



Guide sur l'utilisation de l'intelligence artificielle agentive

Publié : le 5/18/2026

© Sa Majesté le Roi du chef du Canada,
représentée par le président du Conseil du Trésor 2026,

Publié par le Secrétariat du Conseil du Trésor du Canada
90 rue Elgin, Ottawa, Ontario, K1A 0R5, Canada

No de catalogue BT22-298/2026F-PDF
ISBN: 978-1-100-00733-5

Ce document est disponible sur Canada.ca, le site Web du gouvernement du Canada.

Ce document est disponible en médias substituts sur demande.

Nota : Pour ne pas alourdir le texte français, le masculin est utilisé
pour désigner tant les hommes que les femmes.

Also available in English under the title: Guide on the Use of Agentic Artificial Intelligence

Guide sur l'utilisation de l'intelligence artificielle agentive

Sur cette page

- [1. Aperçu](#)
- [2. Qu'est-ce que l'IA agentive?](#)
- [3. Possibilités et risques](#)
- [4. Principes d'utilisation responsable](#)
- [5. Conseils pratiques pour le personnel du gouvernement du Canada \(avant, pendant, après\)](#)
- [6. Études de cas : utilisation responsable de l'IA agentive](#)
- [Nous joindre](#)
- [Glossaire](#)
- [Liens connexes](#)

1. Aperçu

L'intelligence artificielle évolue rapidement, et le gouvernement du Canada commence à s'intéresser à une nouvelle catégorie de systèmes d'IA dont les capacités vont au-delà de la simple création de contenu. Ces systèmes

d'IA peuvent également exécuter des opérations. L'IA agentive représente ce nouveau tournant.

Alors que l'IA générative produit divers résultats en réponse à une requête, l'IA agentive s'appuie sur ces capacités pour exécuter des tâches, enchaîner des étapes, interagir avec des systèmes numériques et viser des objectifs définis dans le cadre de limites établies. En termes simples, si l'IA générative peut décrire ce qui doit être fait, l'IA agentive peut quant à elle tenter de le faire.

Cette distinction est importante pour la fonction publique fédérale. L'IA agentive soulève toutes les mêmes préoccupations que celles associées à l'IA générative, comme l'exactitude, la protection de la vie privée, l'équité et la sécurité. Cependant, à ces préoccupations s'ajoutent de nouveaux risques liés à sa capacité d'agir. Ces risques comprennent les interopérations imprévues avec les systèmes, l'exécution inadéquate des tâches ainsi que la nécessité d'une surveillance et d'une traçabilité accrues.

Par conséquent, l'IA agentive nécessite un cadre de gouvernance qui s'étend au-delà du contenu du Guide sur l'utilisation de l'intelligence artificielle générative, tout en demeurant pleinement conforme aux principes qui y sont énoncés.

Le présent guide fournit aux ministères et organismes des directives pratiques leur permettant d'évaluer, de gérer et de régir ces systèmes d'IA.

L'objectif consiste à aider les organisations fédérales à utiliser les nouvelles capacités de l'IA de manière responsable, sécuritaire et transparente, conformément à l'engagement du gouvernement du Canada visant la prestation de services publics responsable et axée sur les personnes.

Si vous envisagez d'utiliser des systèmes d'IA agentive, veuillez à consulter systématiquement les personnes concernées afin de déterminer l'applicabilité de la Directive sur la prise de décisions automatisée ainsi que

les autres mesures de diligence requises pour respecter diverses autres obligations (par exemple, les obligations juridiques et celles relatives à la protection de la vie privée, à la sécurité, aux droits de la personne, aux droits linguistiques ainsi qu'à la gestion de l'information et de la technologie de l'information [GI-TI]).

Vous ne devriez envisager l'IA agentive que lorsque :

- les résultats visés sont clairement définis;
- les limites décisionnelles sont explicites;
- la responsabilité est clairement définie et attribuée;
- les risques peuvent être testés, surveillés et gérés tout au long du cycle de vie du système.

Le présent guide pourrait être mis à jour régulièrement afin de prendre en compte les évolutions technologiques.

2. Qu'est-ce que l'IA agentive?

De manière générale, l'IA agentive se définit davantage par ce que le système fait (ses opérations) que par ce qu'il produit (du contenu). L'IA agentive tend à être plus proactive (elle exécute des opérations et coordonne des étapes vers un objectif), tandis que l'IA générative est plus réactive (elle produit des résultats uniquement en réponse aux requêtes des personnes qui l'utilisent).

Les systèmes d'IA acquièrent une meilleure capacité à exécuter des opérations lorsqu'ils gagnent en autonomie, adoptent un fonctionnement orienté vers un objectif et peuvent exécuter des opérations sans supervision ni intervention humaine continue.

