une motivation financière, et non pédagogique. Ainsi, l'introduction de ces outils en enseignement risque d'être assujettie à ces objectifs, et non en premier lieu aux intérêts des étudiants ou de leurs enseignants. Il importe par conséquent d'éviter de céder à un enthousiasme aveugle face aux promesses qu'offrent ces entreprises et de faire preuve de prudence et de vigilance dans l'intégration de ces outils aux pratiques éducatives, considérant le fait que l'adoption massive d'outils d'IA générative en enseignement est susceptible d'accroître la dépendance du réseau de l'éducation aux géants du numérique.

« Le fait que les modèles d'IA générative tendent à générer du contenu similaire à leurs données d'entraînement comporte le risque de réduire la diversité des idées qui sont présentées aux étudiants, par exemple en privilégiant des perspectives dominantes au détriment de voix plus marginales. »

« ...l'arrivée de l'IA générative ne peut être considérée comme un événement isolé ou idiosyncratique»

2.4 La commercialisation de l'IA générative

Si le déploiement de l'IA générative dans nos sociétés soulève de nombreux enjeux liés à son utilisation, comme nous venons de les aborder, il fait aussi poindre des enjeux non négligeables relativement à la façon dont celle-ci est commercialisée. Or, ces enjeux sont souvent relégués à l'arrière-plan dans les débats entourant l'IA générative. Compte tenu de leurs implications pour le développement actuel et futur de cette technologie disruptive, nous considérions toutefois important de les ramener ici au centre de la réflexion.

Afin de mettre en lumière ces enjeux de commercialisation, une lecture du contexte économique et politique dans lequel s'inscrit l'arrivée des modèles d'IA générative dans nos sociétés est nécessaire. Une telle mise en contexte nous permet de voir que l'arrivée de l'IA générative ne peut être considérée comme un événement isolé ou idiosyncratique et que les entreprises qui produisent les modèles d'IA générative s'inscrivent dans des contextes sociaux, économiques, politiques et technologiques complexes et hautement imbriqués, sur le fond desquels ils mènent ce que l'on appelle maintenant une « course mondiale à l'IA ». Cette course, dont on commence aujourd'hui seulement à mesurer l'ampleur, a débuté il y a déjà plusieurs années, et se déroule à l'intérieur du périmètre d'un marché mondial hautement compétitif et axé sur les profits, où la concentration des ressources et des pouvoirs réduit à quelques joueurs seulement les compétiteurs ayant une réelle chance de franchir le fil d'arrivée - pensons ici notamment à Microsoft, Google (Alphabet), Meta, Baidu, IBM ou Alibaba.

Ainsi, malgré le fait que les technologies d'IA projettent l'impression d'être issues de laboratoires et de centres de recherche où œuvrent des chercheuses et chercheurs dédiés à l'avancement de l'humanité, ou encore de petites entreprises « startup » animées par de nobles missions, la réalité s'avère souvent bien différente. En effet, les importants capitaux nécessaires pour développer l'IA générative et permettre aux petits joueurs d'atteindre leur plein potentiel font en sorte que ceux-ci sont souvent avalés par les plus gros joueurs dès que leur technologie montre des signes de potentiel de marché, de sorte qu'à ce jour, la propriété intellectuelle des innovations en IA est principalement concentrée dans les mains de grandes entreprises[6].

Le choix de cette trajectoire signifie souvent, pour les chercheuses et chercheurs et les petites entreprises d'IA générative, qu'ils devront sacrifier certains de leurs objectifs et intentions initiaux au profit des intérêts des entreprises qui les acquièrent. En d'autres termes, les entreprises d'IA générative se révèlent, comme c'est le cas dans tous les autres secteurs d'activités économiques, soumises à d'importants intérêts économiques, commerciaux, voire politiques, intérêts qui exercent un pouvoir significatif sur les orientations que prennent le développement et la commercialisation de l'IA.

Du même coup, cela invite à mettre en perspective les discours d'IA « responsable », maintenant généralisés chez les développeurs d'IA générative. Dans ce qui suit, nous verrons comment l'exemple de la commercialisation de ChatGPT nous amène notamment à questionner les prétentions d'OpenAI à poursuivre l'intérêt général et à aligner ses modèles d'IA générative avec les valeurs que nos sociétés jugent importantes.

Les entreprises d'IA générative se révèlent soumises à d'importants intérêts économiques, commerciaux, voire politiques, intérêts qui exercent un pouvoir significatif sur les orientations que prennent le développement et la commercialisation de l'IA.

Une IA générative dans l'intérêt de qui?

La récente saga autour du congédiement, puis du retour de Sam Altman à la tête d'OpenAI (Metz et al., 2023) nous a permis de voir avec une acuité particulière ce pouvoir important que les intérêts économiques et commerciaux peuvent exercer sur le développement et les façons de commercialiser l'IA générative. En effet, ChatGPT, la première IA générative à avoir été lancée dans nos sociétés, a été développée par une jeune entreprise, Open AI, qui était, à ses débuts en 2015, structurée en OBNL. Celle-ci se disait animée par une mission sociale forte, celle de produire une IA qui serait bénéfique pour toute l'humanité :

> *OpenAI's mission is to ensure that artificial general intelligence (AGI)—by which we mean highly autonomous systems that outperform humans at most economically valuable work—benefits all of humanity. (OpenAI, 2018).*

À peine trois années plus tard, soit en 2019, OpenAI a créé une filiale à « but lucratif plafonné » sur laquelle elle détient le contrôle en vue d'attirer des capitaux lui permettant de poursuivre le développement de ses technologies en vue de développer une IA générale (OpenAI, 2023d). C'est ce qui lui a permis, dans la foulée de la sortie de ChatGPT en fin 2022, de recevoir des investissements de l'ordre de 13 milliards US$ du géant Microsoft (Chafkin et Metz, 2023), soit bien plus que tout ce qu'elle avait réussi à lever jusque-là (OpenAI, 2023d).

6 Selon McKenna (2023, 6 avril), au Canada, la plupart des innovations en IA sont faites par des chercheuses et chercheurs canadiens, mais la plupart de ces brevets sont détenus par des entreprises étrangères, principalement Microsoft et IBM – https://www.ledevoir.com/economie/788117/techno-l-ia-canadienne-limitee-dans-son-developpement

Or, dans le contexte de cette association avec un acteur aussi puissant que Microsoft, qui détiendrait pas moins de 49% de sa filiale (Chafkin et Metz, 2023), on peut toutefois douter de la capacité effective d'OpenAI à poursuivre sa mission orientée vers la poursuite de l'intérêt général. En effet, malgré les discours de son équipe, et plus particulièrement de son PDG, Sam Altman, qui ont cherché à nous rassurer tout au long de la dernière année sur le maintien de cette mission initiale, on constate que les objectifs des deux entreprises divergent de façon substantielle, ce qui risque nécessairement d'influencer les pratiques d'OpenAI.

Cet écart ressort d'ailleurs avec une acuité particulière à travers certaines actions récentes de Microsoft, comme le démantèlement de son équipe « éthique et société », dont la fonction était d'anticiper les impacts potentiels des technologies développées, face à la pression de la direction pour « sortir » ses applications d'IA le plus rapidement possible sur le marché, et ainsi prendre de l'avance dans la course à l'IA (Schiffer et Newton, 2023). On le note aussi à travers son approche résolument axée sur le marketing et la vente de son nouveau programme « Copilot », un assistant virtuel mobilisant l'IA générative qui est depuis le 1er novembre intégré à sa suite Office 365. Or, ce recours habile à la métaphore de l'outil (passif) et de l'utilisateur (actif) dans la conception d'un « copilote » laisse entendre que l'IA générative n'est qu'un assistant personnel, et que le professionnel demeurera toujours aux commandes pour superviser l'assistant et engager sa propre responsabilité, vise à nous réconforter face à la profonde transformation du travail qui vient d'être décrétée par Microsoft, en voilant la substitution cognitive du travail qu'elle engendrera dans les faits (Anctil, 2023b). Un tel discours marketing tait cependant l'incidence d'un tel programme sur la société dans son ensemble, et ne fait état d'aucune étude sur les risques prévisibles et imprévisibles d'en étendre l'utilisation à tous les secteurs sensibles des organisations et professions utilisant ses logiciels.

L'influence de Microsoft et de son approche résolument axée sur la performance et l'anticipation des besoins de ses clients, ainsi que son objectif d'aller plus vite que ses concurrents, notamment en étant toujours le premier à mettre sur le marché les produits d'IA les plus avancés sur OpenAI et plus particulièrement sur Sam Altman, se fait d'ailleurs progressivement sentir depuis la dernière année. La proximité entre ce dernier et Microsoft, n'est sans doute pas étrangère au renvoi momentané de Altman au poste de PDG d'OpenAI. Si celui-ci a été officiellement justifié par le conseil d'administration par un manque de transparence de sa part, la thèse d'une tension plus fondamentale entre la volonté d'arriver à l'IAG le plus rapidement possible, incarnée par Altman, et le souhait des membres du CA d'avancer plus prudemment, de façon à limiter les risques et de se recentrer sur sa mission initiale visant à produire une IA générale sécuritaire et bénéfique pour tous (OpenAI, 2023e) est avancée par plusieurs observateurs.

« Ce recours habile à la métaphore de l'outil (passif) et de l'utilisateur (actif) dans la conception d'un « copilote » laisse entendre que l'IA générative n'est qu'un assistant personnel, et que le professionnel demeurera toujours aux commandes pour superviser l'assistant et engager sa propre responsabilité, vise à nous réconforter face à la profonde transformation du travail… »

Dans ses récentes entrevues, Satya Nadella, le PDG de Microsoft, exprimait pour sa part clairement sa désapprobation du renvoi de Sam Altman et considérait qu'il aurait dû être consulté sur cette décision, compte tenu de sa participation financière significative dans OpenAI (Duhigg, 2023). Ces événements ont d'ailleurs entraîné une intervention rapide de Microsoft auprès de son partenaire stratégique et suscité un rebrassage important de la gouvernance de OpenAI, incluant le départ des personnes qui avaient initié le renvoi d'Altman. À la suite de cette restructuration, le géant de la technologie ne détient toujours pas de siège votant sur le C.A. de OpenAI (OpenAI, 2023d; Field, 2023), mais il y gagne néanmoins un siège d'observateur sans droit de vote, faisant en sorte qu'il aura à partir de maintenant une bien meilleure vision du fonctionnement interne et des décisions de son partenaire (Heath, 2023), constituant un autre parmi les nombreux signes du pouvoir croissant qu'exerce Microsoft sur OpenAI.

Comme on peut le voir, les pratiques d'OpenAI ainsi que la façon pour le moins cavalière dont elle a lancé ChatGPT dans nos sociétés et s'est ensuite associée avec un des géants des GAFAM mettent en lumière plusieurs enjeux importants qui doivent eux aussi être abordés dans le débat autour de l'IA génératives et ses impacts sociétaux, soit ceux de l'influence des modèles d'affaires des entreprises qui les développent et de la structure de l'industrie de l'IA sur l'orientation de ces développements et de leur commercialisation. Et ce que l'on observe à cet égard, c'est que la structure de l'industrie de l'IA, par son orientation hautement compétitive et par la concentration du pouvoir et des intérêts qui la caractérisent (Verdegem, 2021), ne diffère pas fondamentalement des autres secteurs de l'économie.

« les pratiques d'OpenAI ainsi que la façon pour le moins cavalière dont elle a lancé ChatGPT dans nos sociétés et s'est ensuite associée avec un des géants des GAFAM mettent en lumière plusieurs enjeux importants qui doivent eux aussi être abordés dans le débat autour de l'IA génératives et ses impacts sociétaux »

Dans un tel contexte, on peut s'attendre à ce que la priorisation de l'intérêt général dans le déploiement de l'IA générative soit sérieusement compromise, surtout lorsque celui-ci ne convergera pas avec les intérêts commerciaux des entreprises qui les produisent, et ce, malgré les déclarations d'intentions à cet effet.

Quel alignement avec quelles valeurs?

Depuis l'arrivée de l'IA générative dans nos sociétés, on parle de plus en plus du « problème de l'alignement » de ces modèles d'IA avec les objectifs et valeurs importantes pour nos sociétés. Si la notion d'alignement s'avère une formule relativement nouvelle, l'idée qu'elle traduit, elle, ne l'est certainement pas. En effet, au cours des dix dernières années, on a vu émerger un grand nombre de déclarations de valeurs et chartes éthiques proposées par différents pays et organismes multipartites (Jobin et al. 2019; Tidjon et Khomh, 2022) qui ont justement pour objectif d'énoncer quelles sont ces valeurs auxquelles nous tenons et sur lesquelles nous souhaitons voir l'IA « s'aligner » pour être considéré « responsable ». Malgré quelques angles morts et une effectivité somme toute limitée de par leur nature volontaire et non-contraignante (Jobin et al., 2019; Mittelstadt, 2019; Greene et al., 2019; Buruk et Al., 2020; Adams, 2021; Dignum, 2022), ces déclarations et chartes ont le mérite d'expliciter les valeurs et principes importants non seulement selon le point de vue des entreprises qui produisent de l'IA, mais aussi pour les autres acteurs sociaux qui seront impactés d'une façon ou d'une autre par le déploiement de l'IA dans nos sociétés. Certaines de ces chartes et déclarations ont d'ailleurs été produites sur la base de consultations de la population, comme c'est notamment le cas de la Déclaration de Montréal pour une développement responsable de l'IA (2018). Elles devraient donc constituer un guide incontournable pour les développeurs d'IA qui souhaitent s'attaquer de façon sérieuse au « problème de l'alignement ».

