Code de conduite volontaire visant un développement et une gestion responsables des systèmes d'IA générative avancés

De : Innovation, Sciences et Développement économique Canada

Septembre 2023

Les systèmes d'intelligence artificielle (IA) avancés capables de générer du contenu, comme ChatGPT, DALL·E 2 et Midjourney, ont capté l'attention du monde entier. Les capacités générales de ces systèmes d'IA offrent un potentiel énorme d'innovation dans un bon nombre de domaines, et ces systèmes sont déjà adoptés et utilisés dans divers contextes. Ces systèmes avancés peuvent être utilisés pour effectuer de nombreuses tâches, comme rédiger des courriels, répondre à des questions complexes, produire des images ou des vidéos réalistes, ou rédiger du code de logiciel.

Bien qu'ils présentent de nombreux avantages, les systèmes d'IA générative avancés comportent également un profil de risque qui est manifestement considérable. Cela s'explique par la vaste portée des données au moyen desquels ils sont entraînés, le large éventail d'utilisations potentielles des systèmes et l'ampleur de leur déploiement. Les systèmes qui sont accessibles au public pour un éventail d'utilisations peuvent présenter des risques pour la santé et la sécurité, propager des préjugés et avoir des répercussions sociétales plus vastes, particulièrement lorsqu'ils sont utilisés par des auteurs malveillants. Par exemple, la capacité de produire des images et des vidéos réalistes ou de se faire passer pour de vraies personnes peut permettre des tromperies d'une envergure qui peut nuire à d'importantes institutions, notamment aux systèmes de justice démocratique et pénale. Ces systèmes peuvent également avoir une incidence importante sur les droits individuels en matière de protection de la vie privée, comme le souligne la Déclaration sur l'IA générative des autorités de protection des données et de la vie privée du G7.

Les organisations peuvent également adapter les systèmes génératifs pour des utilisations précises – comme les applications de gestion des connaissances organisationnelles ou les outils de service à la clientèle – qui présentent généralement un éventail plus restreint de risques. Malgré tout, les développeurs et les gestionnaires de tels systèmes devraient prendre un certain nombre de mesures pour veiller à ce que les risques soient bien cernés et atténués.

Afin de gérer et d'atténuer ces risques, les signataires du présent code s'engagent à adopter les mesures définies. Le code décrit les mesures qui devraient être appliquées par toute organisation qui développe[Note de bas de page 1] ou gère les opérations[Note de bas de page 2] d'un système d'IA génératif ayant des capacités générales, ainsi que les mesures supplémentaires qui devraient être prises par toute organisation qui

développe ou gère les opérations d'un tel système rendu accessible à un vaste public, soit des systèmes dont l'éventail d'utilisations potentiellement nuisibles ou inappropriées est plus vaste. Les organisations qui développent et gèrent des systèmes génératifs de pointe jouent des rôles importants et complémentaires. Les développeurs et les gestionnaires doivent coopérer pour veiller à ce que les répercussions négatives soient examinées par l'acteur approprié.

Bien que le cadre décrit ici soit propre aux systèmes d'IA génératifs avancés, bon nombre des mesures peuvent être appliquées en grande partie à divers systèmes d'IA à incidence élevée et peuvent être facilement adaptées par les organisations de l'écosystème canadien d'IA. Il est également important de noter que les directives ne changent en rien les obligations juridiques que les organisations peuvent avoir, par exemple, au titre de la *Loi sur la protection des renseignements personnels et les documents électroniques*.

Dans le cadre de cet engagement volontaire, les développeurs et les gestionnaires de systèmes génératifs avancés s'engagent à s'efforcer d'atteindre les résultats suivants :

