Publications du Fellowship en politiques IA

Concevoir des agents conversationnels d'IA plus sûrs

Helen A. Hayes

Résumé

Les agents conversationnels d'IA se sont rapidement intégrés dans la vie quotidienne en fonctionnant comme des tuteurs, des confidents et des compagnons. Pour les jeunes en particulier, ces systèmes reproduisent les qualités affectives et relationnelles de la conversation humaine avec suffisamment de fidélité pour façonner la confiance, le comportement et l'attachement émotionnel. Or, les cadres de gouvernance numérique existants sont mal outillés pour faire face aux risques que représentent les agents conversationnels d'IA pour les jeunes, notamment la dépendance émotive, le délestage cognitif et l'exposition accrue à des contenus préjudiciables.

S'appuyant sur les observations de Gen(Z)AI une assemblée nationale de jeunes du Canada de 17 à 23 ans, ainsi que sur une analyse comparative de la réglementation internationale, ce document présente des interventions politiques ciblées visant à protéger les jeunes dans les écosystèmes d'intelligence artificielle conversationnelle. Plus précisément, il propose de modifier deux cadres fédéraux précédemment présentés, la Loi sur les préjudices en ligne (projet de loi C-63) et la Loi sur la protection de la vie privée des consommateurs, afin d'imposer des obligations en matière de sécurisation dès la conception et de conception adaptée à l'âge des agents conversationnels d'IA, de renforcer les protections à la vie privée des enfants et de créer un organisme de surveillance indépendant. Une telle approche permettrait non seulement de réduire les préjudices, mais aussi de positionner le Canada en tant que meneur en matière de sécurité de l'IA centrée sur les jeunes.

Mila

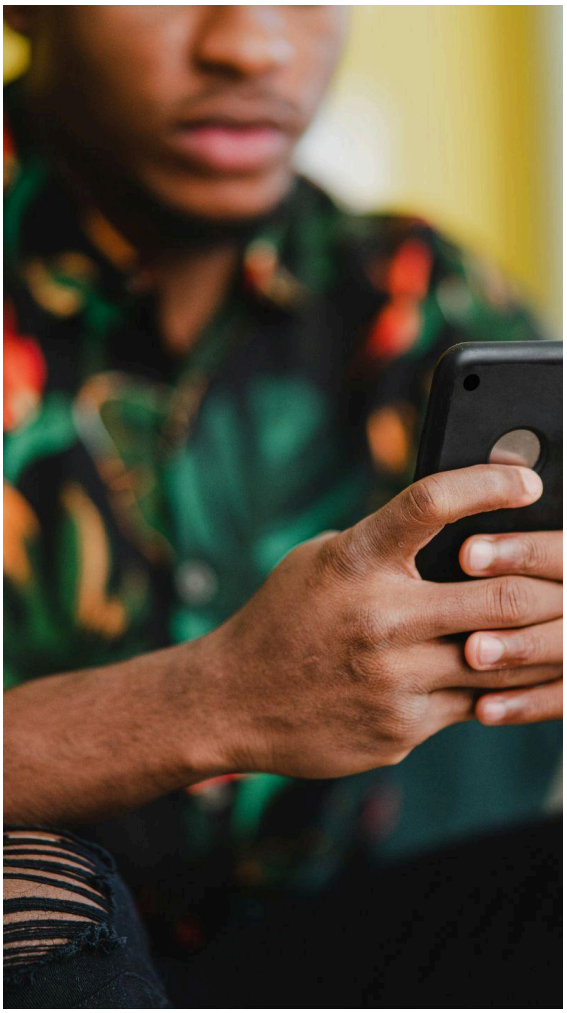
Collaborations et remerciements :

Cette note politique a été réalisée avec le soutien consultatif de Taylor Lynn Curtis et du Dr Adam Oberman dans le cadre de la bourse de recherche en politiques de l'IA de Mila. Les recherches primaires ont été menées dans le cadre de Gen(Z)AI : A National Youth Assembly on Artificial Intelligence, une initiative conjointe du Centre pour les médias, la technologie et la démocratie, du *Dialogue on Technology Project* et de Mila. Gen(Z)AI est financé par la Fondation Waltons, la Fondation Ronald S. Roadburg et l'ICRA.