« Mais OpenAI nous demande plutôt de faire confiance à ses nobles intentions et promesses, et ce, sans fournir les informations nécessaires pour nous permettre d'évaluer si les mesures mises en place correspondent effectivement aux valeurs et attentes des parties prenantes [...] et si nous les considérons satisfaisantes. »

Si l'on prend encore ici l'exemple d'OpenAI, cette entreprise et son PDG nous expliquent, depuis maintenant une année, l'importance qu'ils accordent au problème de l'alignement entre les performances de ChatGPT et les objectifs humains. Ils demeurent en contrepartie très évasifs sur le comment ils s'y prennent pour résoudre ce problème. Quels valeurs et objectifs considèrent-ils importants? Comment sont-ils déterminés et comment OpenAI s'assure qu'ils sont respectés de façon satisfaisante? À cet égard, ce développeur d'IA aurait pu s'appuyer sur les chartes et déclarations déjà existantes et mettre en place une approche plus ouverte impliquant les parties prenantes et la société civile pour évaluer de façon structurée et rigoureuse quelles mesures devraient être mises en place et si celles-ci sont satisfaisantes ou non. Mais OpenAI nous demande plutôt de faire confiance à ses nobles intentions et promesses, et ce, sans fournir les informations nécessaires pour nous permettre d'évaluer si les mesures mises en place correspondent effectivement aux valeurs et attentes des parties prenantes qui seront touchées par l'IA générative et si nous les considérons satisfaisantes.

« il semble que l'ouverture et la transparence, pourtant au cœur de la vision initiale d'OpenAI, ont perdu beaucoup de leur poids effectif dans les décisions de l'entreprise qui a créé ChatGPT. »

Cette façon de faire est inquiétante d'abord parce qu'elle ne s'avère pas une garantie suffisante pour nous assurer que les modèles d'IA génératives soient adéquatement alignés avec nos valeurs et objectifs. Mais elle est aussi inquiétante parce que, comme plusieurs autres pratiques de l'entreprise, elle n'est elle-même pas alignée avec les valeurs énoncées dans les chartes éthiques en matière d'IA. Elle néglige notamment la valeur de **participation démocratique,** qui est au cœur de la Déclaration de Montréal (2018) et qui stipule que les systèmes d'IA « doivent pouvoir être soumis à un examen, un débat et un contrôle démocratiques ».

Les pratiques d'OpenAI sont aussi problématiques en regard de la valeur de **transparence**, principe si fondamental qu'il est celui qui revient le plus souvent dans les déclarations et chartes éthiques en matière d'IA (Jobin et al., 2019). Si on vient de voir que la problématique de l'alignement des modèles d'IA générative est gérée de façon trop opaque, il semble que l'ouverture et la transparence, pourtant au cœur de la vision initiale d'OpenAI, ont perdu beaucoup de leur poids effectif dans les décisions de l'entreprise qui a créé ChatGPT. On peut ici prendre comme autre exemple le recours au « *open sourcing* », soit l'utilisation et le partage de code informatique libre de droit, qui s'avérait particulièrement intéressant pour tenter de minimiser les conséquences néfastes potentielles de l'IA générative :

> *even experts do not understand when or how AI might become powerful enough to cause harm, damage or injury. Open sourcing of code allows many people to think through the consequences both individually and together. Ideally, that effort will advance software that is increasingly powerful and useful, but also broadly understandable in its mechanisms and their implications.* (Shafto, 2016).

Or, l'arrivée de GPT4, la nouvelle version de cette IA générative, a aussi signé la fin du *opensourcing* : « c'était une erreur », affirme maintenant Altman (cité dans Vincent, 2023).

On peut ajouter à cela que c'est pour manque de transparence envers son conseil d'administration que Sam Altman a été congédié de son poste de PDG le 17 novembre dernier (OpenAI, 2023). En fait, plusieurs observateurs, dont Elon Musk, l'un des fondateurs de l'entreprise, considèrent qu'il n'y a aujourd'hui « plus rien d'ouvert dans OpenAI », contrairement à ce que son nom laisse prétendre (Musk, 2023). Une telle affirmation est par ailleurs appuyée par l'étude de Bommasani et ses collègues (2023), qui ont attribué un score de seulement 47% à OpenAI pour la transparence de l'ensemble de ses pratiques. Les auteurs notent qu'il s'agit là d'une lacune importante chez tous les principaux joueurs de l'IA générative tels Google, Anthropic et Amazon, et dans l'ensemble de l'écosystème, où on observe une baisse générale de transparence. Ils rappellent que des pratiques transparentes sont pourtant essentielles pour un développement d'une IA générative qui ne soit pas dommageable :

Transparency is a broadly-necessary condition for other more substantive societal progress, and without improvement opaque foundation models are likely to contribute to harm. Foundation models are being developed, deployed, and adopted at a frenetic pace : for this technology to advance the public interest, real change must be made to rectify the fundamental lack of transparency in the ecosystem. (Bommasani et al., 2023, p.63).

Quant aux principes de **prudence** et de **sécurité**, qui figurent aussi parmi les valeurs les plus fréquemment énoncées dans les chartes éthiques en IA, on peut encore ici questionner l'alignement des pratiques d'OpenAI avec ceux-ci. À commencer par le lancement de ChatGPT et de ses versions plus récentes, que plusieurs observateurs, comme l'ancienne conseillère d'OpenAI Gillian Hadfield, considèrent avoir été fait de façon trop précipitée, sans en avoir mesuré adéquatement les conséquences potentielles avant de les rendre accessibles au grand public (2023). Si nous avons vu plus tôt que c'est le cas de façon particulièrement criante en ce qui concerne la propriété intellectuelle et la diffusion de fausses informations, au-delà de ces impacts somme toute relativement visibles, l'impact plus profond et possiblement non-réversible sur le travail et sur l'apprentissage a encore moins été anticipé. Aux yeux de nombre d'observatrices et d'observateurs, nous sommes en fait ici les cobayes d'une technologie qui est en train de modifier l'humanité, mais dont les conséquences ont été bien peu évaluées. Or, c'est précisément le contraire qu'appelle la Déclaration de Montréal, lorsqu'elle énonce que les personnes impliquées dans le développement de l'IA doivent agir avec prudence, c'est-à-dire en « anticipant autant que possible les conséquences néfastes de l'utilisation des SIA et en prenant des mesures appropriées pour les éviter » (Déclaration de Montréal, 2018).

« Aux yeux de nombre d'observatrices et d'observateurs, nous sommes en fait ici les cobayes d'une technologie qui est en train de modifier l'humanité, mais dont les conséquences ont été bien peu évaluées. »

Altman affirme pour sa part que malgré tout le travail d'anticipation réalisé par OpenAI, ils ne peuvent à eux seuls prévoir tous les impacts de leur technologie. C'est précisément pour cette raison qu'ils ont souhaité lancer ChatGPT dans le monde, c'est-à-dire pour bénéficier du feedback de milliers d'utilisateurs et le tester dans le monde réel, décuplant ainsi l'efficacité de cette expérimentation, comme il l'expliquait notamment le 17 mars dernier en entrevue : « *I think it doesn't work to do all this in the lab, you've got to get these products out into the world and make contact with reality and make our mistakes while the stakes are low* ». (Altman, 2023). Or, s'il est nécessairement ardu, voire impossible de prévoir tous les effets réels possibles d'une innovation, il n'en demeure pas moins qu'un niveau suffisant d'anticipation est nécessaire afin d'assurer que les SIA soient « robustes, sûrs et sécurisés tout au long de leur cycle de vie, de sorte que, dans des conditions d'utilisation normales ou prévisibles, ou en cas d'utilisation abusive ou de conditions défavorables, ils soient à même de fonctionner convenablement, et ne fassent pas peser un risque de sécurité démesuré », comme l'exige le principe de sécurité de l'OCDE (2023). Lorsque prise au sérieux, la valeur de sécurité risque inévitablement de ralentir la vitesse de l'innovation. Pourtant, bien que la sécurité soit au cœur des discours d'OpenAI et de son PDG (OpenAI, 2018), ce n'est dans les faits pas la voie qu'a choisie ce pionnier de l'IA générative, contrairement à son concurrent Anthropic, qui avait plutôt opté pour éviter d'accélérer la course mondiale à l'IA dans un contexte où les risques, en termes de sécurité notamment, étaient encore trop incertains. Car malgré ce que promeut la populaire devise de la Silicon Valley, « *AI Doesn't Need to Move Fast and Break Thing*s. » (Collins, 2023).

Le **bien-être** (Déclaration de Montréal, 2018) est une autre valeur dont on peut se demander si elle se traduit dans les pratiques commerciales d'OpenAI. En effet, bien qu'elle soit au cœur de la mission de ce pionnier de l'IA générative, qui dit vouloir faire profiter à tous de ses bienfaits, il n'en demeure pas moins que le bien-être des personnes qui risquent de perdre leur emploi ou qui voient leur conditions de travail se dégrader, ou encore des « travailleurs du clic » embauchés par OpenAI au Kenya (Perrigo, 2023) pour étiqueter, répertorier et identifier les données et les images, travaillant pour des salaires infimes et s'exposant à d'importants risques de traumatisme, est loin d'être assuré avec l'arrivée de chatGPT.

On pourrait ainsi continuer à examiner les pratiques d'OpenAI et des autres développeurs d'IA générative à l'aune des déclarations et chartes éthiques en matière d'IA, et on verrait que nombreuses sont les valeurs que nous considérons importantes à avoir été négligées par les façons dont cette technologie révolutionnaire est en train d'être commercialisée.

Cela nous fait ainsi prendre conscience que « l'alignement » n'interpelle pas que les caractéristiques intrinsèques au SIA lui-même, mais aussi ses méthodes de commercialisation, et que par ailleurs la détermination des valeurs et objectifs devant guider cet alignement ne concerne pas seulement les entreprises, mais aussi (et même surtout) les parties prenantes qui seront impactées de près ou de loin par l'IA générative, c'est-à-dire toute la société.

Les risques des discours de responsabilité des entreprises d'IA générative

Les entreprises qui développent l'IA ont bien compris qu'elles ne pourraient commercialiser leurs produits, notamment leurs modèles d'IA générative, si elles ne s'engageaient pas à le faire de façon responsable. Elles ont ainsi dès le départ adopté sans hésitation des discours sur l'IA responsable, comme on l'a bien vu avec OpenAI notamment. S'il s'agit là d'un changement majeur comparativement aux entreprises technologiques qui, il y a à peine vingt ans encore, résistaient à l'idée d'assumer des « responsabilités sociales » élargies - pensons ici entre autres aux domaines de la bioingénierie (OGM) ou des nanotechnologies -, les nombreuses recherches menées sur le sujet au cours des deux dernières décennies nous ont appris qu'il faut néanmoins demeurer vigilants face à ces discours. En effet, ils recèlent un piège, que nous pourrions appeler une « illusion de vertu », découlant de l'utilisation croissante des notions d'éthique, de responsabilité sociale et de bien commun par les entreprises, car ils tendent à créer une aura de moralité autour de celles-ci, leur conférant un capital de sympathie, voire une légitimité plus grande auprès du public et de nos décideurs. Lorsque cela se produit, on constate une plus grande confiance et moins de vigilance envers ces entreprises, ce qui leur permet en retour d'agir avec plus de marge de manœuvre quant aux produits qu'ils commercialisent ou aux pratiques qu'ils adoptent, mais aussi face aux régulateurs. Ainsi, ces discours d'éthique et de responsabilité ont le potentiel de conférer un pouvoir supplémentaire aux entreprises qui les mobilisent, ce qui s'avère particulièrement problématique lorsque ces discours ne s'accompagnent pas de pratiques conséquentes (Marchildon, 2016; 2017).

Ces risques d'usage essentiellement performatif des discours de responsabilité sociale ne sont pas spécifiques aux entreprises d'IA, mais ils découlent plus largement du modèle d'affaires et du système économique dans lequel elles évoluent et qui en orientent fortement la trajectoire. À la lumière de ces observations, il n'est donc pas surprenant de constater ces mêmes risques dans le domaine de l'IA générative, où les discours d'IA responsable sont monnaie courante. En effet, Sam Altman n'hésite pas à multiplier les tribunes sur lesquelles il présente OpenAI comme un développeur d'IA générative particulièrement responsable, entre autres parce qu'ils ont mis en place des mesures et garde-fous pour minimiser les usages problématiques de ChatGPT et en assurer les ajustements en cas de problème détecté à l'interne ou mis en lumière par ses utilisateurs. Il souligne dans la foulée que leurs compétiteurs n'en font pas nécessairement autant et n'auront pas tous le même souci de l'intérêt général qu'OpenAI. Comme il le disait le 17 mars dernier : « *A thing that I do worry about is we're not gonna be the only creator of this technology. There will be other people who won't put some of the safety limits that we put on it.* » (Altman, 2023). Interrogé par le congrès américain, il a par ailleurs invité les régulateurs à utiliser cette occasion pour réfléchir aux impacts de l'arrivée des IA génératives dans nos sociétés et la réguler de façon appropriée, mais en ne contraignant pas l'innovation.