Les **agents d'IA** sont des systèmes qui peuvent comprendre leur environnement et exécuter des opérations en conséquence, souvent de manière autonome, pour atteindre des objectifs précis et adapter leur fonctionnement en fonction des changements de données ou de contextes ¹.

L'**IA agentive** désigne des systèmes composés d'un ou de plusieurs agents d'IA coordonnés, capables de décomposer des tâches, de collaborer, d'utiliser des outils externes et d'atteindre des objectifs sur de longues périodes sous une supervision humaine limitée.

Pour mieux comprendre les différents niveaux d'autonomie des systèmes d'AI agentive, examinons un exemple comportant quatre niveaux différents d'IA agentive.

Exemple : planification d'une réunion en utilisant une IA de plus en plus agentive

Niveau 1 – Assistance : à l'aide d'une IA qui agit plutôt comme une aide, le logiciel suggère des créneaux horaires disponibles pour une réunion, mais la personne doit choisir celui qui lui convient et envoyer les invitations. Cette fonction est souvent offerte dans les outils de calendrier existants.

Niveau 2 – Semi-autonome : au niveau suivant, le système d'IA est plus performant. Il propose un créneau horaire pour la réunion et rédige une invitation qui doit être approuvée.

Niveau 3 – Haute autonomie : un système d'IA planifie la réunion en fonction des autorisations déléguées et d'une instruction générale (« Planifie une réunion avec Marc, Lisa et Darren la semaine prochaine »). Il consigne ses opérations et avise la personne.

Niveau 4 – Autonomie adaptative : le système d’IA surveille les changements dans la disponibilité des personnes participant à la réunion, propose ou fixe un nouveau créneau horaire selon les contraintes définies et signale les exceptions.

La progression d’un outil d’assistance à un agent d’IA proactif et autonome illustre bien l’efficacité de l’IA agentive. Elle ne se contente pas de vous aider à accomplir des tâches, elle s’en charge pour vous, conformément aux autorisations déléguées que vous avez configurées, en s’adaptant aux changements de circonstances. Cette hausse des capacités offre plus de possibilités, mais comporte également plus de risques.

3. Possibilités et risques

► Dans cette section

L’IA agentive peut vous aider à accélérer votre travail, à accroître la charge de travail gérée et à améliorer la cohérence. Toutefois, elle peut également engendrer de nouveaux risques, car elle est capable d’effectuer des opérations (et pas seulement générer du contenu) dans plusieurs systèmes et selon différents niveaux d’autorisation, ce qui soulève des préoccupations tant ciblées que générales.

En général, à mesure que l’autonomie augmente, les possibilités et les risques augmentent aussi. Cela est particulièrement vrai pour les systèmes d’AI agentive, qui peuvent multiplier grandement les possibilités et les risques comparativement à l’utilisation d’un seul agent.

Dans de nombreux contextes du gouvernement du Canada, l’IA agentive est utilisée au mieux dans des flux de travail internes très ciblés, dont les limites sont claires et les autorisations restreintes, et où ce sont des fonctionnaires responsables qui prennent les décisions finales et autorisent les opérations qui s’ensuivent.

Possibilités

L'IA agentive peut créer de la valeur en réduisant le temps consacré au travail de routine, en prenant en charge des opérations de plus grande envergure et en aidant les équipes à adopter des approches cohérentes. Par exemple, un système d'AI agentive peut accomplir les tâches suivantes :

- rassembler des ressources publiques et des directives internes pour élaborer des projets de politiques;
- compiler des données provenant de sources approuvées et rédiger un texte explicatif visant à appuyer une analyse existante;
- exécuter des flux de travail plus complexes, tels que l'examen des demandes entrantes, la vérification de l'exhaustivité selon les règles d'admissibilité et l'extraction de fichiers pertinents à partir de systèmes internes approuvés;
- résumer ou signaler les renseignements contenus dans les fichiers à l'intention des personnes chargées de la prise de décision ou formuler des recommandations, par exemple en transmettant automatiquement les dossiers ambigus ou incertains pour examen immédiat et complet par des personnes chargées de la prise de décision.

Ces utilisations permettent de gagner du temps pour de tâches à plus grande valeur ajoutée, améliorer la capacité de réponse, réduire le fardeau administratif et aider le personnel à faire preuve de plus de discernement, à améliorer la qualité de leur travail et à mieux cerner les besoins de la clientèle.

Les systèmes d'AI agentive peuvent également faciliter la continuité du travail et la gestion des connaissances en transformant les tâches répétitives en listes de contrôle et en modèles, en trouvant des précédents pertinents et en consignant la justification des recommandations afin de garantir la traçabilité.

Dans l'ensemble, l'IA agentive est généralement plus utile lorsqu'il s'agit de tâches répétitives, chronophages et vérifiables, dont la supervision et la responsabilité des décisions prises incombent toujours aux humains.