1. Le contexte

Les systèmes d'IA conversationnelle produisent en temps réel des réponses personnalisées qui simulent l'empathie et la présence sociale, brouillant ainsi les frontières entre la transmission d'informations et l'interaction sociale. Un tiers des adolescent·e·s américain·e·s déclarent préférer se confier à un compagnon d'IA plutôt qu'à un être humain, et deux tiers des enfants britanniques utilisent des agents conversationnels d'IA pour obtenir des conseils d'ordre émotionnel, plus d'un tiers d'entre eux décrivant l'expérience comme équivalente à celle d'une conversation avec un ami ou une amie. Ce sont ces capacités qui rendent les agents conversationnels d'IA attrayants, mais ce sont précisément ces caractéristiques qui posent également les risques les plus élevés pour les jeunes qui les utilisent.

Les enquêtes ont documenté des cas où les agents conversationnels de Meta ont engagé des personnes mineures dans des conversations sexuellement suggestives. My AI de Snapchat a donné des conseils explicites à des enfants en se faisant passer pour un personnage de 13 ans et Character.AI a autorisé des jeux de rôle impliquant des agressions sexuelles et des conversations sur le thème du suicide. Ces cas démontrent que des philosophies de conception préjudiciables sont intégrées dans les systèmes populaires d'agents conversationnels d'IA et qu'elles privilégient un engagement soutenu au détriment de la sécurité de la personne utilisatrice.



Les Canadiens et les Canadiennes en ont pris bonne note. Les données d'un sondage représentatif à l'échelle nationale recueillies par le Centre pour les médias, la technologie et la démocratie montrent que plus de 70 % des répondant·e·s se disent préoccupé·e·s par toutes les grandes catégories de risques liés aux agents conversationnels, notamment en ce qui concerne la santé mentale et l'isolement social, la dépendance émotive et l'incitation à l'automutilation. Leur préoccupation semble porter sur les principes de conception sous-jacents, notamment la maximisation de l'engagement, l'anthropomorphisme et la simulation de l'intimité. En outre, les Canadiens et Canadiennes estiment manifestement que la responsabilité de ces préjudices incombe d'abord aux entreprises d'IA, 69 % des personnes répondantes étant favorables à une réglementation fédérale plus stricte en matière d'IA. Il existe donc un mandat populaire clair pour une intervention politique.

2. Une analyse des politiques selon la juridiction

Le contexte réglementaire canadien en matière de systèmes d'IA se compose d'un ensemble fragmenté de mécanismes qui ne parviennent pas à traiter de manière adéquate les risques distincts posés par les agents conversationnels d'IA pour les jeunes utilisateurs et utilisatrices. Bien que les débats politiques récents signalent une prise de conscience politique croissante des préjudices liés à l'IA, les outils législatifs existants ne s'attaquent pas de manière ciblée aux caractéristiques adaptatives et psychologiquement immersives des agents conversationnels. Trois initiatives législatives antérieures — la Loi sur l'intelligence artificielle et les données (projet de loi C-27), la Loi sur la protection de la vie privée des consommateurs (projet

de loi C-27) et la Loi sur les préjudices en ligne (projet de loi C-63) — ont chacune abordé certains aspects précis des défis de gouvernance décrits dans la présente note politique, mais elles sont devenues caduques suite à la prorogation du Parlement au début de l'année 2025. Ce qui reste, c'est une mosaïque de lois obsolètes, qui ont été conçues pour un environnement technologique fondamentalement différent, et qui laissent les jeunes du Canada sans protection réelle contre les risques distincts que posent les agents conversationnels d'IA.

2.1 Les lacunes réglementaires canadiennes

La loi canadienne qui traite de la protection de la confidentialité dans le secteur privé, la Loi sur la protection des renseignements personnels et les documents électroniques (LPRPDE), a été conçue il y a plus de 20 ans pour un environnement technologique qui ne ressemble guère à l'écosystème de l'IA d'aujourd'hui. Le fait que la LPRPDE exclut les données inférées de sa définition des renseignements personnels est d'une importance cruciale pour la gouvernance des agents conversationnels d'IA[1]. À cause de cette exclusion, les profils comportementaux et psychologiques détaillés générés par les agents conversationnels d'IA à partir des conversations des jeunes utilisateurs et utilisatrices échappent totalement à ses protections. La loi ne contient pas non plus de dispositions régissant la prise de décision automatisée, y compris le droit d'obtenir une explication ou de contester les décisions lorsque les systèmes d'intelligence artificielle procèdent à des déterminations substantielles. La Déclaration sur l'intelligence artificielle et les enfants du Commissariat à la protection de la vie privée, publiée en 2024, énonçait trois défis fondamentaux non résolus qui découlent directement de ces lacunes : la prise de décision opaque de l'IA, qui produit des résultats discriminatoires sans possibilité de contestation ; l'importante capacité de manipulation des systèmes d'IA déployés en tant qu'outils réactifs sur le plan émotif ; et l'utilisation courante, pour entraîner les modèles d'IA, de données relatives aux enfants, souvent extraites de sources accessibles au public ou saisies à partir d'appareils connectés. L'adoption proposée, mais non concrétisée, de la Loi sur la protection de la vie privée des consommateurs aurait remédié directement à plusieurs de ces lacunes, notamment