« ...[ces discours] recèlent un piège, que nous pourrions appeler une *illusion de vertu*, découlant de l'utilisation croissante des notions d'éthique, de responsabilité sociale et de bien commun par les entreprises, car ils tendent à créer une aura de moralité autour de celles-ci... »

Or, à la lumière de ce qui vient d'être dit concernant les risques associés aux discours de responsabilité sociale, tant le public que les décideurs doivent demeurer prudents devant une telle rhétorique, et toujours questionner dans quelle mesure les discours et les promesses se traduisent réellement dans les pratiques. Par exemple, les mesures et garde-fous mis en place par OpenAI sont-ils suffisants ou peuvent-ils aisément être contournés? Ces mesures sont-elles effectivement plus rigoureuses que celles de leurs compétiteurs? L'entreprise nous fournit-elle, en toute transparence, les détails quant à ces mécanismes et garde-fous afin de nous permettre d'évaluer s'ils correspondent aux valeurs et attentes des parties prenantes qui seront touchées par l'IA générative ainsi que pour vérifier si elles sont adéquatement mises en pratique? Car au-delà des pratiques elles-mêmes, la possibilité, pour les acteurs externes, de participer aux orientations et à l'évaluation de ces pratiques est essentielle à l'exercice d'une responsabilité envers la société.

Or, à la lumière des risques associés aux discours de responsabilité sociale, tant le public que les décideurs doivent demeurer prudents devant une telle rhétorique, et toujours questionner dans quelle mesure les discours et les promesses se traduisent réellement dans les pratiques.

Bref, les discours et les initiatives concrètes des entreprises d'IA doivent être questionnées et scrutées, et elles doivent être critiquées si elles ne s'avèrent à la hauteur de ce que nous considérons responsable. Et pour les bonifier, on ne compte plus les outils et de critères qui sont proposés par des chercheuses et chercheurs et des organismes étatiques ou de la société civile pour favoriser l'innovation responsable et durable en général et l'IA en particulier, dont près de 700 sont recensés dans le catalogue de l'OCDE[7]. Pour n'en nommer que quelques-uns, pensons notamment au *Code de conduite volontaire visant un développement et une gestion responsables des systèmes d'IA générative avancés* tout récemment développé par Innovation, Sciences et développement économique Canada (ISDE, 2023), le cadre du *Partnership on AI* pour le déploiement sécuritaire des modèles de fondation (PAI, 2023), le cadre d'innovation responsable de Stilgoe et al. (2013), ou encore la mesure de l'impact de l'innovation sur les 17 objectifs de développement durable de l'ONU suggérée par Vinuesa et al. (2020).

« Les discours et les initiatives concrètes des entreprises d'IA doivent être questionnées et scrutées, et elles doivent être critiquées si elles ne s'avèrent à la hauteur de ce que nous considérons responsable. »

Considérant l'ampleur des conséquences et les effets potentiellement systémiques des IA génératives d'une part, et les problèmes soulevés par les pratiques de commercialisation de l'IA générative d'autre part, le recours à de tels outils reconnus par la communauté internationales et à une évaluation des pratiques par des acteurs externes et par les parties prenantes s'avèrent par conséquent nécessaires pour que le déploiement de l'IA générative puisse être considéré responsable. Cette responsabilité des entreprises, avance Vogel (2006), va par ailleurs jusqu'à appuyer les initiatives législatives visant à mieux encadrer leur secteur d'activité, et ce même si cela peut nuire à leurs profits ou ralentir leur croissance et l'innovation, tout le contraire du lobbying qu'on a notamment pu observer lors des négociations finales de l'AI Act européen (Axiotes, 2023).

7 Voir le *Catalogue of Tools et Metrics for Trustworthy AI* de l'OCDE: https://oecd.ai/fr/catalogue/tools.

Les questions et enjeux que nous venons de soulever rappellent donc qu'il importe de demeurer particulièrement vigilants et de ne pas nous limiter aux discours et déclarations des entreprises qui produisent les systèmes d'IA générative pour évaluer le bien-fondé et l'impact de leurs pratiques, aussi séduisants et rassurants soient-ils. La recherche et l'expérience montrent par ailleurs que l'autorégulation n'est généralement pas suffisante pour assurer des pratiques responsables, et ce, même si elle provient d'une petite entreprise avec une « mission sociale » (et non d'un géant du GAFAM) qui affirme avoir le potentiel de trouver une alternative plus intéressante au capitalisme, voire de le « briser » (cité dans Konrad et Cai, 2023). Bref, en matière d'IA générative, les balises et contrôles sociétaux demeurent nécessaires afin d'assurer que les SIA qui seront introduits dans nos sociétés dans les mois et années à venir respecteront non seulement certaines exigences minimales, notamment en matière de sécurité, de vie privée, de transparence et de responsabilité, mais aussi qu'ils seront alignés avec les valeurs que nos sociétés jugent importantes, et ce, dans l'intérêt et le respect des êtres vivants. Par conséquent, la mise en place rapide d'initiatives régulatoires exigeantes est essentielle, qu'il s'agisse de balises juridiques, comme nous le verrons dans la section 3, ou encore de mesures régulatoires non-étatiques, telles les certifications ou autres formes de contrôle externes aux entreprises, comme nous le suggérerons dans les pistes d'action et de réflexion à la section 4.

2.5 En conclusion sur les enjeux de l'IA générative

Des discussions précédentes autour des enjeux de l'utilisation et de la commercialisation de l'IA générative, il ressort deux constats qui nous apparaissent fondamentaux. D'abord, qu'il s'agisse du monde du travail ou de l'éducation, nous avons pu voir qu'un des principaux effets de l'IA générative se révèle être un bouleversement profond de notre rapport à la connaissance et à l'expertise. Qu'il s'agisse des compétences que nous acquérons comme apprenants ou que nous mettons à profit comme travailleurs, ou encore de nos compétences comme enseignants ou gestionnaires, l'IA générative vient chambouler nos schémas traditionnels en remettant en question l'expertise humaine.

Celles et ceux ayant acquis une expertise peuvent se sentir dépassés devant les prouesses de ce que peut accomplir une IA générative, qui s'impose de plus en plus comme source d'information et de connaissances et qui modifie profondément ce que signifie savoir et apprendre, mais sans que cette IA ne soit toutefois en mesure d'assumer les responsabilités qui viennent avec ce pouvoir.

Ensuite, tant dans les sphères d'activités spécifiques que nous avons discutées qu'en ce qui a trait à la façon dont est commercialisée l'IA générative par les entreprises qui la produisent, on observe que les enjeux de pouvoir sont omniprésents. De par les importantes ressources financières requises pour développer l'IA générative, ces modèles sont surtout commercialisés par de grandes entreprises, qui ont des capacités - puissance de calcul, talents et données - et sont en mesure d'exercer un plus grand pouvoir tant sur les citoyens et la société civile, que sur les entreprises plus traditionnelles et nos institutions, notamment celles du milieu de l'éducation. Ces rapports de pouvoir étaient certes déjà bien présents - et souvent dénoncés d'ailleurs - dans nos sociétés gouvernées par l'économie de marché. Mais le survol des enjeux que nous avons présenté plus tôt met en lumière le fait que l'arrivée de l'IA générative vient renforcer et accentuer ces rapports de pouvoir déjà hautement asymétriques et contribue, ce faisant, à mettre à risque la démocratie.

Dans la section qui suit, nous verrons comment certains pays tentent de s'attaquer à ces enjeux de taille que soulève l'IA générative par le biais de la mise place d'initiatives législatives à cet effet.

« Qu'il s'agisse des compétences que nous acquérons comme apprenants ou que nous mettons à profit comme travailleurs, ou encore de nos compétences comme enseignants ou gestionnaires, l'IA générative vient chambouler nos schémas traditionnels en remettant en question l'expertise humaine. »

« ...le survol des enjeux que nous avons présenté plus tôt met en lumière le fait que l'arrivée de l'IA générative vient renforcer et accentuer ces rapports de pouvoir déjà hautement asymétriques et contribue, ce faisant, à mettre à risque la démocratie. »

3
Encadrement juridique de l'IA générative à ce jour

3. L'encadrement juridique de l'IA générative à ce jour

3.1 Les initiatives canadiennes

Le 16 juin 2022, le ministre de l'Innovation, des Sciences, et de l'Industrie du Canada a déposé à la Chambre des communes le projet de loi C-27. Celui-ci comporte trois parties : la première sur la vie privée des consommateurs, la deuxième sur le tribunal de la protection des renseignements personnels, et la dernière intitulée *Loi sur l'intelligence artificielle et les données* (LIAD). La version initiale de la LIAD ne portait pas spécifiquement sur l'IA générative ni sur les modèles de fondation (*Projet de loi C-27*, 2022). Cependant, elle s'accompagne d'un document complémentaire (Gouvernement du Canada, 2023a), qui évoque aussi les enjeux de responsabilité en lien avec les systèmes d'IA générative.

Ce document complémentaire (Gouvernement du Canada, 2023a) indique que :

« Certains systèmes d'IA exécutent des fonctions généralement applicables – telles que la génération de texte, audio ou vidéo – et peuvent être utilisés dans une variété de contextes différents. Étant donné que les utilisateurs finaux des systèmes à usage général ont une influence limitée sur le fonctionnement de ces systèmes, les développeurs de systèmes à usage général devraient s'assurer que les risques liés aux biais ou au contenu préjudiciable sont documentés et traités. »

« Les entreprises impliquées uniquement dans la conception ou le développement d'un système d'IA à incidence élevée mais sans capacité pratique de surveiller le système après le développement auraient des obligations différentes à cet égard de celles qui gèrent ses opérations. On ne s'attendrait pas à ce que les employés individuels soient responsables des obligations associées à l'entreprise dans son ensemble. Outre les obligations relatives à l'évaluation et à l'atténuation des risques, les entités responsables des activités réglementées associées à un système à incidence élevée seraient également tenues d'informer le ministre lorsqu'un système cause un préjudice important ou est susceptible de le faire. »

La LIAD concerne les personnes qui conçoivent, développent ou rendent disponible et gèrent l'exploitation des SIA. Son champ d'application se limite au secteur privé et ne concerne pas l'usage des SIA dans le cadre public. La loi comporte par ailleurs deux objectifs : d'une part, elle doit réguler le commerce international et interprovincial en matière d'IA; d'autre part, elle interdit certaines pratiques pouvant causer des préjudices sérieux aux individus ou à leurs intérêts ainsi qu'entraîner des risques de biais. Enfin, elle introduit la notion d'incidence : certains systèmes, considérés à incidence élevée, doivent être soumis à des mesures préventives.

Toutefois, le texte ne contenait aucune précision sur la définition des systèmes à incidence élevée dans sa version initiale. Ceci dénotait une volonté de proposer un texte flexible et adaptable à une technologie en pleine évolution. Ainsi, des règlements plus précis pourront ultérieurement être adoptés pour, notamment, établir les critères quant à l'application de la définition de système à incidence élevée, définir ce qui constitue ou non un préjudice important, de même qu'instituer ce qui constitue ou non une justification pour l'application de la définition de résultat biaisé.

Selon la LIAD, c'est le responsable d'un SIA, soit la personne qui le conçoit, le développe, ou le rend disponible, qui devra déterminer si celui-ci a une incidence élevée. Si tel est le cas, cette personne - ce qui inclut « les fiducies, les sociétés de personnes, les coentreprises, les associations non dotées de la personnalité morale et toute autre entité juridique » (*Projet de loi C-27,* 2022, art. 2) - devra mettre en place des mesures pour atténuer les risques de résultats biaisés ou de préjudices, ainsi qu'évaluer l'efficacité de ces mesures (*Projet de loi C-27,* 2022, art. 7 à 9). Toute personne gérant un SIA à incidence élevée devra également publier, sur un site web accessible, des informations qui incluent l'utilisation faite du SIA, le contenu qu'il génère ou les décisions qu'il prend et les mesures d'atténuation établies (*Projet de loi C-27,* 2022, art. 11).

En cas de suspicion de contravention, le ou la ministre chargé(e) de l'application de la loi peut prendre différentes mesures par ordonnance, par exemple ordonner que la personne en question fournisse des documents, imposer une vérification du SIA par un tiers indépendant, ou même faire cesser l'utilisation du SIA ou sa mise sur le marché (art. 13 à 17). En cas de violation de la loi, une amende peut être imposée à la personne qui en est responsable.

La LIAD, dans sa version originale, présentait de nombreuses limitations. Tout d'abord, l'absence de définition d'incidence élevée a suscité de vives critiques de la part de membres de l'industrie et de la société civile, en raison de l'incertitude quant à quels systèmes seraient concernés. Le flou similaire entourant les notions de préjudice et de biais créait lui aussi une incertitude juridique. À la suite de nombreuses critiques, le Gouvernment canadien, à l'origine du texte de loi, a introduit des amendements pour définir ce que signifie l'incidence élevée et proposer des mesures spécifiques concernant l'IA à usage général et l'IA générative. Le gouvernement a ainsi proposé lui-même des amendements à son propre texte devant le comité INDU (Comité permanent de l'industrie et de la technologie) à la Chambre des Communes (Champagne, 2023).