Risques

Les principaux risques de l'IA agentive ne concernent pas seulement la qualité des résultats, qui pourrait être compromise par des biais et des erreurs, ou comporter des contenus préjudiciables. Ces risques concernent aussi les opérations non autorisées, les autorisations ambiguës, la responsabilité et la traçabilité. Les systèmes plus autonomes peuvent outrepasser leurs limites, tourner en boucle ou effectuer des opérations non souhaitées si les limites ne sont pas clairement définies ou si le système tombe sur du contenu non fiable. Il peut s'agir d'une information factuelle incorrecte ou des attaques par infiltration de requêtes.

Les risques augmentent aussi lorsqu'un agent se voit accorder un accès étendu aux systèmes du gouvernement du Canada ou à des renseignements sensibles sans configuration appropriée, lui permettant ainsi d'extraire ou de divulguer des renseignements qu'il ne devrait pas, ou d'effectuer des opérations que la personne n'avait pas l'intention de faire.

Dans des contextes où les conséquences sont importantes, comme l'octroi de subventions, l'approvisionnement, les activités de réglementation ou les décisions administratives touchant les droits, les finances ou l'accès aux services, d'autres risques s'ajoutent. Il pourrait notamment s'agir d'une atteinte possible à l'équité, d'un manque de transparence et de la difficulté à expliquer la façon dont les résultats ont été produits et utilisés. Ces conséquences pourraient entraîner des risques structurels plus importants, comme un risque juridique accru, la non-conformité à la loi ou la prise de mesures contraires à l'éthique. Ces dernières pourraient comprendre des

atteintes importantes à la vie privée, comme l'accès non autorisé à des renseignements personnels ou sensibles ou la communication de tels renseignements.

Enfin, la pression liée à la charge de travail peut mener, au fil du temps, à une dépendance excessive. Le personnel pourrait alors commencer à considérer les recommandations ou les classements obtenus comme des décisions réelles. Ce risque est présent, que les résultats de l'IA soient fiables ou non.

Le transfert continu de décisions aux agents d'IA peut entraîner une atrophie des connaissances et des compétences et un affaiblissement du jugement du personnel. Ce dernier devient moins habile dans les tâches essentielles et met plus de temps à remarquer quand quelque chose ne va pas. Ces risques peuvent toucher la qualité du service, la protection de la vie privée et la sécurité, la capacité de défendre juridiquement les décisions et la confiance du public.

Le maintien des compétences appropriées chez les personnes qui utilisent et déploient l'IA agentive constitue une condition préalable à une surveillance humaine efficace. Par conséquent, les fonctionnaires doivent avoir reçu une formation appropriée et bien comprendre l'utilisation responsable de l'IA agentive, y compris les limites quant aux situations dans lesquelles elle peut être utilisée et à la manière de le faire conformément aux pouvoirs existants. Cela contribue à réduire au minimum les risques pour la population, le personnel et le gouvernement du Canada, à atténuer la probabilité d'une utilisation non conforme et à favoriser l'adoption appropriée de l'IA agentive tout en évitant une utilisation abusive ou insuffisante de celle-ci.

4. Principes d'utilisation responsable

► Dans cette section

Pour atténuer les risques liés aux agents et aux systèmes d'IA, deux principes axés sur les agents d'IA ont été établis : une autonomie limitée et la récupérabilité.

Ces principes s'ajoutent aux principes PRETES (pertinente, responsable, équitable, transparente, éclairée et sécurisée) relatifs à l'utilisation responsable de l'IA générative. Ces principes vont également de pair avec les cadres de gouvernance existants du gouvernement du Canada et devraient être pris en compte parallèlement à ceux-ci.

1) Autonomie limitée

Ce que cela signifie

- Des paramètres stricts et explicites doivent régir le fonctionnement des agents d'IA afin de limiter l'accès aux données et aux outils, les autorisations et leur champ d'action, et de permettre une utilisation efficace sans qu'ils disposent de pouvoirs étendus et illimités.
- Le niveau d'autorisation des agents d'IA doit être clairement défini (par exemple, « ébauche seulement » ou « lecture seulement ») et une liste concise, rédigée en langage clair, de ce qu'ils peuvent et ne peuvent pas faire doit être affichée. Ces renseignements doivent être présentés de manière bien visible chaque fois qu'une personne utilise un agent.
- Les rôles des propriétaires et des responsables de chaque agent doivent être clairement désignés, enseignés et consignés. Cela comprend la responsabilité des agents d'IA qui ont leurs propres agents d'IA (ou sous-agents d'IA) pour accomplir des tâches.