[1] Le paragraphe 2(1) de la Loi sur la protection des renseignements personnels et les documents électroniques (2000, c. 5) stipule que par « renseignement personnel », on entend « Tout renseignement concernant un individu identifiable ».

en classant les renseignements personnels des personnes mineures comme « de nature sensible » par défaut (art. 2), en renforçant les exigences en matière de consentement pour mettre l'accent sur la compréhensibilité adaptée aux capacités (art. 15 à 18), et en introduisant un droit renforcé au retrait des renseignements personnels, explicitement encadré pour traiter les risques à long terme liés à la persistance des données des enfants (art. 55).

Les lacunes des lois canadiennes en matière de sécurité en ligne sont tout aussi importantes. Le projet de loi C-63, première tentative législative canadienne sur les préjudices en ligne qui n'a pas été adoptée, énonçait les obligations des services en ligne réglementés d'agir de manière responsable, de protéger les enfants et de rendre certains contenus inaccessibles, et prévoyait une commission de la sécurité numérique dotée de vastes pouvoirs d'enquête et d'application de la loi. Toutefois, le projet de loi définissait les services réglementés comme étant ceux qui hébergent ou mettent à disposition des contenus générés par les utilisateurs et utilisatrices (art. 2)[2], ce qui exclut explicitement de son champ d'application les agents conversationnels d'IA et les systèmes d'IA générative orientés vers les consommateurs et consommatrices. Cette exclusion reflète un décalage conceptuel plus profond entre le paradigme canadien de modération de contenu et la nature des préjudices causés par les agents conversationnels d'IA. Les mécanismes traditionnels de signalement et de retrait sont structurellement mal adaptés pour traiter les préjudices qui émergent de la génération personnalisée en temps réel : des préjudices qui sont contextuels, éphémères et produits par la plateforme elle-même plutôt qu'hébergés par des sources externes.

2.2 Les approches internationales

L'évolution internationale montre que la sécurité des enfants en ligne exige de réglementer la conception même des systèmes, et non seulement le contenu qu'ils affichent. Ce passage de la modération de contenu à l'intégration de la responsabilité

[2] Le projet de loi définissait un « service de média social » comme un « [s]ite Web ou application accessible au Canada dont le but principal est de faciliter la communication en ligne à l'échelle interprovinciale ou internationale entre les utilisateurs du site Web ou de l'application en leur permettant d'avoir accès à du contenu et d'en partager. » Le paragraphe 2(2) apportait des précisions supplémentaires : les services de contenu pour adultes, comme les sites Web pornographiques et les services de diffusion en direct, sont également couverts par la définition d'un « service de média social ».

dès la conception s'est accéléré dans de nombreux territoires et pointe directement vers les types d'action recommandés dans cette note politique.

Dans l'Union européenne, cette approche est à un stade de développement très avancé. La Loi sur l'intelligence artificielle de l'Union européenne classe les enfants parmi les personnes vulnérables et interdit explicitement les systèmes d'IA qui déploient des techniques subliminales, manipulatrices ou trompeuses ou qui exploitent les vulnérabilités liées à l'âge pour influencer le comportement (art. 5). Parallèlement à la Loi sur l'intelligence artificielle, le Règlement sur les services numériques (RSN) exige des très grandes plateformes en ligne qu'elles procèdent à des évaluations systémiques des risques en matière de dépendance, de santé mentale et de sécurité des enfants (art. 34-35), impose des mécanismes de signalement adaptés aux enfants et restreint la publicité ciblée destinée aux personnes mineures (art. 28). Ensemble, ces instruments créent une architecture cohérente d'obligations en amont, de responsabilité systémique et d'interdictions exécutoires dont le Canada est actuellement dépourvu.