L'ensemble du texte et les amendements étant encore en discussion, il n'est pas utile d'entrer ici dans les détails, mais de signaler simplement que le gouvernement canadien propose de définir les systèmes d'IA à incidence élevée par catégories d'usage, sur le modèle de la proposition de règlement de l'AI Act de l'Union européenne. Selon un des amendements proposés, il s'agirait de suivre une approche sectorielle et d'instaurer des catégories de systèmes ayant une incidence élevée dans sept secteurs d'activité des SIA : l'emploi, la détermination de l'accès aux services, les données biométriques, la modération des plateformes de communication en ligne, la santé, les procédures judiciaires ou administratives, le contrôle d'application des lois par les officiers de la paix.

Dans la version originale de la LIAD, l'absence de définition d'incidence élevée a suscité de vives critiques de la part de membres de l'industrie et de la société civile, en raison de l'incertitude quant à quels systèmes seraient concernés. Le flou similaire entourant les notions de préjudice et de biais créait lui aussi une incertitude juridique.

Cette approche sectorielle s'est révélée problématique dans l'Union Européenne au vu de l'arrivée sur le marché de systèmes sans finalité unique (Boine, 2022), mais le gouvernement canadien tient compte de ces évolutions technologiques puisqu'un amendement propose que la LIAD impose des obligations spécifiques dans le cadre de l'IA à usage général (GPAI), sans les préciser à ce stade. Comme l'IA générative est un type de GPAI, les obligations qui seront imposées aux GPAI s'appliqueront à cette dernière. En outre, le gouvernement chercherait à proposer des modifications qui permettraient de s'assurer que les Canadiens puissent identifier le contenu généré par l'IA s'il y a un risque raisonnablement prévisible qu'une personne communiquant avec un système puisse croire qu'il est humain. La personne qui gère les opérations de ce système devra informer la personne qu'il ne l'est pas. De même, l'amendement propose d'ajouter dans la loi que « les personnes qui mettent au point des systèmes à usage général qui produisent du texte ou du contenu audiovisuel doivent s'efforcer de s'assurer que le public puisse l'identifier » (Champagne, 2023).

Outre le projet de loi C-27, le gouvernement fédéral a également publié des normes non obligatoires pour encadrer l'IA générative utilisée par les institutions fédérales et par les entreprises. Il a ainsi dévoilé :

- Guide sur l'utilisation de l'IA générative adressé aux institutions fédérales (Gouvernement du Canada, 2023b);
- Code de conduite volontaire visant un développement et une gestion responsables des systèmes d'IA générative avancés adressé aux entreprises (ISDE, 2023)

En résumé : depuis la proposition de la LIAD dans le projet de loi C-27 révélée en juin 2022, le gouvernement canadien a bien pris conscience des enjeux spécifiques générés par l'IA générative et, plus largement, l'IA à usage général, en proposant aux entreprises et institutions fédérales de suivre des normes de conduite non obligatoires et en proposant des amendements au projet de loi C-27 qui restent encore à définir dans le détail lors des débats parlementaires.

3.2 Les initiatives au sein de l'Union européenne

Le 21 avril 2021, la Commission européenne a proposé la première législation transversale au monde pour encadrer l'intelligence artificielle (2021), appelée le "AI Act". Cependant, bien qu'avant-gardiste, ce texte législatif ne prévoyait pas la régulation de l'IA générative en tant que telle. D'une part, la proposition de règlement assujettissait les développeurs de SIA à des mesures qui dépendent de la finalité du système, ce qui posait donc un problème lorsque le système n'a pas de finalité propre, comme c'est le cas pour l'IA générative (Boine, 2022). D'autre part, la proposition de règlement visait principalement un type de système en particulier, les statistiques d'aide à la décision, qui est très éloigné de l'IA générative (Boine et Rolnick, 2023). L'arrivée de ChatGPT sur le marché en novembre 2022 a ainsi remis en question l'approche adoptée par la Commission Européenne. En effet, face à la réalité, les législateurs européens ont vite constaté que leur proposition de règlement n'était pas adaptée à cette nouvelle technologie, ce qui a obligé à modifier le texte de la proposition pour l'adapter à l'IA générative d'abord dans la version votée par le Conseil de l'Union européenne en décembre 2022, puis dans celle adoptée par le Parlement européen en juin 2023. Finalement, un compromis a été trouvé par le trilogue, qui a adopté un accord politique le 9 décembre 2023. À noter qu'il reste toutefois des étapes institutionnelles pour un vote définitif. Le texte final n'est en outre pas encore publié à la date de rédaction de ce rapport. Les trois institutions (Commission européenne, Conseil et Parlement européen) doivent désormais trouver un compromis pour l'adoption d'une version unique et définitive du texte, ce qui pourrait survenir au début de 2024[8].

Le Conseil et le Parlement européen tentent d'encadrer l'**IA à usage général** (**GPAI** ou **general purpose AI**) ou à finalités multiples (*multi-purpose AI*). Plus récemment, le Parlement européen cherchait même à définir les modèles de fondation (ou modèles fondationnels) par opposition aux modèles simples. Outre la définition de ces différents systèmes, l'enjeu principal tient à la détermination des responsables dans une chaîne de valeur plus complexe que ce qui était compris initialement. Quant aux définitions, le texte du Parlement européen définit les modèles de fondation et les modèles à finalité générale (**article 3**) :

(1a) « modèle de fondation » : un modèle de système d'intelligence artificielle qui est entraîné sur de vastes données à grande échelle, qui est conçu pour la généralité des résultats et qui peut être adapté à un large éventail de tâches distinctes ;

(1b) « système d'IA à usage général », un système d'IA qui peut être utilisé et adapté à un large éventail d'applications pour lesquelles il n'a pas été intentionnellement et spécifiquement conçu ;

(1c) On entend par « grandes séries d'entraînement » le processus de production de modèles d'IA puissants qui nécessitent des ressources informatiques supérieures à un seuil très élevé. La difficulté à ce jour est que le législateur européen n'a pas encore voté définitivement cette définition ni surtout défini le seuil.

8 Voir Castets-Renard (2023)

Quant à la **responsabilité**, les règles de responsabilité tentent de tenir compte de la chaîne de valeur des SIA à finalité générale : les « déployeurs » (utilisateurs) sont responsables, tout comme les fournisseurs, les importateurs, distributeurs ou autres parties tierces qui attribuent un objectif risqué (annexe III) à l'IA à usage général. Ils seraient responsables comme des fournisseurs d'IA à haut risque.

Plus spécifiquement, **le projet d'article 28b** prévoit des obligations propres au fournisseur d'un modèle de fondation qui devra par exemple s'assurer de ne traiter et incorporer que des ensembles de données soumis à des mesures appropriées de gouvernance des données pour les modèles de fondation, en particulier des mesures visant à examiner l'adéquation des sources de données et les éventuels biais et erreurs, et les mesures d'atténuation appropriées. Il devra aussi élaborer une documentation technique extensive et documenter un résumé de l'utilisation des données d'entraînement protégées par le droit d'auteur. Ces dispositions font encore l'objet d'intenses négociations dans le cadre du trilogue entre la Commission européenne, le Conseil et le Parlement européen.

3.3 Des difficultés sans frontières

L'arrivée de ChatGPT sur le marché a également conduit d'autres juridictions à entrevoir l'adoption de lois sur l'IA. Ainsi, les États-Unis envisagent la création d'une agence administrative chargée de l'IA et ont adopté en octobre 2023 un décret présidentiel (*executive order*) imposant des mesures de sécurité aux développeurs de SIA (The White House, 2023). Le Brésil a présenté une proposition de loi encadrant l'IA (Uechi et Moraes, 2023) reprenant sensiblement l'AI Act européen. La Chine a adopté une loi sur l'IA générative en juillet 2023 (Agence France-Presse, 2023). Elle prévoit aussi d'adopter une autre loi sur les enjeux de sécurité de l'IA générative (McFarland, 2023). Enfin, plusieurs États dont le Canada, la Chine, les États-Unis et l'UE ont signé début novembre au Royaume-Uni la déclaration de Bletchley pour un développement « sûr » de l'intelligence artificielle (IA) visant en particulier l'IA générative, bien que la portée de ce texte demeure symbolique car il est vague dans son énoncé et ne présente aucune force obligatoire.

Cela dit, les débats législatifs concernant le droit de l'IA dans les différents pays partagent certaines difficultés. La première concerne la définition large de l'intelligence artificielle. En effet, en incluant dans la définition des technologies différentes aux fonctionnements différents, il est impossible pour les législateurs de trouver des mesures efficaces et adaptées. La seconde problématique réside dans la tension entre finalité générale et usages particuliers. Ainsi, les SIA à finalité générale qui peuvent être utilisés dans une multitude de contextes particuliers sont particulièrement imprévisibles car leurs mises en marché dépendent de leur interaction avec le contexte de déploiement. Or, il est impossible de tester le système dans tous les contextes possibles pour s'assurer que le déploiement ne présente pas de risque majeur. Il est donc particulièrement difficile de trouver des formes d'encadrement adéquates pour de tels systèmes. Celles proposées aujourd'hui, notamment par le gouvernement canadien via les amendements, se verront probablement insuffisantes pour encadrer de façon satisfaisante les SIA à venir.

4
Un appel à la vigilance et à l'action

4. Un appel à la vigilance et à l'action

Nous avons jusqu'ici mis en lumière les enjeux et risques des IA génératives qui nous semblaient les plus importants et préoccupants à l'heure actuelle, mais aussi pour l'avenir de nos sociétés. Nous avons aussi tenté de faire ressortir que tous ces enjeux et risques appellent la nécessité d'une réflexion sérieuse quant à leur utilisation potentielle, ainsi qu'à des actions pour repenser et mieux répartir les pouvoirs et responsabilités qui y sont associés.

Cet appel à la vigilance et à l'action ne se limite pas aux IA génératives. Nous avons précédemment abordé certaines initiatives qui ont déjà été mises en place ou qui sont en cours de développement, comme les initiatives législatives canadiennes et européennes en cours d'élaboration, ainsi que les chartes et cadres éthiques proposés par les entreprises et diverses instances, gouvernementales ou non, un peu partout dans le monde. Or, à la lumière des enjeux que nous avons identifiés, on constate que ces initiatives actuelles ne sont pas suffisantes pour tenir compte et minimiser les risques qui seront potentiellement créés par l'IA générative et les pratiques en regard de celle-ci.

Des actions supplémentaires doivent par conséquent être rapidement envisagées, et ce, par plusieurs acteurs et à différents niveaux. Afin de couvrir la grande diversité de besoins et d'enjeux liés à l'IA, il nous semble de plus essentiel d'éviter de chercher à développer des solutions englobantes qui tenteraient de répondre à l'ensemble de ces enjeux et des besoins, pour plutôt développer un ensemble d'initiatives aux fonctions et visées différenciées, mais de façon complémentaire. Ainsi, on souhaitera déployer d'une part des initiatives dont la **fonction en est une de contrôle des pratiques** et d'autre part déployer des initiatives dont la **fonction en sera plutôt une de capacitation** des actrices et acteurs et des organisations.

Afin de couvrir la grande diversité de besoins et d'enjeux liés à l'IA, il nous semble essentiel de développer un ensemble d'initiatives aux fonctions et visées différenciées de façon complémentaire.

Dans cette perspective, trois niveaux d'intervention complémentaires nous apparaissent prioritaires :

- les **initiatives normatives et législatives** en matière d'IA (fonction de contrôle);
- les **initiatives éthiques** en matière d'IA (fonction de capacitation);
- les **initiatives de démocratisation** de l'IA (fonction de capacitation).

Nous proposons par conséquent dans ce qui suit des pistes d'action susceptibles d'inspirer des actions complémentaires à chacun de ces trois niveaux d'initiatives.

4.1 Les initiatives normatives et législatives en matière d'IA

Piste d'action 1 : Mettre en place des normes de droit reconnues au niveau national et international

S'il est un constat que le déploiement des applications d'IA génératives nous a permis de dresser au cours de la dernière année, c'est que notre contexte réglementaire actuel, en dépit du fait que l'IA soit assujettie aux régimes de droit en place, est peu adapté à ces nouvelles technologies, notamment parce qu'elles soulèvent des enjeux inédits et évoluent à un rythme sans précédent. Il est par conséquent indéniable que des réformes législatives et de nouvelles règles doivent rapidement être mises en place afin d'encadrer adéquatement ces nouvelles technologies. Sans empêcher toute innovation et développement de l'IA, ces modifications doivent être suffisamment exigeantes pour prévenir les risques et impacts préjudiciables de ces technologies. À cet égard, des mesures obligatoires permettant d'anticiper et de minimiser les risques relatifs à la sécurité, à la santé et aux droits fondamentaux associés au déploiement et à l'utilisation de l'IA, en portant une attention particulière aux personnes ou populations vulnérables, doivent être mises en place et ce, le plus rapidement possible.