- Il s'agit d'un contrôle préventif qui réduit le risque que des opérations dangereuses soient effectuées avant même qu'elles n'aient lieu. Les contraintes techniques qui limitent les capacités des agents devraient s'ajouter aux contrôles exercés par le personnel responsable. Ensemble, ils favorisent l'utilisation responsable de l'IA agentive.

Pourquoi c'est important pour les agents d'IA

Parce que les agents d'IA peuvent planifier et utiliser des outils, une petite erreur peut prendre des proportions importantes (par exemple, courriels de masse, mises à jour en bloc). L'établissement au préalable de limites claires et strictes permet de réduire les erreurs, de limiter les coûts et de minimiser les préjudices potentiels en cas de problème.

De plus, l'établissement de champs d'activité restreints, encadrés par des garde-fous, facilite la mise en œuvre des agents d'IA dans des contextes de production, permet de démontrer plus rapidement leur valeur et offre la possibilité d'accroître progressivement les autorisations à mesure que la confiance et le savoir-faire en matière de supervision du personnel s'améliorent.

2) Récupérabilité

Ce que cela signifie

- Les agents d'IA doivent être conçus de manière qu'ils puissent être facilement guidés, suspendus ou désactivés au besoin, et remis rapidement à un état sûr et stable. Pour ce faire, toutes les opérations effectuées par ceux-ci devraient être intégralement consignées dans un système qui ne peut être modifié par l'agent.
- Dans les situations où une « annulation » complète des opérations n'est pas possible, le personnel doit être en mesure d'assurer la fluidité

et le contrôle des opérations en présentant des aperçus, en procédant à des validations humaines et en proposant des options claires pour corriger les résultats inattendus, par exemple en les annulant, en les modifiant ou en informant les personnes concernées de manière fiable. Des examens juridiques devraient être menés pour recenser les cas où le recours à l'IA agentive serait inapproprié si une « annulation » complète des opérations n'est pas possible, ou lorsque les cadres législatifs ou réglementaires applicables ne permettent pas une telle inversion.

- De manière générale, les systèmes d'IA agentive devraient être conçus pour échouer de manière sécuritaire, dans l'hypothèse où les agents, les outils ou les justificatifs d'identité peuvent éventuellement être compromis.

Pourquoi c'est important pour les agents d'IA

Les agents d'IA bien conçus aident le personnel à obtenir des résultats responsables, même lorsque les tâches sont complexes. Les résultats générés peuvent varier, des modifications pourraient devoir être apportées aux plans, et l'utilisation d'outils pourrait avoir une incidence sur d'autres systèmes. Des choix de conception judicieux permettent de garantir une gestion des résultats imprévus.

En limitant les répercussions et en permettant une récupération rapide, il est possible de maintenir la confiance et de favoriser une utilisation sûre et efficace des capacités des agents.

Le fait d'accorder une importance particulière à la récupérabilité réduit l'incidence des erreurs et favorise une mise à l'essai et une intégration plus sécuritaires des agents dans les activités quotidiennes du gouvernement. Cela permet également aux gestionnaires de confirmer rapidement l'efficacité des garde-fous (par exemple, le délai d'interruption, le délai de

correction, le nombre de cas touchés). Enfin, la récupérabilité renforce aussi l’auditabilité en produisant un registre complet et horodaté des opérations et des interventions, ce qui facilite la reconstitution de ce qui s’est produit, la vérification des garde-fous et la description des mesures correctives.

5. Conseils pratiques pour le personnel du gouvernement du Canada (avant, pendant, après)

► Dans cette section

Les systèmes d’IA agentive peuvent planifier, exécuter des opérations et s’adapter à différents outils ou ensembles de données sans grande supervision. Ces capacités considérables offrent de grandes possibilités, mais soulèvent aussi des risques importants.

Mis à part la recherche et l’utilisation personnelle qui n’ont pas d’incidence sur la clientèle, l’utilisation, la mise à l’essai, l’acquisition ou le déploiement d’une IA agentive doit toujours faire l’objet d’une consultation préalable auprès des spécialistes (tels que des spécialistes du domaine de la protection de la vie privée, de la sécurité, des services juridiques ou de la gestion de l’information et les propriétaires de programmes) afin que les décisions de conception qui influencent les résultats et les risques puissent être convenablement traitées et atténuées.

Dans les contextes gouvernementaux, trois enjeux doivent être rigoureusement pris en compte lors de l’utilisation de l’IA agentive :

- **Interactions imprévues entre les systèmes** : lorsque les systèmes enchaînent des tâches ou utilisent des outils et des données externes d’une manière que nous n’avons pas anticipée. Ils peuvent alors

ou dépasser leur champ d'action prévu et/ou rendre publique de l'information sensible, exclusive ou protégée.