En ce qui concerne la protection de la vie privée, le Règlement général sur la protection des données (RGPD) de l'UE a établi le principe fondamental selon lequel les protections de la vie privée doivent être intégrées dans l'architecture du système. L'Age Appropriate Design Code du Royaume-Uni, publié en vertu de la Data Protection Act de 2018, concrétise ce principe en exigeant que les services susceptibles d'être consultés par des enfants appliquent par défaut les paramètres de confidentialité les plus élevés possibles. En outre, le commissariat à l'information du Royaume-Uni a proposé d'exiger des organisations qu'elles établissent des objectifs précis, explicites et légitimes de collecte de données à chaque étape du développement de l'IA afin d'éviter que des justifications générales comme « l'amélioration des capacités de l'IA » ne servent de licence fourre-tout pour collecter les renseignements personnels des enfants. Les lois des États américains — notamment la loi californienne California Age Appropriate Design Code Act et les lois parallèles du Maryland, du Nebraska, du Nouveau-Mexique et du Vermont[3] — reflètent la même approche, en exigeant des

[3] Le California Age Appropriate Design Code Act a établi le cadre fondamental adopté par les lois subséquentes de l'État. Le Connecticut interdit le défilement sans fin ; le Maryland exige des pratiques de surveillance par les tuteurs et tutrices et une conception dans l'intérêt supérieur des enfants ; le Nebraska impose des fils d'actualités chronologiques, des pauses des notifications pendant la nuit et les jours d'école, ainsi que des options de limitation du temps d'utilisation ; le Nouveau-Mexique

plateformes qu'elles donnent la priorité à l'intérêt supérieur de l'enfant, qu'elles procèdent à des évaluations de l'impact sur la protection des données et qu'elles configurent les paramètres par défaut de manière à garantir le niveau de confidentialité le plus élevé. De même, le projet de loi brésilien sur la protection numérique des enfants étend ces principes au développement de l'IA en particulier, en interdisant l'utilisation d'images d'enfants dans la formation de l'apprentissage automatique sans le consentement des parents.

L'Australie est allée le plus loin, en désignant explicitement les agents conversationnels d'IA comme des systèmes à haut risque. Son commissaire à la sécurité en ligne a classé les agents conversationnels et les compagnons d'IA comme des technologies à haut risque et publié des orientations demandant aux développeurs et développeuses d'adopter des principes de sécurisation dès la conception, tandis que la Online Safety Act exige des plateformes qu'elles prennent des mesures raisonnables pour empêcher l'exposition des enfants à du matériel préjudiciable, l'évaluation des risques pour la sécurité des enfants étant l'une de ces mesures.

Plusieurs thèmes se dégagent de l'ensemble de ces territoires. Premièrement, une gouvernance efficace considère la conception comme le lieu principal à réglementer, en imposant des obligations aux développeurs et développeuses, donc en amont, avant le déploiement plutôt qu'en aval, quand le résultat a causé un préjudice. Deuxièmement, les enfants sont systématiquement nommés comme formant une catégorie justifiant une protection accrue, que ce soit par l'entremise de classifications de données sensibles, d'exigences de protection de la vie privée par défaut ou d'interdictions totales de conception manipulatrice. Troisièmement, les cadres les plus robustes se caractérisent par la présence d'organismes de surveillance indépendants dotés d'un véritable pouvoir de vérification et d'application, ce qui témoigne de la reconnaissance par les gouvernements du fait que la complexité et l'opacité des systèmes d'IA nécessitent des organismes de réglementation qui ont une expertise spécifique et des pouvoirs d'envergure.

désactiverait les contacts avec des utilisateurs et utilisatrices inconnu·e·s et les notifications entre 22 h et 6 h ; et le Vermont se concentre sur la réduction des fonctionnalités addictives et nuisibles.

3. Les implications politiques et les données probantes

Les idées de politiques présentées dans cette note politique sont tirées d'une analyse comparative des approches réglementaires canadiennes et internationales (voir la section 2) et des conclusions originales tirées de Gen(Z)AI, une assemblée nationale de jeunes sur l'intelligence artificielle qui a réuni 100 Canadiens et Canadiennes de 17 à 23 ans pour délibérer et rédiger des recommandations politiques afin de faire face aux risques des agents conversationnels d'IA. Par l'entremise de délibérations facilitées, les jeunes participants et participantes ont convergé vers trois domaines de risque interdépendants : (1) la dépendance relationnelle ; (2) le délestage cognitif et (3) les préjudices liés au contenu.