- **Responsabilité** – Les organisations comprennent leur rôle à l'égard des systèmes qu'elles développent ou gèrent, mettent en place des systèmes appropriés de gestion des risques et collaborent avec d'autres organisations au besoin pour éviter qu'il y ait des lacunes.
- **Sécurité** – Des évaluations des risques doivent être réalisées pour les systèmes, et les mesures d'atténuation nécessaires doivent être prises avant le déploiement pour veiller à ce que l'exploitation des systèmes soit sécuritaire.
- **Justice et équité** – L'incidence potentielle en matière de justice et d'équité est évaluée et gérée à différentes étapes de l'élaboration et du déploiement des systèmes.
- **Transparence** – Suffisamment de renseignements sont publiés pour permettre aux consommateurs de prendre des décisions éclairées et aux experts d'évaluer si les risques ont été adéquatement gérés.
- **Surveillance humaine** – L'utilisation du système est surveillée après le déploiement, et des mises à jour sont mises en œuvre au besoin pour gérer les risques qui se matérialisent.
- **Validité et fiabilité** – Les systèmes fonctionnent comme prévu, sont sécurisés contre les cyberattaques, et leur comportement en réponse aux diverses tâches ou situations auxquelles ils sont susceptibles d'être exposés est compris.

Les signataires s'engagent également à soutenir le développement continu d'un écosystème d'IA fiable et responsable au Canada. Notamment, la contribution à l'élaboration et à l'application de normes, la transmission d'information et de pratiques exemplaires à d'autres membres de l'écosystème de l'IA, la collaboration avec des chercheurs qui travaillent pour l'avancement de l'IA responsable et la collaboration avec d'autres intervenants, y compris les gouvernements, pour appuyer la sensibilisation et l'éducation du public à l'égard de l'IA. Les signataires s'engagent également à élaborer et à déployer des systèmes d'IA de manière à favoriser une croissance axée sur l'inclusion et la durabilité au Canada, notamment en accordant la priorité aux droits de la personne, à l'accessibilité et à la durabilité de environnementale, et à exploiter le potentiel de l'IA pour relever les défis mondiaux les plus urgents de notre époque.

Ressources

- Foire aux questions sur le Code de conduite volontaire visant les systèmes d'IA générative avancés.
- Guide de mise en œuvre pour les gestionnaires de systèmes d'intelligence artificielle

Signataires du Code de conduite

Signataires
Ada
AlayaCare
Alberta Machine Intelligence Institute (Amii)
Alloprof
AltaML
Appen
BlackBerry
BlueDot
CGI
CIBC
Clir
Cofomo Inc.
Cohere
Conseil canadien des innovateurs
Coveo
Dayforce Canada Ltée.
Dimonoff Inc.
Geotab Inc.

Signataires
Hewlett Packard Enterprise
IBM
INCA
Institut Vecteur
Intel Corporation
Interac Corp.
Jolera Inc.
kama.ai
Kyndryl
Lenovo
Levio
MaRS Discovery District
Mastercard
Mila
Nuvei
OpenText
Organisme d'autoréglementation du courtage immobilier du Québec (OACIQ)
PaymentEvolution
Protexxa Inc.
Ranovus
Resemble AI
Responsible Artificial Intelligence Institute

Signataires
Salesforce
SAP Canada
Scale AI
TELUS
TELUS Numérique
Workday

Mesures précises à prendre conformément au code de conduite

Principe	Mesures	Systèmes génératifs avancés		Systèmes génératifs avancés qui sont accessibles au public	
		Développeurs	Gestionnaires	Développeurs	Gestionnaires
Responsabilité	**Mettre en œuvre un cadre complet de gestion des risques adapté à la nature et au profil de risque des activités.** Ce cadre comprend la mise en place de politiques, de procédures et de formations pour veiller à ce que les employés connaissent bien leurs responsabilités et les pratiques de gestion des	Oui	Oui	Oui	Oui

Principe	Mesures	Systèmes génératifs avancés		Systèmes génératifs avancés qui sont accessibles au public	
		Développeurs	Gestionnaires	Développeurs	Gestionnaires
	risques de l'organisation.				
	Transmettre l'information et les pratiques exemplaires visant la gestion des risques aux entreprises qui jouent des rôles complémentaires dans l'écosystème.	Oui	Oui	Oui	Oui
	Utiliser plusieurs lignes de défense, notamment des vérifications par des tiers avant le lancement.	Non	Non	Oui	Non
Sécurité	**Effectuer une évaluation complète des répercussions négatives potentielles raisonnablement prévisibles**, notamment des risques associés à une utilisation inappropriée ou malveillante du système.	Oui	Oui	Oui	Oui