- **Exécution inadéquate des tâches** : même lorsque les requêtes semblent claires, le système peut mal cibler ses efforts, privilégier des raccourcis ou automatiser à l'excès des étapes qui nécessitent un jugement.
- **Surveillance et auditabilité** : si nous ne pouvons pas observer, expliquer ou reproduire ce que le système a fait, notre capacité à respecter les exigences de transparence et de responsabilité peut être limitée. Cela est particulièrement essentiel en cas de litige, d'audit ou d'enquête.

Les étapes proposées ci-dessous constituent des mesures concrètes qui peuvent être mises en œuvre **avant, pendant et après** l'utilisation de l'IA pour gérer ces risques tout en tirant parti de ses avantages. Ce guide porte principalement sur la définition de limites claires dès le départ, le maintien d'une supervision humaine pendant le fonctionnement du système et la consignation d'informations permettant de valider, d'expliquer et d'améliorer les résultats. Ces mesures concrètes sont suivies de deux études de cas illustrant les avantages d'une utilisation responsable de l'IA agentive.

Avant de commencer : accordez une attention particulière à la conception et établissez des limites sécuritaires²

a) Commencez par un cas d'utilisation ciblé et bien défini

- Définissez clairement l'objectif de l'agent d'IA. Quels sont les objectifs précis de votre initiative? Le recours à un agent d'IA ou un système d'IA agentive constitue-t-il la meilleure façon de résoudre le problème?

- Le gouvernement du Canada dispose-t-il déjà d'un agent ou d'un système d'IA agentive similaire qui pourrait être utilisé ou reproduit?
- Le cas échéant, utilisez des approches communes à l'échelle organisationnelle pour réduire le double emploi, assurer l'interopérabilité et favoriser une adoption sécuritaire et évolutive dans l'ensemble du gouvernement du Canada.
- Avant d'envisager des cas d'utilisation à forte incidence susceptibles de nécessiter l'automatisation, examinez les outils d'aide à la décision et les flux de travail internes à faible risque.
- Schématisez le processus en détail et consultez les parties prenantes et les partenaires concernés pour comprendre les restrictions juridiques et stratégiques relatives aux problèmes prévus et les réponses possibles aux problèmes imprévus (comme les spécialistes de programmes, les services juridiques, les services de protection de la vie privée, les services de sécurité, les services de GI-TI).
- Pensez à la traçabilité en vous assurant de pouvoir expliquer la manière dont le système d'IA agentive prend des décisions et produit des résultats.
- Déterminez si votre cas d'utilisation est visé par la Directive sur la prise de décisions automatisée.

b) Concevez en fonction d'une autonomie limitée

- Limitez et recensez les données auxquelles l'agent peut avoir accès, les outils qu'il peut utiliser et les opérations qu'il peut exécuter. Il s'agit notamment de mettre en place des limites en matière de données et de débit afin d'empêcher toute utilisation abusive par les agents d'IA.
 - Les droits d'accès accordés aux agents d'IA doivent s'appuyer sur une analyse des risques (quels préjudices pourraient résulter de l'octroi de ces accès?).

- Utilisez le mode « lecture seule par défaut » (afficher et rédiger une ébauche) dans la mesure du possible. Ajoutez la capacité de modifier ou d'exécuter des opérations (envoyer, mettre à jour, publier) au besoin et seulement pour cette étape précise.
- Établissez des identifiants d'agent clairs qui vous permettront de reconnaître facilement chaque agent d'IA, de le gérer et de suivre ses opérations.
- Veillez à ce qu'il soit facile de distinguer les cas où l'agent **donne une suggestion** de ceux où il **exécute une opération**.

c) Intégrez des points de contrôle effectués par des spécialistes (intervention humaine)

- Lorsque l'opération exécutée par un agent d'IA modifie l'état d'un système (envoi, publication, autorisation, dépense, mise à jour de dossiers), intégrez un processus de vérification/confirmation à la conception, sauf si les effets attendus sont manifestement minimes et que l'opération est facilement réversible.
 - Vous devez surveiller la sécurité et l'exactitude d'un agent d'IA pour déterminer s'il est fiable, en mettant en place des procédures de vérification/confirmation et des alertes en cas de fonctionnement inattendu. Si le rendement demeure constant au fil du temps, le risque résiduel peut être plus faible, mais il nécessite tout de même une surveillance continue (et une supervision de cette surveillance), surtout lorsque les conditions changent.
- Expliquez clairement la façon dont la surveillance et la responsabilité humaines sont intégrées dans le processus de l'agent et comment les problèmes sont signalés à l'échelon supérieur et à qui.
- Élaborez un plan précisant le niveau de supervision requis et les ressources qui y seront consacrées. Ce plan doit prendre en compte la

fatigue liée à l'intervention humaine et proposer des mécanismes de supervision évolutifs qui restent efficaces à mesure que l'utilisation de l'IA agentive progresse.

d) Privilégiez la réversibilité dans la conception

- Veillez à ce que les opérations soient réversibles dans la mesure du possible (par exemple, gestion des versions, options d'annulation, demandes de validation avant l'exécution d'opérations importantes). Si les opérations ne sont pas réversibles, évaluez l'acceptabilité des effets des opérations irréversibles.

e) Planifiez la façon dont vous effectuerez les essais

- Effectuez des essais en utilisant des cas limites réalistes et les pires scénarios possibles (y compris du contenu non fiable et clairement malveillant) afin d'évaluer les problèmes, les vulnérabilités et la nature des agents d'IA.
- Commencez vos essais dans un environnement contrôlé et élargissez-les progressivement pour mieux comprendre le fonctionnement des agents.