Gen(Z)AI : les points de vue des jeunes au Canada

Les jeunes qui ont participé à Gen(Z)AI ont proposé que le gouvernement fédéral : (1) oblige les plateformes à lutter contre la conception addictive des agents conversationnels grâce à des filtres de contenu, à la suppression facultative du cache de données et à des contrôles réglables par l'utilisateur ou l'utilisatrice pour la réactivité et la conversationnalité ; (2) mette en place un système de signalement accessible et oblige les plateformes à signaler rapidement les interactions préjudiciables à un organisme indépendant chargé de faire respecter la loi ; et (3) crée un nouvel organisme gouvernemental indépendant chargé de faire respecter les normes de sécurité en matière d'IA, de réaliser des vérifications algorithmiques et des évaluations des risques, et de fournir aux utilisateurs et utilisatrices des mécanismes de résolution des litiges et de recours.

Ces recommandations s'inscrivent directement dans les mécanismes émergents de conception et de changement des systèmes dans la politique internationale, y compris le RSN et la Loi sur l'intelligence artificielle de l'Union européenne, ainsi que le cadre de sécurisation dès la conception de l'Australie. Elles mettent en évidence un écart national évident entre les engagements du Canada pour une IA responsable et son incapacité à rendre opérationnelles les approches fondées sur la conception dans le cadre de la gouvernance numérique.

4. Des perspectives à concrétiser : des solutions politiques et techniques

Pour soutenir la conception et la réglementation d'agents conversationnels d'IA plus sûrs, cette orientation politique propose de modifier deux cadres politiques fédéraux précédemment déposés : (1) la Loi sur les préjudices en ligne, le projet de loi C-63, et (2) la Loi sur la protection de la vie privée des consommateurs. Cette proposition vise à :

1. Imposer des obligations de sécurité dès la conception et de conception adaptée à l'âge pour les systèmes d'IA, y compris les agents conversationnels d'IA, qui sont susceptibles d'être consultés par des enfants ;
2. Classer par défaut les données personnelles des enfants comme sensibles et étendre ces protections aux données inférées et dérivées ;
3. Créer un organisme indépendant ayant le pouvoir d'imposer l'accès aux données, de mener des vérifications et de faire respecter les obligations.

Cela permettrait non seulement de rapprocher le Canada des meilleures pratiques et normes internationales, mais aussi de rendre opérationnelles bon nombre des recommandations proposées par les jeunes Canadiens et Canadiennes qui ont participé au processus Gen(Z)AI.

4.1 Les obligations en matière de sécurisation dès la conception et de conception adaptée à l'âge

L'inclusion explicite des agents conversationnels d'IA dans le champ d'application d'une nouvelle Loi sur les préjudices en ligne, avec des obligations de sécurisation dès la conception adaptées à leurs caractéristiques distinctes, exigerait des développeurs et développeuses de procéder, avant le déploiement et périodiquement, à des évaluations des risques axées sur les enfants — y compris des évaluations de l'impact sur les droits de l'enfant — avant que les agents conversationnels d'IA susceptibles d'être consultés par des enfants ne soient mis à la disposition du public. Il s'agirait également d'établir des interdictions contraignantes sur les pratiques de conception qui favorisent la dépendance émotive des personnes mineures, y compris les caractéristiques qui simulent l'intimité ou déploient des signaux anthropomorphiques calibrés pour maximiser la confiance dans les systèmes des agents conversationnels.

D’un point de vue technique, ces obligations devraient se traduire par des exigences de conception concrètes et applicables, notamment les suivantes :

1. Des limites obligatoires à la persistance conversationnelle ;
2. Des contrôles réglables par l’utilisateur ou l’utilisatrice pour la réactivité de l’agent conversationnel, les indices anthropomorphiques et l’effet de miroir émotionnel ;
3. L’option de supprimer le cache de données et de minimiser la mémoire par défaut.

4.2 Les données relatives aux enfants : sensibles par défaut

Les cadres de protection de la vie privée fondés sur le consentement ne peuvent pas gérer de manière adéquate les systèmes d’IA qui déduisent, prédisent et influencent le comportement. Pour aller plus loin, les dispositions relatives à la protection de la vie privée du projet de loi C-27 devraient être relancées et réformées, avec deux ajouts essentiels par rapport à ce que proposait le projet de loi. Premièrement, les données personnelles des enfants devraient être classées comme sensibles par défaut, ce qui entraînerait des exigences accrues en matière de consentement, de transparence et de limitation des finalités pour tout traitement de telles données. Deuxièmement, ces protections devraient s’étendre explicitement aux données déduites et dérivées — les profils psychologiques et comportementaux que les agents conversationnels d’IA génèrent par l’analyse de la conversation.