Plusieurs axes d'action sont à considérer :

- **Énoncer des règles de sanctions et accorder des pouvoirs de contrôle à des autorités administratives** pour encadrer l'IA au sens large et l'IA à usage général (incluant l'IA générative)

Toutes les mesures d'encadrement légal devront être sévèrement sanctionnées par des juridictions (sanctions civiles et pénales), mais devraient l'être aussi par des organes administratifs (sanctions administratives) qu'il faudra doter de pouvoirs. À cet égard, le *Commissaire à l'AI et aux données* envisagées dans la LIAD n'est pas suffisamment indépendant des services d'Innovation, Sciences et Développement économique Canada (ISDE) et de son ministre, ni suffisamment puissant puisque ses moyens ne sont pas définis. Il n'a pas non plus de pouvoir de sanction dans le projet de loi actuel. Or, il faudrait le doter d'un tel pouvoir car les sanctions administratives sont plus faciles à mettre en œuvre que des sanctions judiciaires.

Énoncer des règles de droit propres à l'IA à usage général (incluant l'IA générative), en plus des règles visant à encadrer l'IA au sens large

Les règles concernant l'IA au sens large reposent sur une approche par les risques, liés à des secteurs d'activité et des usages déterminés. Doivent s'ajouter à cela des règles propres aux systèmes à IA général (incluant l'IA générative) qui, par définition, peuvent être mis en œuvre pour différents types d'usage.

- **Énoncer aussi des règles propres aux modèles de fondation** dont le respect devrait être vérifié par un audit externe et indépendant (voir sur ce dernier point la piste d'action 3)

Compte tenu de l'ampleur de ces modèles, il est indispensable de prendre des précautions dès leur mise sur le marché et avant qu'ils soient repris et modifiés à large échelle par les « déployeurs » pour des usages potentiellement très nombreux. L'autorégulation en la matière ne serait pas suffisamment protectrice. Le coût de l'audit devrait être supporté par les entreprises (généralement de grande taille) qui les mettent sur le marché.

« Compte tenu de l'ampleur de ces modèles, il est indispensable de prendre des précautions dès leur mise sur le marché. […] L'autorégulation en la matière ne serait pas suffisamment protectrice. »

- **Renforcer en particulier la protection du droit d'auteur et des données personnelles**, alors que les risques d'atteinte sont amplifiés par les IA génératives

Un des défis majeurs auquel sera confronté le législateur est la cohérence et l'intégration de ces nouvelles dispositions législatives avec le droit existant. La protection des données personnelles et du droit d'auteur devra être renforcée par les législations spéciales les concernant.

- **Rendre compatibles les lois adoptées au Canada avec ses provinces et à l'échelle internationale** pour plus d'efficacité

Les questions de compétence entre le gouvernement fédéral et les provinces et territoires sont en cours d'analyse. Suivant les secteurs d'activité concernés, les provinces et territoires peuvent conserver une compétence, mais pour éviter un patchwork normatif inefficace, nous appelons à adopter une approche relevant d'un « constitutionnalisme coopératif » (Pelletier, 2023)[9] entre les différents paliers de gouvernement.

2 Piste d'action 2 : Mettre en place des structures de gouvernance adaptées à la nature évolutive de l'IA

Il est par ailleurs essentiel que les initiatives législatives en développement s'appuient sur des structures de gouvernance adaptées à la réalité de l'IA et de l'IA générative.

Un des défis est la mise en place d'un lieu d'échange, de coordination et de prise de décision impliquant une diversité d'expertises et d'expériences, qui vont des sciences sociales, aux sciences politiques à l'éthique, au droit et à la participation citoyenne et démocratique.

D'abord, il importera que ces structures de gouvernance capitalisent sur les instances déjà en place dont l'expertise est pertinente en matière d'IA, notamment, au Québec : la Commission d'accès à l'information, la Commission des droits de la personne, l'Office de la protection des consommateurs ou encore la Commission de l'éthique en science et en technologie (CEST). Des mécanismes devront par ailleurs permettre de faire les ponts entre les différents ministères, organismes publics et paliers gouvernementaux qui s'intéressent à l'IA et ont entamé une réflexion sur la question.

La gouvernance de l'IA interpelle un vaste ensemble d'acteurs et d'organisations. Un des défis de la mise en place d'une gouvernance appropriée à l'IA sera donc la mise en place d'un lieu d'échange, de coordination et de prise de décision impliquant une diversité d'expertises et d'expériences, qui vont des sciences sociales, aux sciences politiques à l'éthique, au droit et à la participation citoyenne et démocratique. Il faudra donc penser cette gouvernance comme une structure agile et évolutive spécifiquement dédiée à l'encadrement du développement, du déploiement et de l'utilisation de l'IA s'appuyant sur des mécanismes démocratiques et consultatifs.

9 Cet article fait partie du dossier spécial "Comment encadrer l'intelligence artificielle?" dirigé par Céline Castets-Renard dans *Options politiques* et accessible au : https://policyoptions.irpp.org/fr/magazines/september-2023/comment-legiferer-sur-lintelligence-artificielle/ « après Option politiques

Compte tenu de la nature évolutive de l'IA et des enjeux spécifiques qu'elle générera dans différents domaines d'activité, sa gouvernance nécessitera par ailleurs une approche contextuelle et adaptée. Dans cette optique, on pourrait vouloir l'appuyer non seulement sur les règles de droit, mais aussi explorer les possibilités que comportent des approches nouvelles comme le droit de la gouvernance (Lasserrre, 2015), qui implique l'intégration de différentes sources de normativités à l'encadrement de l'IA, ou encore les lois à exigence réflexive (Lalonde et Bernatchez, 2016), qui impliquent une participation active des destinataires de la norme à la production des règles. Dans tous les cas, de nouveaux moyens et mesures gagneront à être mis à la disposition des acteurs pour soutenir adéquatement la gouvernance de l'IA et il importera que les règles et structures de gouvernance mises en place pour l'encadrer fassent l'objet d'ajustements constants à la lumière de l'évolution de la technologie et des expériences de toutes les parties prenantes.

Piste d'action 3 : Intégrer des obligations d'audit et de reddition de comptes pour favoriser la responsabilité des entreprises productrices d'IA

Les nouvelles règles ne peuvent se limiter à encadrer les technologies d'IA elles-mêmes : elles doivent aussi venir mieux baliser les pratiques des entreprises qui les développent et commercialisent, celles-ci ayant à l'heure actuelle une latitude presque totale pour mettre en société des produits dont les effets sont non seulement encore inconnus, mais qui sont parfois même encore en développement, tel qu'on le constate avec ChatGPT et ses versions ultérieures.

La seule voie du recours aux tribunaux lorsqu'il y a faute de la part d'une entreprise est trop lente et lourde pour être satisfaisante.

Comme nous l'avons vu précédemment avec le cas d'OpenAI, les initiatives volontaires comportent trop de limites pour constituer une solution viable pour assurer que les entreprises productrices d'IA assument adéquatement leurs responsabilités envers la société. Il faudra donc, pour ce faire, intégrer aux nouvelles lois touchant l'IA des obligations claires et des mesures de reddition de comptes obligatoires de la part de ces entreprises. À cela devront s'ajouter des mécanismes d'imputabilité et de sanctions qui pourront être facilement et rapidement mis en œuvre en cas de dérapage ou de dommages, la seule voie du recours aux tribunaux lorsqu'il y a faute de la part d'une entreprise étant trop lente et lourde pour être satisfaisante.

Ces mesures légales doivent par ailleurs être complémentées par d'autres types de mesures normatives, via des certifications et des normes émises par des organismes reconnus, comme le proposent déjà ISO et le Conseil canadien des normes (CCN). D'autres mécanismes qui se révéleraient pertinents pour encadrer les pratiques d'entreprises qui développent de l'IA ou en homologuer les produits gagneront aussi à être considérés.

On voudra ainsi s'assurer de contrebalancer les importants pouvoirs détenus par les grandes entreprises d'IA, en exigeant qu'ils assument des responsabilités envers la société qui y soient proportionnelles, que les responsabilités des entreprises d'IA ne soient pas qu'optionnelles et tributaires de la bonne volonté de leurs dirigeants, mais qu'elles soient balisées par des obligations claires sur lesquelles elles devront rendre des comptes à l'ensemble de la société.

4.2 Les initiatives éthiques en matière d'IA

Piste d'action 4 : S'appuyer sur des initiatives et outils éthiques reconnus

Considérant l'ampleur des conséquences et des effets potentiels des IA génératives, il est essentiel que leur développement soit aligné avec les valeurs et objectifs de nos sociétés et de l'humanité. Cet alignement ne peut toutefois être assuré en privé par les seules entreprises qui les développent : il doit impérativement se faire de façon transparente et s'appuyer sur des principes, objectifs et balises édictés et reconnus par les parties prenantes et la communauté internationale. À cet égard, on dénombre déjà quelques centaines de chartes et déclarations éthiques pour guider le développement de l'IA proposées au cours des dernières années et près de 700 outils pour accompagner leur mise en œuvre ont été recensés dans un catalogue tenu par l'OCDE. Pour être crédible et adéquat, l'alignement des systèmes d'IA générative doit par conséquent s'appuyer sur de tels balises, méthodes et outils ayant fait l'objet d'une délibération et d'une reconnaissance publiques.

Pour être crédible et adéquat, l'alignement du développement des systèmes d'IA générative doit s'appuyer sur des balises, méthodes et outils ayant déjà fait l'objet d'une délibération et d'une reconnaissance publiques.

Piste d'action 5 : Envisager les initiatives éthiques comme un processus plutôt que comme une liste de principes

Pour que le déploiement de l'IA générative puisse être considéré responsable, ses développeurs ne peuvent toutefois se limiter à endosser symboliquement un ensemble de principes éthiques ou à faire des déclarations d'intentions. Si ceux-ci ne s'accompagnent pas de pratiques conséquentes, ils risqueront alors d'être considérés comme de simples discours de façade (Floridi, 2019). Il s'agit d'ailleurs là d'une critique que l'on voit adressée de façon croissante aux discours éthiques dans le domaine de l'IA et qui a donné lieu à un mouvement de scepticisme face aux initiatives éthiques en IA (Bietti, 2020; Munn, 2023).

Pour assurer l'authenticité et la crédibilité des engagements éthiques dans le domaine de l'IA, il sera donc important d'approfondir de façon substantielle la teneur et la portée des initiatives éthiques qui y sont déployées afin d'en exploiter tout leur potentiel réflexif, collaboratif et participatif, tant au sein des organisations qui développent l'IA, qu'au sein de celles qui l'utilisent ou qui l'encadrent. Il s'agit, bien au-delà des seules déclarations d'intentions, d'envisager **l'éthique comme un processus collaboratif de réflexion critique, d'enquête et d'évaluation** qui gagnera à être activé **tout au long du cycle de vie des applications d'IA, de leur idéation (en amont) jusqu'à leur introduction en société (en aval)** (Marchildon, 2021, 2023). On voudra que ces processus soient par ailleurs **inclusifs des différentes parties prenantes sociétales potentiellement impactées** (de près ou de loin) par les SIA, et non seulement réservés aux seuls développeurs, experts, et parfois aussi aux régulateurs, comme c'est souvent le cas à l'heure actuelle[10].

10 Voir *L'éthique au coeur de l'IA* (Langlois et al., 2023) sur l'importance de la réflexion et de la délibération en éthique de l'IA.

Car pour que ces processus se rapprochent de ce que l'on pourrait considérer comme un dialogue éthique, il faudra qu'elles permettent d'évaluer, à partir de plusieurs perspectives, dans quelle mesure les SIA qu'on envisage introduire dans la société **permettront, par leurs impacts, d'actualiser (ou non) les valeurs auxquelles nous tenons** et de faire advenir le type de société que nous souhaitons toutes et tous (et non la seule vision d'un cercle restreint d'experts, de techno-optimistes ou d'entrepreneurs).

Enfin, pour permettre ces conversations inclusives, celles-ci devront se dérouler simultanément ou parallèlement dans divers lieux – tant au sein de l'État et de l'industrie, que dans la sphère publique, les milieux de recherche et les entreprises privées – et ce, afin d'en faciliter l'accès à différents participant.e.s et groupes et ainsi maximiser la participation d'un plus grand nombre et d'une plus grande diversité d'acteur.trice.s sociaux possibles.

Il est à cet égard important de situer l'éthique comme complément au droit positif. Il s'agit de penser l'intervention de l'éthique dans la gouvernance de l'IA d'une part en amont du droit, c'est-à-dire comme réflexion axiologique sur le bien-fondé des lois et sur la potentielle nécessité de les réformer, mais aussi au sein même du droit ainsi qu'en aval de celui-ci. En effet, au-delà de l'espace limité de la prescription ou de l'interdiction législative, qui porte sur le caractère souhaitable ou non du déploiement de l'IA dans nos différentes sphères d'activités, cette réflexion peut mener au développement d'outils d'évaluation éthiques contextuels et réflexifs, adaptés à la mission, aux valeurs ou aux orientations des organisations qui intègrent l'IA générative à leurs pratiques.