Pendant l'utilisation : travaillez de manière sécuritaire, misez sur la surveillance et évitez la « dérive de l'automatisation »

f) Facilitez la surveillance

- Veillez à ce qu'un rôle clairement défini soit attribué à la personne responsable des résultats et de la surveillance de l'agent d'IA, à ce qu'une procédure de renvoi aux niveaux supérieurs soit documentée et à ce qu'un remplaçant soit désigné en cas d'absence ou de mutation.

- La responsabilité incombe toujours en fin de compte à la personne responsable désignée, même lorsque l'agent fonctionne de manière autonome dans le cadre des autorisations approuvées.
- Mettez en place un processus permettant de passer en revue et de mettre à jour la liste des responsables lorsque les rôles changent, de même qu'une démarche claire pour suspendre ou désactiver la fonction de l'agent d'IA lorsque la responsabilité n'est pas clairement définie.
- Veillez à ce que la procédure de départ prévoie la suspension, à la réaffectation ou à la désactivation des agents d'IA utilisés par une personne responsable lorsqu'elle quitte son poste, son ministère ou le gouvernement du Canada, à moins que la responsabilité ne soit explicitement transférée à une autre personne désignée (cela s'applique aussi bien aux agents d'IA au niveau du programme qu'aux agents d'IA personnels).

g) Consignez les opérations et décisions clés

- Conservez les journaux des outils utilisés, des opérations exécutées, des autorisations accordées et des données d'entrée et de sortie clé. Pour les demandes d'outils, consignez le nom de l'outil, l'heure, le résultat et le type de paramètres transmis.
 - Les valeurs brutes des paramètres ne devraient être consignées que lorsque cela est nécessaire pour la surveillance, la sécurité ou l'intégrité du programme, et elles doivent respecter les contrôles ministériels en matière de protection de la vie privée et de sécurité.
- Assurez-vous que les journaux sont faciles à comprendre pour les audits et les enquêtes en cas d'incident.
- Assurez-vous que les journaux sont traités de manière appropriée lorsqu'ils sont sensibles.

- Assurez-vous de la conformité aux exigences liées à la gestion de l'information, à la tenue de documents et aux protocoles de conservation en cas de litige.

h) Surveillez la dérive de l'automatisation

- Surveillez le comportement humain pour détecter toute dépendance excessive et actualisez les attentes et la formation au besoin. Il peut arriver que les équipes en viennent à se fier aux évaluations ou aux résumés des agents IA comme s'il s'agissait de décisions réelles, même lorsque ces derniers sont conçus pour l'aide à la décision. De plus, les systèmes d'IA agentive peuvent également dériver au-delà de leur champ d'application prévu et doivent donc être encadrés et surveillés.
- Surveillez les dérives dans la qualité. Les résultats et le raisonnement peuvent progressivement devenir moins exacts, complets ou cohérents au fil du temps à mesure que les données, les outils ou les configurations changent. Les équipes devraient effectuer des vérifications ponctuelles et comparer les résultats aux attentes pour déceler rapidement tout écart.
- Demandez occasionnellement à des spécialistes d'effectuer manuellement les tâches de l'IA agentive afin de comparer leur rendement à celui de l'agent, et de déterminer s'il est nécessaire de former à nouveau le personnel ou l'agent IA, en fonction des résultats.

i) Traitez le contenu non fiable avec prudence

- Ne permettez pas aux agents d'IA de suivre automatiquement les directives intégrées, car le contenu non fiable peut contenir des instructions cachées (« infiltration de requêtes ») visant à modifier son fonctionnement. Traitez les textes externes ou fournis par les personnes qui l'utilisent comme des données à analyser, et non comme des instructions à suivre.

j) Soyez prêts à interrompre l'agent ou le système

- Maintenez un mécanisme de suspension et de désactivation (« interrupteur d'urgence ») indépendant de l'agent d'IA ou du système d'IA agentive ainsi qu'un plan de récupération approprié pour les opérations imprévues.