Principe	Mesures	Systèmes génératifs avancés		Systèmes génératifs avancés qui sont accessibles au public	
		Développeurs	Gestionnaires	Développeurs	Gestionnaires
	Mettre en œuvre des mesures adaptées pour atténuer les risques de biais, notamment en créant des mesures de protection contre l'utilisation malveillante.	Oui	Non	Oui	Non
	Mettre à la disposition des développeurs et des gestionnaires en aval des conseils sur l'utilisation appropriée du système, notamment des renseignements sur les mesures prises pour gérer les risques.	Oui	Non	Oui	Non

Principe	Mesures	Systèmes génératifs avancés		Systèmes génératifs avancés qui sont accessibles au public	
		Développeurs	Gestionnaires	Développeurs	Gestionnaires
Justice et équité	**Évaluer et organiser les ensembles de données utilisés pour l'entraînement** afin de gérer la qualité des données et les biais potentiels.	Oui	Non	Oui	Non
	Mettre en œuvre diverses méthodes de test et mesures pour évaluer et atténuer le risque d'obtenir des résultats biaisés avant le lancement.	Oui	Non	Oui	Non
Transparence	**Publier de l'information sur les capacités et les limites du système.**	Non	Non	Oui	Non
	Élaborer et mettre en œuvre une méthode fiable et disponible gratuitement pour détecter le contenu généré par le système, et	Non	Non	Oui	Non

Principe	Mesures	Systèmes génératifs avancés		Systèmes génératifs avancés qui sont accessibles au public	
		Développeurs	Gestionnaires	Développeurs	Gestionnaires
	ce, en mettant l'accent à court terme sur le contenu audiovisuel (p. ex. le tatouage numérique).				
	Publier une description des types de données d'entraînement utilisées pour développer le système ainsi que des mesures prises pour déterminer et gérer les risques.	Non	Non	Oui	Non
	Veiller à ce que les systèmes qui pourraient être confondus avec des êtres humains soient clairement et visiblement identifiés comme des systèmes d'IA.	Non	Oui	Non	Oui
Surveillance humaine	**Surveiller le fonctionnement du système pour s'assurer qu'il n'est pas utilisé à**	Non	Oui	Non	Oui

Principe	Mesures	Systèmes génératifs avancés		Systèmes génératifs avancés qui sont accessibles au public	
		Développeurs	Gestionnaires	Développeurs	Gestionnaires
	des fins nuisibles ou qu'il n'a pas des répercussions néfastes après qu'on l'ait rendu accessible, y compris par l'intermédiaire de canaux de rétroaction tiers, et informer le développeur et/ou mettre en œuvre des contrôles d'utilisation au besoin pour atténuer les biais.				
	Maintenir une base de données sur les incidents signalés après le déploiement et fournir des mises à jour au besoin pour veiller à l'efficacité des mesures d'atténuation.	Oui	Non	Oui	Non
Validité et fiabilité	**Utiliser avant le déploiement une grande variété de méthodes de test dans un**	Oui	Non	Oui	Non

Principe	Mesures	Systèmes génératifs avancés		Systèmes génératifs avancés qui sont accessibles au public	
		Développeurs	Gestionnaires	Développeurs	Gestionnaires
	ensemble de tâches et de contextes pour mesurer le rendement et garantir la fiabilité.				
	Avoir recours à des tests adversatifs (c.-à-d. la méthode de l'équipe rouge) pour cerner les vulnérabilités.	Non	Non	Oui	Non
	Effectuer une évaluation des risques liés à la cybersécurité et mettre en œuvre des mesures adaptées pour atténuer les risques, notamment en ce qui a trait à l'empoisonnement des données.	Oui	Non	Oui	Oui
	Effectuer des analyses comparatives pour mesurer le rendement du modèle par rapport aux	Oui	Non	Oui	Non

Principe	Mesures	Systèmes génératifs avancés		Systèmes génératifs avancés qui sont accessibles au public	
		Développeurs	Gestionnaires	Développeurs	Gestionnaires
	normes reconnues.				

Notes de bas de page

Note de bas de page 1

> Le développement comprend la sélection de méthodologies, la collecte et le traitement d'ensembles de données, la création de modèles et les tests.
>
> Retour à la référence de la note de bas de page 1

Note de bas de page 2

> La gestion des opérations comprend la mise en service d'un système, le contrôle des paramètres de son fonctionnement, le contrôle des accès et la surveillance de son fonctionnement.
>
> Retour à la référence de la note de bas de page 2