À cet égard, on parle à l'heure actuelle de « l'alignement moral » au regard des objectifs des systèmes d'IA – et en particulier de ChatGPT - avec les valeurs humaines. Cela signifie de faire en sorte que les SIA agissent en fonction de ce qui est considéré comme acceptable ou souhaitable. Les limites de ce qui est acceptable et l'horizon du souhaitable doivent être discutés avec une diversité d'acteurs et d'actrices. Or, à l'heure actuelle, ce sont les entreprises elles-mêmes qui en déterminent les orientations. Ainsi, lorsqu'OpenAI utilise le renforcement par feedback humain, c'est elle qui décide de ce qui constitue un contenu approprié (acceptable) ou non. Cela s'avère toutefois problématique d'un point de vue éthique, puisqu'une telle décision nécessite l'ouverture d'un débat inclusif et démocratique sur l'alignement moral des systèmes d'IA, afin que les entreprises ne définissent pas les buts et valeurs des SIA en vase clos, mais bien en concertation avec celles et ceux sur qui et avec qui ils seront utilisés. Si cet alignement constitue une tâche ardue, il doit néanmoins être au cœur des préoccupations en matière de sécurité et d'éthique de l'IA. Par conséquent, un accompagnement et des outils peuvent être offerts aux chercheuses et chercheurs ainsi qu'aux entreprises qui développent les SIA, qui n'ont pas toujours les compétences éthiques nécessaires pour intégrer adéquatement les besoins et attentes sociétales dans leurs innovations technologiques.

Sans ces adaptations concrètes et effectives des SIA aux préoccupations éthiques et sociales, les réflexions, discussions et déclarations de valeurs risquent fort d'être perçues comme des discours vides et des consultations bidon. Au lieu de renforcer la confiance, c'est alors au contraire le scepticisme à l'égard de cette technologie et des spécialistes et institutions qui la soutiennent qui s'en verra alimentée.

4.3 Les initiatives de démocratisation de l'IA

6 Piste d'action 6 : Favoriser le développement de compétences pour une participation plus large et plus diversifiée

Être aux discusssions sur l'IA est essentiel pour favoriser la participation des actrices et acteurs sociaux, mais cela n'en constitue tout de même que la première étape. Car pour favoriser une pleine participation aux conversations autour de l'IA, il importe que chaque personne puisse aussi y contribuer de manière significative et informée. Ainsi, il sera nécessaire de **favoriser la capacitation des actrices et acteurs** que l'on souhaitera voir participer aux discussions sur l'IA, ce qui implique des efforts de littératie en matière d'IA, mais aussi le développement de leur compétence éthique. De telles initiatives de capacitation devront être orientées vers le public de façon large – incluant les jeunes et les personnes âgées –, mais aussi vers les régulateurs et décideurs publics, qui accusent souvent un déficit important de connaissances lorsqu'il est question de technologies de pointe comme l'IA, ce que l'on constate d'ailleurs avec acuité face à l'arrivée des IA génératives. En ce qui concerne les développeurs d'IA, que ce soit dans les milieux académiques ou en entreprise privée, c'est dans leur cas sur le développement de leur sensibilité et compétence éthique que la capacitation devra être axée. À cet égard, tant le milieu de l'éducation que les milieux de travail ont un rôle particulièrement important à jouer, que ce soit en termes de connaissances fondamentales, de littératie en matière d'IA, de développement de compétences professionnelles en phase avec la nouvelle réalité de l'IA ou de compétences éthiques et démocratiques. Accepter de relever ce défi leur permettra par ailleurs de ne pas se cantonner dans une position de victime du développement de l'IA, pour plutôt se poser en acteurs qui en orientent la trajectoire, et ainsi regagner du pouvoir face à cette technologie et à ceux qui la commercialisent.

« … tant le milieu de l'éducation que les milieux de travail ont un rôle particulièrement important à jouer, que ce soit en termes de connaissances fondamentales, de littératie en matière d'IA, de développement de compétences… »

7 Piste d'action 7 : Favoriser la participation citoyenne dans l'évaluation de l'acceptabilité (sociale) et dans l'orientation des SIA

Dans une perspective de démocratisation - où la démocratie ne serait pas seulement représentative, mais aussi inclusive et participative - on souhaitera **que le public et la société civile prennent part aux initiatives visant à évaluer et orienter le développement de l'IA**.

De telles mesures de démocratisation de l'IA seront nécessaires afin que son développement et sa commercialisation se fassent dans l'intérêt des individus et des collectivités, et non seulement dans l'intérêt de ceux qui la développent, bref, dans une perspective d'intérêt public. Dans cette perspective, il semble important de souligner la nécessité (de ne pas cesser) d'investir dans le développement d'une **IA publique**, c'est-à-dire une IA financée par le secteur public de façon à s'assurer qu'elle soit bel et bien orientée vers l'intérêt public.

« … que le public et la société civile prennent part aux initiatives visant à évaluer et orienter le développement de l'IA. »

5
Lexique

Lexique [11]

Apprentissage automatique / *machine learning* : Branche de l'IA étudiant la manière dont les systèmes informatiques peuvent améliorer leur perception, leurs connaissances, leurs décisions ou leurs actions sur la base de l'expérience ou des données.

Apprentissage profond / *deep learning* : Forme avancée **d'apprentissage automatique (*Machine Learning*)** utilisant de grands réseaux neuronaux (artificiels) multicouches qui calculent avec des représentations continues (nombres réels), de manière semblable aux neurones organisés hiérarchiquement dans le cerveau humain.

Apprentissage supervisé / *supervised learning*: Forme d'**apprentissage automatique** où l'ordinateur apprend à faire des prédictions à partir de données préalablement étiquetées, c'est-à-dire catégorisées, par l'être humain.

Intelligence artificielle (IA) / système d'intelligence artificielle (SIA) : « Un système d'intelligence artificielle (ou système d'IA) est un système automatisé qui, pour des objectifs explicites ou implicites, déduit, à partir d'entrées reçues, comment générer des résultats en sortie tels que des prévisions, des contenus, des recommandations ou des décisions qui peuvent influer sur des environnements physiques ou virtuels. Différents systèmes d'IA présentent des degrés variables d'autonomie et d'adaptabilité après déploiement. » (OCDE, 2023).

IA à usage général / *general purpose AI* (GPAI) : Le parlement européen définit un système d'IA à usage général comme étant « un système d'IA qui peut être utilisé et adapté à un large éventail d'applications pour lesquelles il n'a pas été intentionnellement et spécifiquement conçu » (2023). Ce terme est souvent utilisé de manière interchangeable avec « **modèles de fondation** ».

Intelligence artificielle générale / *artifical general intelligence* (AGI) : Systèmes qui démontrent de vastes capacités d'intelligence (égales ou supérieures au niveau humain), incluant le raisonnement, la planification et la capacité à apprendre par l'expérience. À noter que ce terme fait référence à une forme d'IA qui n'existe pas actuellement, bien que plusieurs entreprises poursuivent l'objectif de développer de tels systèmes.

IA générative : Type de système d'IA pouvant créer une grande variété de données et de documents, telles que des images, des vidéos, du son, du texte et des modèles 3D.

Modèles multimodaux / IA générative multimodale : Modèles en mesure de produire des résultats dans plusieurs modalités (texte, image, audio, etc.) ET de recevoir des requêtes (*prompts*) provenant de ces différentes modalités.

Modèles *Transformers* : Type d'architecture de réseau neuronal pouvant suivre les relations dans les données séquentielles et étant à la base de la plupart des **modèles de fondation** actuels.

Grands modèles de langages / *Large language models* (LLMs) : Type de système d'IA entraîné sur de vastes quantités de données textuelles et capable de générer des réponses en langage naturel à un large éventail d'entrées.

Modèles de fondation / *foundation models* : Type de réseau neuronal d'IA entraîné sur de vastes données à grande échelle et pouvant être adapté à un large éventail de tâches et servant de base pour une multitude d'applications d'IA, y compris l'IA générative. Ce terme est souvent utilisé de manière interchangeable avec « **IA à usage général (GPAI)** ».

11 Toutes les définitions du lexique sont tirées et traduites du glossaire d'une publication du *Ada Lovelace Institute* (Jones, 2023), à l'exception des termes « apprentissage automatique », « Apprentissage profond » et « Apprentissage supervisé » qui proviennent du *Stanford Institute for Human-Centered AI* (Manning, 2022), ainsi qu'« Intelligence artificielle » qui correspond à la nouvelle définition de l'OCDE (2023).

6 Bibliographie

Bibliographie

Adams, R. (2021). Can artificial intelligence be decolonized? *Interdisciplinary Science Reviews*, 46(1-2), 176–197. https://doi.org/10.1080/03080188.2020.1840225

Agence France-Presse. (2023). L'intelligence artificielle encadrée par une nouvelle réglementation en Chine. *Agence France-Presse* dans *Le Journal de Montréal.* https://www.journaldemontreal.com/2023/08/18/lintelligence-artificielle-encadree-par-une-nouvelle-reglementation-en-chine#:~:text=L%E2%80%99intelligence%20artificielle%20encadr%C3%A9e%20par%20une%20nouvelle%20r%C3%A9glementation%20en%20Chine,-Photo%20AFP&text=La%20Chine%20a%20lanc%C3%A9%20cette,en%20maintenant%20un%20strict%20contr%C3%B4le.

Agostinelli, A., Denk T. I., Borsos, Z., Engel, J., Verzetti, M., Caillon, A., Huang, Q., Jansen, A., Roberts, A., Tagliasacchi, M., Sharifi, M., Zeghidour, N. et Frank, C. (2023). MusicLM: Generating Music From Text. *arXiv.*15p. https://doi.org/10.48550/arXiv.2301.11325

Altman, S. (2023). *OpenAI CEO, CTO on risks and how AI will reshape society* [entrevue de Rebecca Jarvis] . *ABC News*. 17 mars. https://abcnews.go.com/Technology/video/openai-ceo-cto-risks-ai-reshape-society-97949497

Anctil, D. (2023a). L'éducation supérieure à l'ère de l'IA générative. *Pédagogie collégiale*. 36(3). https://eduq.info/xmlui/bitstream/handle/11515/38833/Anctil-36-3-23.pdf?sequence=2

Anctil, D. (2023b). Copilot, l'assistant virtuel qui s'apprête à bouleverser le monde du travail [entrevue d'Annie Labrecque]. *Québec Science.* 27 octobre. https://www.quebecscience.qc.ca/technologie/copilot-assistant-monde-travail/

Axiotes, C. (2023). Lobbying for Loopholes: The Battle Over Foundation Models in the EU AI Act. *Euractiv.* https://www.euractiv.com/section/digital/opinion/lobbying-for-loopholes-the-battle-over-foundation-models-in-the-eu-ai-act/

Boudreau LeBlanc, A., Monteferrante, E. et Verreault, G. (2021). Écosystème de gouvernance et technologie: une source d'innovation ou de confusion ?, Éthique publique. 23(2). https://doi.org/10.4000/ethiquepublique.6563

Buruk, B., Ekmekci, P. E., et Arda, B. (2020). A critical perspective on guidelines for responsible and trustworthy artificial intelligence. *Medicine, Health Care and Philosophy*, 23(3), 387-399.

Bietti, E. (2020). From ethics washing to ethics bashing: a view on tech ethics from within moral philosophy. Dans *Proceedings of the 2020 conference on fairness, accountability, and transparency*. https://doi.org/10.1145/3351095.3372860

Boine, C. (2022). « L'IA générale et la proposition de règlement de la Commission européenne », Dalloz IP/IT, N°2, Février 2022. https://www.dalloz-revues.fr/revues/Dalloz_IP_IT-750.htm

Boine, C. et Rolnick, D. (2023). *General Purpose AI Systems in the AI Act: trying to fit a square peg into a round hole.* Proceedings of We Robot 2023. https://www.bu.edu/law/files/2023/09/General-Purpose-AI-systems-in-the-AI-Act.pdf

Boine, C. et Castets-Renard, C. (2023). ChatGPT : le plagiat n'est que l'arbre qui cache la forêt. *The Conversation.* https://theconversation.com/chatgpt-le-plagiat-nest-que-larbre-qui-cache-la-foret-198972

Bommasani, R. et al. (2021). On the opportunities and risks of foundation models. *arXiv*. 214p. https://doi.org/10.48550/arXiv.2108.07258

Borji, A. (2023). A categorical archive of chatgpt failures. *arXiv* (preprint). https://doi.org/10.48550/arXiv.2302.03494

Briggs, J. et, Kodnani, D. (2023) The potentially large effects of artificial intelligence on economic growth. *Goldman Sachs*. https://www.gspublishing.com/content/research/en/reports/2023/03/27/d64e052b-0f6e-45d7-967b-d7be35fabd16.html

Brynjolfsson, E., Li, D. et Raymond, L.D. (2023). Generative AI at Work. *National Bureau of Economic Research*. NBER Working Paper No. 31161. DOI 10.3386/w31161

Bubek, S. et Chandrasekaran, V., Eldan, R.,Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y.T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M.T. et Zhang, Y. (2023). Sparks of Artificial General Intelligence: Early experiments with GPT-4. *arXiv.* https://doi.org/10.48550/arXiv.2303.12712

Buolamwini, J. et Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification.Proceedings of the 1st Conference on Fairness, Accountability and Transparency, PMLR 81. https://proceedings.mlr.press/v81/buolamwini18a.html

Bussières McNicoll, F. (2023). Plusieurs universités hésitent à se lancer dans une « course aux armements » contre l'IA. *Radio-Canada.* https://ici.radio-canada.ca/nouvelle/1997383/intelligence-plagiat-universites-etudiants

Cardon, P.W., Getchell, K., Carradini, S., Fleischmann, C. et Stapp, J. (2023). Generative AI in the Workplace: Employee Perspectives of ChatGPT Benefits and Organizational Policies . *SocArXiv* (preprint). https://doi.org/10.31235/osf.io/b3ezy.