Après le déploiement : tirez des leçons, améliorez les agents et mettez-les hors service de manière responsable

k) Évaluez le rendement et les incidences

- Examiner les schémas d'erreurs, les preuves de partialité, les taux de renvoi aux niveaux supérieurs, les commentaires des utilisatrices et utilisateurs et si les résultats diffèrent selon les groupes ou les contextes. Déterminez si les objectifs ont été atteints de manière efficace et appropriée.
- Prévoyez des vérifications et des évaluations régulières (et des plans d'action connexes) pour l'utilisation continue de l'agent d'IA.

l) Réévaluez en cas de changement

- Réévaluez les risques et les contrôles si les outils, les sources de données, les autorisations ou le champ d'application de l'agent d'IA changent. De même, réévaluez les risques et les contrôles si le cadre juridique ou stratégique pertinent de l'agent d'IA change.

m) Mettez les agents hors service de manière sécuritaire

- Si un agent d'IA ou un système d'IA agentive n'est plus nécessaire, supprimez-en l'accès, archivez les documents requis de manière appropriée et consignez les leçons apprises. La communication des

leçons apprises et des pratiques exemplaires aide à renforcer la capacité et la compréhension de l'IA au gouvernement du Canada.

6. Études de cas : utilisation responsable de l'IA agentive

► Dans cette section

Les études de cas suivantes illustrent comment l'utilisation responsable de l'IA agentive peut efficacement atténuer les schémas d'échec courants dans les systèmes d'AI agentive.

Étude de cas 1 : l'agent de service « serviable » qui outrepassa ses fonctions

Scénario d'échec : un agent de triage des demandes de service rédige des réponses et achemine des cas. Pendant une période d'afflux, il commence à clore des billets après avoir rédigé des réponses, et ce, même si sa conception ne prévoyait que la rédaction de réponses. Dans le système de gestion des billets, il arrive que certaines opérations relatives à la rédaction de réponses ou au transfert des dossiers aient pour effet de marquer automatiquement un billet comme « résolu ou fermé » à moins d'une indication contraire explicite. Le service fourni à la clientèle est alors incomplet et le personnel doit rouvrir le dossier.

Utilisation responsable : avant de lancer l'agent de triage de demandes de service, les équipes prévoient la réalisation de mesures rigoureuses de protection, de surveillance et de contrôles par le personnel. L'agent fonctionne en mode « ébauche seulement », sans capacité d'envoyer des réponses définitives ou de changer le statut des billets. Seule une courte liste d'opérations est autorisée. Si l'agent d'IA tente un changement de

statut semblable à une fermeture ou déclenche un flux de travail à cette fin, il se met automatiquement en pause et alerte l'équipe. Le personnel dispose d'un bouton d'arrêt en un clic et d'une procédure simple et testée au préalable pour rouvrir les billets touchés.

Lorsque l'augmentation soudaine du nombre de demandes de service a amené l'agent d'IA à déclencher la fermeture de certains dossiers après la rédaction des réponses, les mesures préventives mises en place ont fonctionné comme prévu. Le problème a été immédiatement détecté, signalé et rapidement corrigé. Toutes les opérations ont été consignées dans un journal d'activités. Le personnel a rouvert les dossiers touchés, assurant ainsi la prestation d'un service complet à la clientèle. L'incident a permis de valider l'approche de conception proactive et de recueillir des renseignements utiles qui ont encore renforcé la fiabilité du système quant à la gestion des périodes de pointe à venir.

Étude de cas 2 : le système est induit en erreur par du contenu non fiable

Scénario d'échec : un agent de surveillance analyse des documents et suit les instructions intégrées dans le contenu qu'il lit, ce qui l'amène à communiquer des renseignements ou à exécuter des opérations qu'il ne devrait pas. Des renseignements sensibles sont communiqués.

Utilisation responsable : lors de la conception de l'agent d'IA, le personnel a pris conscience du risque que représente le contenu non fiable et des problèmes potentiels qu'il peut causer. Avant le déploiement, il a analysé les faiblesses potentielles de l'agent d'IA au moyen de requêtes conçues pour l'induire en erreur afin de s'assurer qu'il fonctionnait de manière appropriée. Le personnel a conçu l'agent d'IA pour qu'il traite tout ce qu'il lit comme non fiable et pour qu'il ignore toute instruction contenue dans les documents. De plus, l'agent d'IA ne peut exécuter qu'une courte

liste d'opérations autorisée (par exemple, ajouter une étiquette, créer une alerte, acheminer pour examen). Il dispose d'un accès en lecture seule et la fonction de communication à l'externe est désactivée par défaut. De plus, à mesure que l'agent traite l'information, il épure les données saisies (texte brut seulement, aucun lien, script ou macro), et limite son analyse aux sources approuvées.