Le corollaire technique est tout aussi important et devrait exiger ce qui suit :

1. La minimisation des données au niveau du modèle et de l’interaction ;
2. L’interdiction de l’utilisation secondaire des données conversationnelles à des fins de formation ou de profilage sans lien avec le sujet ;
3. Des mécanismes de liaison aux objectifs qui limitent la réutilisation en aval des données d’interaction des enfants.

4.3 Le contrôle indépendant par une autorité de vérification et d’application

Une réglementation efficace doit également inclure la création ou la désignation d'un organisme existant en tant qu'organisme de réglementation indépendant. Au minimum, cet organisme de réglementation devrait être habilité à : mener des évaluations de systèmes, des vérifications algorithmiques et des évaluations des risques des agents conversationnels d'IA ; exiger des mesures correctives lorsque des risques pour les enfants sont repérés ; recevoir et traiter les plaintes des utilisateurs et utilisatrices par l'entremise de mécanismes accessibles de règlement des différends ; et coordonner ses actions avec le Commissariat à la protection de la vie privée et tout futur organisme de réglementation responsable de lutter contre les préjudices en ligne, afin d'éviter une fragmentation de la surveillance.

Une surveillance techniquement efficace devrait également exiger que l'organisme de réglementation et les scientifiques accrédités aient accès aux renseignements suivants provenant des plateformes, dans le cadre d'un plan de sécurité numérique : la documentation sur la conception du modèle, les décisions de mise au point, les garde-fous et la logique d'interaction avec l'utilisateur ou l'utilisatrice, y compris les registres d'interaction anonymisés.

5. En guise de conclusion

Les preuves sont claires : les agents conversationnels d'IA susceptibles d'être consultés par les jeunes devraient faire l'objet d'une réglementation. Concrètement, il s'agit d'intégrer les agents conversationnels d'IA dans le champ d'application d'une nouvelle Loi sur les préjudices en ligne — avec des obligations exécutoires en matière de sécurisation dès la conception et de conception adaptée à l'âge — et de relancer et renforcer les réformes en matière de protection de la vie privée proposées dans le projet de loi C-27. Il s'agit également de créer un organisme de réglementation indépendant capable de vérifier, d'évaluer et de faire respecter les normes de sécurité. Bien menée, une telle approche permettrait non seulement de réduire les préjudices, mais aussi de positionner le Canada en tant que chef de file mondial pour ce qui est de garantir que les jeunes puissent utiliser les systèmes d'IA en toute sécurité, avec un esprit critique et en toute confiance.

Clause de non-responsabilité

Les opinions exprimées dans cet article sont strictement celles des auteurs et ne représentent ni ne reflètent nécessairement les politiques ou positions officielles de Mila, de ses affiliés, de ses administrateurs ou de ses bailleurs de fonds. Les auteurs assument l'entière responsabilité de l'exactitude et de l'intégrité de ce travail.

Annexes

Annexe 1. Gen(Z)AI : une assemblée nationale de jeunes sur l'IA

Gen (Z)AI est une initiative nationale de recherche délibérative qui examine comment les jeunes Canadiens et Canadiennes de 17 à 23 ans comprennent, expérimentent et évaluent les défis émergents de la gouvernance de l'IA. De l'automne 2025 au printemps 2026, le projet devait réunir 100 participants et participantes de 17 à 23 ans dans 4 villes du Canada pour délibérer sur une série de domaines politiques interconnectés : les agents conversationnels d'IA (Toronto), l'intégrité de l'information (Montréal), la confidentialité des données (Vancouver) et la garantie de l'âge (Halifax). Le projet utilise un nouveau modèle d'assemblée citoyenne et un processus de rétroaction itératif pour générer des recommandations politiques consensuelles éclairées par les jeunes. Les résultats de chaque forum sont diffusés par l'intermédiaire d'une plateforme numérique nationale hébergée par Make.org afin de solliciter les réactions de milliers d'autres Canadiens et Canadiennes appartenant à la même catégorie d'âge, les idées étant intégrées dans un rapport national final présenté aux décideurs et décideuses politiques, aux organismes de réglementation et aux acteurs et actrices de la société civile. L'initiative Gen(Z)AI est codirigée par Helen Hayes et Fergus Linley-Mota et est animée par les boursiers et boursières 2025/2026 du Centre for Media, Technology, and Democracy : Madeleine Case, Nonso Morah, Julian Lam et Alexander Martin.