Castets-Renard, C. (2021), "AI and the Law in the E.U. and the U.S." in F. Martin-Bariteau et T. Scassa (eds.) *Artificial Intelligence and the Law in Canada*, LexisNexis Canada, pp. 398-421

Castets-Renard, C. (2023). Proposition de règlement sur l'intelligence artificielle (derniers développements). *Recueil Dalloz,* (13), 680. https://www.dalloz.fr/documentation/Document?id=RECUEIL/CHRON/2023/0565

Castets-Renard, C. et Eynard, J. (dir.). (2023). *Droit de l'intelligence artificielle : entre règles sectorielles et régime général - Perspectives de droit comparé*, Bruylant, 990 p.

Chafkin, M. et Metz, R. (2023). What We Know So Far About Why OpenAI Fired Sam Altman. *Time.* https://time.com/6337437/sam-altman-openai-fired-why-microsoft-musk/

Christian, J. (2023). Amazing "Jailbreak" Bypasses ChatGPT's Ethics Safeguards. *Futurism.* https://futurism.com/amazing-jailbreak-chatgpt

Champagne, F.P. (2023). Correspondance de l'honorable François-Philippe Champagne, ministre de l'Innovation, des Sciences et de l'Industrie. *Comité permanent de l'industrie et de la technologie.* https://www.ourcommons.ca/DocumentViewer/fr/44-1/INDU/document-pertinent/12600809

Chui, M., Roberts, R., Hazan, E., Singla, A., Smaje, K., Sukharevsky, A., Yee, L et Zemmel, R. (2023). The economic potential of generative AI: The next productivity frontier. *McKinsey & Company.* https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-economic-potential-of-generative-AI-the-next-productivity-frontier#introduction

Collins, Sara. (2023). AI Doesn't Need to Move Fast and Break Things. *Public Knowledge.* https://publicknowledge.org/ai-doesnt-need-to-move-fast-and-break-things/

Commission de l'éthique en science et en technologie (CEST). (2023). Les enjeux éthiques et pédagogiques de l'utilisation des intelligences artificielles génératives en enseignement supérieur. https://www.ethique.gouv.qc.ca/fr/projets-en-cours/les-enjeux-ethiques-et-pedagogiques-de-l-utilisation-des-intelligences-artificielles-generatives-en-enseignement-superieur/

Commission européenne. (2021). *Règlement du Parlement européen et du Conseil établissant des règles harmonisées concernant l'intelligence artificielle (législation sur l'intelligence artificielle) et modifiant certains actes législatifs de l'Union*. https://eur-lex.europa.eu/legal-content/FR/TXT/?uri=celex%3A52021PC0206

CNN. (2023). *How Microsoft's AI is messing up the news* [vidéo]. Youtube. https://www.youtube.com/watch?v=mGHqz-BJz84

Conseil du statut de la femme (CSF). (2023). Avis sur l'intelligence artificielle : des risques pour l'égalité entre les femmes et les hommes. https://csf.gouv.qc.ca/article/publicationsnum/avis-intelligence-artificielle/

Davis, W. (2023). Sarah Silverman is suing OpenAI and Meta for copyright infringement. *The Verge.* https://www.theverge.com/2023/7/9/23788741/sarah-silverman-openai-meta-chatgpt-llama-copyright-infringement-chatbots-artificial-intelligence-ai

Déclaration de Montréal. (2018). *La Déclaration de Montréal pour un développement responsable de l'intelligence artificielle*. https://declarationmontreal-iaresponsable.com/la-declaration/

Dignum, V. (2022). Relational artificial intelligence. *arXiv preprint*. https://doi.org/10.48550/arXiv.2202.07446

Duhigg, C. (2023). The Inside Story of Microsoft's Partnership with OpenAI. *The New Yorker.* https://www.newyorker.com/magazine/2023/12/11/the-inside-story-of-microsofts-partnership-with-openai

Elhage, N., Nanda, N., Olsson, C., Henighan, T., Joseph, N., Mann, B., ... et Olah, C. (2021). A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 1. https://transformer-circuits.pub/2021/framework/index.html

Ellingrud, K., Sanghvi, S., Singh Dandona, G., Madgavkar, A., Chui, M., White,O. et Hasebe, P. (2023). Generative AI and the future of work in America. *McKinsey Global Institute*. https://www.mckinsey.com/mgi/our-research/generative-ai-and-the-future-of-work-in-america

Eloundou, T., Manning, S., Mishkin, P., et Rock, D. (2023). GPTs are GPTs: An early look at the labor market impact potential of large language models. *arXiv* (preprint). https://doi.org/10.48550/arXiv.2303.10130

Fernández Gibaja, A. (2023). ChatGTP et la démocratie (dossier *Mutations numériques*). *La Grande Conversation*. https://www.lagrandeconversation.com/societe/chatgtp-et-la-democratie/

Felten, E.W., Raj, M., Seamans, R. (2023). Occupational Heterogeneity in Exposure to Generative AI. Disponible sur *SSRN*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4414065

Field, H. (2023). Who's on the OpenAI board — the group behind Sam Altman's ouster. *CNBC*. https://www.cnbc.com/2023/11/18/heres-whos-on-openais-board-the-group-behind-sam-altmans-ouster.html

Floridi, L. (2019). Translating principles into practices of digital ethics: five risks of being unethical. *Philosophy & Technology*, 32(2), 185–193. https://doi.org/10.1007/s13347-019-00354-x

Franceschi-Bicchierai, L. (2023, 20 avril). Jailbreak tricks Discord's new chatbot into sharing napalm and meth instructions. *TechCrunch*. https://techcrunch.com/2023/04/20/jailbreak-tricks-discords-new-chatbot-intosharing-napalm-and-meth-instructions/

Gallienne, R. et Poibeau, T. (2023). Quelques observations sur la notion de biais dans les modèles de langue. 18e Conférence en Recherche d'Information et Applications – 16e Rencontres Jeunes Chercheurs en RI – 30e Conférence sur le Traitement Automatique des Langues Naturelles – 25e Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues, Juin 2023. https://hal.science/hal-04130210/file/480736.pdf

Ganguli, D., Hernandez, D., Lovitt, L., Askell, A., Bai, Y., Chen, A., ... et Clark, J. (2022). Predictability and surprise in large generative models. Dans *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 1747-1764. https://doi.org/10.1145/3531146.3533229

Goldberg, E. (2023). A.I.'s Threat to Jobs Prompts Question of Who Protects Workers. *The New York Times*. https://www.nytimes.com/2023/05/23/business/jobs-protections-artificial-intelligence.html

Gmyrek, P., Berg, J. et Bescond, D. (2023). Generative AI and jobs: A global analysis of potential effects on job quantity and quality. *International Labour Organization*. https://www.ilo.org/wcmsp5/groups/public/---dgreports/---inst/documents/publication/wcms_890761.pdf

Gouvernement du Canada. (2023a). *La Loi sur l'intelligence artificielle et les données (LIAD) – document complémentaire*. Consulté en avril 2023. https://ised-isde.canada.ca/site/innover-meilleur-canada/fr/loi-lintelligence-artificielle-donnees-liad-document-complementaire#s4

Gouvernement du Canada. (2023b). *Guide sur l'utilisation de l'intelligence artificielle générative*. https://www.canada.ca/fr/gouvernement/systeme/gouvernement-numerique/innovations-gouvernementales-numeriques/utilisation-responsable-ai/guide-utilisation-intelligence-artificielle-generative.html

Greene, D., Hoffmann, A. L., et Stark, L. (2019). Better, nicer, clearer, fairer: A critical assessment of the movement for ethical artificial intelligence and machine learning. *Proceedings of the 52nd Hawaii International Conference on System Sciences*. https://scholarspace.manoa.hawaii.edu/items/2c91a03d-caad-4c6f-9069-9bca4f393ce6

Hadfield, G. (2023). OpenAI: « Ce n'était pas une bataille d'entreprise ordinaire ». [entrevue de Chloé Sondervorst]. *Radio-Canada*. https://ici.radio-canada.ca/nouvelle/2030752/openai-bataille-entreprise-chatgpt-gillian-hadfield

Hartmann, J., Schwenzow, J., et Witte, M. (2023). The political ideology of conversational AI: Converging evidence on ChatGPT's pro-environmental, left-libertarian orientation. *arXiv preprint*. https://doi.org/10.48550/arXiv.2301.01768

ISDE (Innovation, Sciences et Développement économique Canada). (2023). *Code de conduite volontaire visant un développement et une gestion responsables des systèmes d'IA générative avancés*. https://ised-isde.canada.ca/site/isde/fr/code-conduite-volontaire-visant-developpement-gestion-responsables-systemes-dia-generative-avances

Jobin, A., Ienca, M., et Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature machine intelligence*, 1(9), 389-399. https://doi.org/10.1038/s42256-019-0088-2

Jones, E. (2023). Explainer: What is a foundation model? *Ada Lovelace Institute*. https://www.adalovelaceinstitute.org/resource/foundation-models-explainer/#_ftn41

Kang, C. et Metz, C. (2023). F.T.C. Opens Investigation Into ChatGPT Maker Over Technology's Potential Harms. *The New York Times.* https://www.nytimes.com/2023/07/13/technology/chatgpt-investigation-ftc-openai.html?emc=edit_na_20230713&ref=cta&nl=breaking-news

Kaplan, J., McCandlish, S., Henighan, T., Brown, T.B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J. et Amodei, D. (2020) Scaling Laws for Neural Language Models . *arXiv.* https://doi.org/10.48550/arXiv.2001.08361

Konrad, A. et Cai, K. (2023). Exclusive Interview: OpenAI's Sam Altman Talks ChatGPT And How Artificial General Intelligence Can 'Break Capitalism. *Forbes.* https://www.forbes.com/sites/alexkonrad/2023/02/03/exclusive-openai-sam-altman-chatgpt-agi-google-search/?sh=28d3953e6a63

KPMG. (2023a). *Répertoire sur l'adoption de l'IA générative*. https://collimateur.uqam.ca/wp-content/uploads/sites/11/2023/09/kpmg-au-canada-repertoire-sur-ladoption-de-lia-generative-2023.pdf.

KPMG. (2023b). Selon 6 étudiants sur 10, utiliser l'IA générative c'est tricher. https://kpmg.com/ca/fr/home/media/press-releases/2023/08/six-in-ten-students-consider-generative-ai-cheating.html

Lalonde, L. et Bernatchez, S. (2016). *La norme juridique reformatée : perspectives québécoises des notions de force normative et de sources revisitées*. Éditions de la Revue de droit de l'Université de Sherbrooke (RDUS).

Langlois, L., Dilhac, M.A., Dratwa, J., Ménissier, T., Ganascia, J.G., Weinstock, D., Bégin, L. et Marchildon, A. (2023). *L'éthique au cœur de l'IA*. Obvia. https://observatoire-ia.ulaval.ca/lethique-au-coeur-de-lia/

Lasserrre, V. (2015). *Le nouvel ordre juridique : Le droit de la gouvernance*. LexisNexis.