En cas d'anomalie, l'exécution est mise en pause et examinée avant que quoi que ce soit ne soit communiqué. Le personnel est prêt à utiliser un simple interrupteur d'arrêt pour arrêter l'agent immédiatement, si nécessaire. Une conception réfléchie a permis à l'agent d'accomplir ses tâches sans que des opérations risquées puissent être provoquées par des manœuvres malveillantes.

Nous joindre

Pour toute question concernant l'interprétation d'un aspect quelconque de ce guide, communiquer avec le Secrétariat du Conseil du Trésor du Canada.

Pour toute question concernant ce guide, le personnel des ministères peut envoyer un courriel à ai-ia@tbs-sct.gc.ca.

Glossaire

Agent d'IA

Systèmes capables de percevoir leur environnement et d'exécuter des opérations en conséquence, souvent de manière autonome, pour atteindre des objectifs spécifiques, et de s'adapter en fonction de l'évolution de l'information ou du contexte.

Dérive de l'automatisation

Changement progressif du fonctionnement ou de l'utilisation d'un système automatisé au fil du temps, si bien qu'il s'écarte de l'intention de

conception initiale ou du champ d'application approuvé.

IA agentive

Systèmes composés de plusieurs agents d'IA coordonnés, capables de décomposer des tâches, de collaborer, d'utiliser des outils externes et d'atteindre des objectifs sur de longues périodes, moyennant une supervision humaine limitée.

Infiltration de requêtes

Instructions malveillantes cachées qui tentent de tromper un agent d'IA de façon à l'amener à ignorer les règles prévues et à exécuter des opérations non autorisées.

Intervention humaine

Contrôle du flux de travail selon lequel un agent d'IA peut rédiger une ébauche ou formuler une recommandation, mais une personne doit examiner et approuver (ou corriger) les résultats ou les opérations proposées à des points définis, surtout avant que des opérations à forte incidence ne soient exécutées.

Limites de fréquence et limites de données

Contrôles techniques qui fixent une limite à la fréquence à laquelle un agent IA peut exécuter des opérations (limites de fréquence) et restreignent les données auxquelles il peut accéder, utiliser ou communiquer (limites de données) afin de réduire la probabilité et les incidences d'un fonctionnement imprévu.

Outils

Capacités externes (comme les opérations liées à la recherche, aux bases de données ou au flux de travail) qu'un système d'IA est autorisé à utiliser pour récupérer des renseignements ou exécuter des opérations dans d'autres systèmes.

Réversibilité

Mesure dans laquelle les opérations d'un agent d'IA peuvent être inversées au moyen de contrôles, comme les aperçus, les autorisations, la gestion des

versions, les points de restauration ou les procédures de retour en arrière (y compris les étapes claires pour les mesures correctives lorsqu'une « annulation » complète n'est pas possible).

Vérifiabilité

Capacité de retracer, d'examiner et de vérifier les résultats et les opérations d'un agent d'IA à l'aide de registres horodatés (comme les journaux et les approbations), y compris ce que l'agent a fait, ce à quoi il a eu accès et ce qu'il a modifié.

Liens connexes

► Dans cette section

Instruments de politique connexes

- [Politique sur les services et le numérique](#)
- [Directive sur la prise de décisions automatisée](#)
- [Guide sur l'utilisation de l'intelligence artificielle générative](#)

Autres références

- [Stratégie en matière d'intelligence artificielle pour la fonction publique fédérale 2025-2027](#)
- [Guide sur les responsabilités ministérielles en matière d'intelligence artificielle](#)
- [Les 10 mesures de sécurité en matière d'intelligence artificielle : Introduction – ITSAP.10.049](#)
- [Executive summary: Artificial intelligence and competitive dynamics in downstream markets](#) (anglais seulement)
- [The agentic AI landscape and its conceptual foundations](#) (anglais seulement)

- [Artificial Intelligence Risk Management Framework \(AI RMF 1.0\)](#).
[anglais seulement]
- [Artificial Intelligence Playbook for the UK Government](#) (anglais seulement)
- [Article 9: Risk Management System | EU Artificial Intelligence Act](#) (anglais seulement)
- [LLM Prompt Injection Prevention](#) (anglais seulement)
- [How Microsoft defends against indirect prompt injection attacks](#) (anglais seulement)
- [Principes pour des technologies de l'intelligence artificielle \(IA\) générative responsables, dignes de confiance et respectueuses de la vie privée](#)

Notes de bas de page

- 1 Les termes « agents d'IA » et « IA agentive » sont tous les deux utilisés conformément aux définitions de l'Organisation de coopération et de développement économiques ([Executive summary: Artificial intelligence and competitive dynamics in downstream markets](#)).
- 2 Les pratiques exemplaires en matière d'IA agentive dans cette section s'ajoutent aux pratiques exemplaires recommandées décrites dans le [Guide sur l'utilisation de l'IA générative](#).