Legros, C. et Balagué, C. (2023). Intelligence artificielle et recrutement : Typologie, controverses et pratiques responsables. *Good In Tech*. https://cf.appdrag.com/goodintech-4900d2/uploads/files/0194b94a-1409-49ff-b70c-fa14326ab337.pdf

Li, P., Yang, J., Islam, M. A., et Ren, S. (2023). Making AI Less» Thirsty»: Uncovering and Addressing the Secret Water Footprint of AI Models. *arXiv* (preprint) https://doi.org/10.48550/arXiv.2304.03271

Liang, W., Yuksekgonul, M., Mao, Y., Wu, E. et Zou, J (2023). GPT detectors are biased against non-native English writers. *arXiv* (preprint). arXiv:2304.02819

Lu, Y. (2023). As Businesses Clamor for Workplace A.I., Tech Companies Rush to Provide It. *The New York Times.* https://www.nytimes.com/2023/07/05/technology/business-ai-technology.html

Luccioni, A. S., Viguier, S., et Ligozat, A. L. (2022). Estimating the carbon footprint of bloom, a 176b parameter language model. *arXiv* (preprint). https://doi.org/10.48550/arXiv.2211.02001

Mangan, D. (2024). Microsoft, OpenAI sued for copyright infringement by nonfiction book authors in class action claim. *CNBC*. https://www.cnbc.com/2024/01/05/microsoft-openai-sued-over-copyright-infringement-by-authors.html

Manning, C. (2022). Artificial Intelligence Definitions. *Stanford Human-Centered Artificial Intelligence (HAI).* https://hai.stanford.edu/sites/default/files/2023-03/AI-Key-Terms-Glossary-Definition.pdf

Marchildon, A. (2016). Corporate responsibility or corporate power? CSR and the shaping of the definitions and solutions to our public problems. *Journal of Political Power*, 9(1), 45-64. https://doi.org/10.1080/2158379X.2016.1149310

Marchildon, A. (2017). Le pouvoir de déployer la compétence éthique. Éthique publique. 19(1). https://doi.org/10.4000/ethiquepublique.2920

Marchildon, A. (2021, 31 août). *Développement et mise en société de nouvelles technologies : Quels enjeux éthiques et comment s'en préoccuper?*[communication orale]. Symposium en interfaces neuronales, Université de Sherbrooke. https://event.fourwaves.com/fr/sisn/resumes/0597658f-fac6-48ce-a34c-3865f8eb9850

Marchildon, A. (2023, 8 juin). *Ethics and AI: from Discourse to Practice* [communication orale], Canadian AI conference-Responsible AI Track, Université McGill. https://www.caiac.ca/en/conferences/canadianai-2023/responsible-ai

McAdoo, T. (2023). How to cite ChatGPT. *APA Style.* https://apastyle.apa.org/blog/how-to-cite-chatgpt

McFarland, A. (2023). La Chine cible la sécurité des données de l'IA générative avec de nouvelles propositions réglementaires. *Unite.AI.* https://www.unite.ai/fr/china-targets-generative-ai-data-security-with-fresh-regulatory-proposals/

Metz, C., Mickle, T., Isaac, M., Weise, K., Rosse, K. (2023). Five Days of Chaos: How Sam Altman Returned to OpenAI. *The New York Times.* https://www.nytimes.com/2023/11/22/technology/how-sam-altman-returned-openai.html

Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI. Nature machine intelligence, 1(11), 501-507. https://doi.org/10.1038/s42256-019-0114-4

Mitchell, M. et Krakauer, D.C. (2023). The Debate Over Understanding in AI's Large Language Models. *arXiv*. https://doi.org/10.48550/arXiv.2210.13966

Munn, L. (2023). The uselessness of AI ethics. *AI and Ethics*, 3(3), 869–877. https://doi.org/10.1007/s43681-022-00209-w

Murgia, M. (2023). OpenAI's red team: the experts hired to 'break' ChatGPT. *Financial Times.* https://www.ft.com/content/0876687a-f8b7-4b39-b513-5fee942831e8

Musk, E. [@elonmusk]. (2023, 17 février). *OpenAI was created as an open source (which is why I named it "Open" AI), non-profit company to serve as a counterweight to Google, but now it has become a closed source, maximum-profit company effectively controlled by Microsoft. Not what I intended at all* [Tweet]. X/Twitter. https://twitter.com/elonmusk/status/1626516035863212034?ref_src=twsrc%5Etfw%7Ctwcamp%5Etweetembed%7Ctwterm%5E1626516035863212034%7Ctwgr%5E8810a59940ae3fac797b82e5f912e6708ccc3214%7Ctwcon%5Es1_&ref_url=https%3A%2F%2Ffortune.com%2F2023%2F02%2F17%2Fchatgpt-elon-musk-openai-microsoft-company-regulator-oversight%2F

Nicoletti, L. et Bass, D. (2023). Humans are biased. Generative AI is even worse. *Bloomberg.* https://www.bloomberg.com/graphics/2023-generative-ai-bias/

Noy, S. et Zhang, W. (2023). Experimental evidence on the productivity effects of generative artificial intelligence. *Science.* https://www.science.org/doi/10.1126/science.adh2586

OCDE. (2023). *Recommandation du Conseil sur l'intelligence artificielle.* (adoptée le 21 mai 2019 et amendée le 7 août 2023). https://legalinstruments.oecd.org/fr/instruments/OECD-LEGAL-0449

OpenAI. (2018). OpenAI Charter. https://openai.com/charter

OpenAI. (2023a). GPT-4 Technical Report. https://cdn.openai.com/papers/gpt-4.pdf

OpenAI. (2023b). GPT-4 System Card. https://cdn.openai.com/papers/gpt-4-system-card.pdf

OpenAI. (2023c). Teaching with AI. https://openai.com/blog/teaching-with-ai

OpenAI. (2023d). Our structure. https://openai.com/our-structure

OpenAI. (2023e). OpenAI announces leadership transition. https://openai.com/blog/openai-announces-leadership-transition

PAI. (2023). *PAI's Guidance for Safe Foundation Model Deployment: A Framework for Collective Action*. https://partnershiponai.org/modeldeployment/#landing

Parlement européen. (2023). *Amendements du Parlement européen, adoptés le 14 juin 2023, à la proposition de règlement du Parlement européen et du Conseil établissant des règles harmonisées concernant l'intelligence artificielle (législation sur l'intelligence artificielle) et modifiant certains actes législatifs de l'Union*. https://www.europarl.europa.eu/doceo/document/TA-9-2023-0236_FR.html.

Pelletier, B. (2023). Qui a compétence sur l'intelligence artificielle : Ottawa ou les provinces? dans Castets-Renard, C. (dir.), Comment encadrer l'intelligence artificielle ? [dossier spécial], *Options politiques* [en ligne] https://policyoptions.irpp.org/fr/magazines/september-2023/comment-legiferer-sur-lintelligence-artificielle/

Perrigo, B. (2023). OpenAI Used Kenyan Workers on Less Than $2 Per Hour to Make ChatGPT Less Toxic. *Time*. https://time.com/6247678/openai-chatgpt-kenya-workers/

Poibeau, T. (2023). Les modèles de langue nous apprennent-ils quelque chose sur le langage ?. *Hypotheses*(Hors-texte). https://horstexte.hypotheses.org/183

Projet de loi C-27 : Loi édictant la Loi sur la protection de la vie privée des consommateurs, la Loi sur le Tribunal de la protection des renseignements personnels et des données et la Loi sur l'intelligence artificielle et les données et apportant des modifications corrélatives et connexes à d'autres lois. Projet de loi C-27 (1ère lecture-16juin 2022), 1ère sess., 44e légis. (Can.). https://www.parl.ca/DocumentViewer/fr/44-1/projet-loi/C-27/premiere-lecture

Rozado, D. (2023). The political biases of ChatGPT. *Social Sciences*, 12(3), 148. https://doi.org/10.3390/socsci12030148

Russell, S. J.et Norvig, P. (2010). Artificial intelligence: a modern approach (3e édition). Prentice Hall.

Sabourin Laflamme, A. (2023). L'enseignement supérieur à l'épreuve de l'IA générative. CScience. https://www.cscience.ca/chroniques/lenseignement-superieur-a-lepreuve-de-lia-generative/

Sabzalieva, E., et Valentini, A. (2023). ChatGPT and artificial intelligence in higher education: quick start guide. *UNESCO*. https://unesdoc.unesco.org/ark:/48223/pf0000385146.locale=en

Sadasivan, V. S., Kumar, A., Balasubramanian, S., Wang, W. et Feizi, S. (2023). Can ai-generated text be reliably detected?. *arXiv* (preprint). arXiv:2303.11156.

Sanders, N.E. et Schneier, B. (2023). How ChatGPT Hijacks Democracy. *The New York Times*. https://www.nytimes.com/2023/01/15/opinion/ai-chatgpt-lobbying-democracy.html

Schiffer, Z. et Newton, C. (2023). Microsoft lays off team that taught employees how to make AI tools responsibly. *The Verge*. https://www.theverge.com/2023/3/13/23638823/microsoft-ethics-society-team-responsible-ai-layoffs

Shafto, P. (2023). Why big tech companies are open-sourcing their AI systems. *The Conversation*. https://theconversation.com/why-big-tech-companies-are-open-sourcing-their-ai-systems-54437

Shrivastava, R. (2023). AI Chatbots Are The New Job Interviewers. *Forbes*. https://www.forbes.com/sites/rashishrivastava/2023/07/26/ai-chatbots-are-the-new-job-interviewers/?sh=57d4b992e3a4

Smalley, S. (2023). Could ChatGPT supercharge false narratives? *Poynter* https://www.poynter.org/ifcn/2023/could-chatgpt-supercharge-false-narratives/

Stanford HAI. (2023). Genretive AI : Perspectives from Stanford Human-Centered Artificial Intelligence. https://hai.stanford.edu/sites/default/files/2023-03/Generative_AI_HAI_Perspectives.pdf

Steinhardt, J. (2022). Future ML systems will be qualitatively different. *Bounded Regret*. https://bounded-regret.ghost.io/future-ml-systems-will-be-qualitatively-different/

Stilgoe, J., Owen, R., et Macnaghten, P. (2013). Developing a framework for responsible innovation. *Research Policy*, 42(9). https://doi.org/10.1016/j.respol.2013.05.008

Srivastava, A., Rastogi, A., Rao, A., Shoeb, A. A. M., Abid, A., Fisch, A., ... et Wang, G. (2022). (2022). Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models. *arXiv*. 100p. https://doi.org/10.48550/arXiv.2206.04615

The White House. (2023). FACT SHEET: President Biden Issues Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence. https://www.whitehouse.gov/briefing-room/statements-releases/2023/10/30/fact-sheet-president-biden-issues-executive-order-on-safe-secure-and-trustworthy-artificial-intelligence/

Tidjon, L. N. et Khomh, F. (2022). The Different Faces of AI Ethics Across the World: A Principle-To-Practice Gap Analysis. *IEEE Transactions on Artificial Intelligence*. https://doi.org/10.1109/TAI.2022.3225132

Toner, H. (2023). What Are Generative AI, Large Language Models, and Foundation Models?. *Center for Security and Emerging Technology (CSET).* https://cset.georgetown.edu/article/what-are-generative-ai-large-language-models-and-foundation-models/

Turing, A.M. (1950). I.—COMPUTING MACHINERY AND INTELLIGENCE. *Mind*, 59(236). https://doi.org/10.1093/mind/LIX.236.433

Uechi, C. A. S. et Moraes, T.G. (2023). Brazil's path to responsible AI. *OECD.AI*. https://oecd.ai/en/wonk/brazils-path-to-responsible-ai

UNESCO. (2023). Enquête de l'UNESCO : moins de 10 % des établissements scolaires et des universités encadrent officiellement l'utilisation de l'IA. https://www.unesco.org/fr/articles/enquete-de-lunesco-moins-de-10-des-etablissements-scolaires-et-des-universites-encadrent

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser et Polosukhin, I. (2017). Attention Is All You Need. *Advances in neural information processing systems*, 30. https://papers.nips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

Verdegem, P. 2021. Introduction: Why We Need Critical Perspectives on AI. In: Verdegem, P (ed.), *AI for Everyone?*. London: University of Westminster Press. https://doi.org/10.16997/book55.a

Vincent, J. (2023). OpenAI co-founder on company's past approach to openly sharing research: 'We were wrong'. *The Verge*. https://www.theverge.com/2023/3/15/23640180/openai-gpt-4-launch-closed-research-ilya-sutskever-interview?fbclid=IwAR2BxUQ1y-wvpRLssSHat2TsfQZwME0dIDQmkQvHWylMgnzvXluL6MODKTa0

Vinuesa, R., Azizpour, H., Leite, I., Balaam, M., Dignum, V., Domisch, S., Felländer, A., Langhans, S. D., Tegmark, M., et Fuso Nerini, F. (2020). The role of artificial intelligence in achieving the sustainable development goals. *Nature Communications*, 11(1). https://doi.org/10.1038/s41467-019-14108-y

Vogel, D. (2006). *The market for virtue: the potential and limits of corporate social responsibility*. Brookings Institution Press.

Weizenbaum, J. (1966). ELIZA: A computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1), 36-45. https://doi.org/10.1145/365153.365168

Wei, J., Tay Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E.H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J. et Fedus, W. (2022). Emergent Abilities of Large Language Models. *Transactions on Machine Learning Research*. https://doi.org/10.48550/arXiv.2206.07682

Williams, T. (2023). Some companies are already replacing workers with ChatGPT, despite warnings it shouldn't be relied on for 'anything important'. *Fortune.* https://fortune.com/2023/02/25/companies-replacing-workers-chatgpt-ai/

Yu, E. (2023). 75% of businesses are implementing or considering bans on ChatGPT. *ZDNET*.https://www-zdnet-com.cdn.ampproject.org/c/s/www.zdnet.com/google-amp/article/75-of-businesses-are-implementing-or-considering-bans-on-chatgpt/

Yip, K. et Balagué, C. (2023). ChatGPT: research evidence based controversies, regulations and solutions. *Good In Tech.* https://cf.appdrag.com/goodintech-4900d2/uploads/files/fd6b27e9-e311-4ebc-a088-8a6dd2b449ab.pdf

Zawacki-Richter, O., Marín, V.I., Bond, M., Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education – where are the educators?. *International Journal of Educational Technology in Higher Education*. 16(39). https://doi.org/10.1186/s41239-019-0171-0

Zhuo, T. Y., Huang, Y., Chen, C., et Xing, Z. (2023). Red teaming ChatGPT via Jailbreaking: Bias, Robustness, Reliability and Toxicity. *arXiv*. https://doi.org/10.48550/arXiv.2301.12867

obvia

www.obvia.ca